Chuzhditsa: a phonological citation register for Cyrillic

Anastas Stoyanovsky*
Unaffiliated anastas@stoyanovs.ky

Abstract

Between ordinary orthography and specialist transcription lies a layer that alphabets fill only patchily: a conventional way to write a foreign word in the native script so that a reader can recover an adequate approximation of its source pronunciation without technical training. No major Slavic Cyrillic orthography occupies that layer — foreign words enter through ad hoc transcription, with systematic indeterminacy as the result (Bulgarian Вашингтон ∼ Уошингтън, Данчев 1982). We present *Chuzhditsa* (Bulg. чуждица, 'foreignism'), a computable, Cyrillic-based *phonological citation register* for foreign words, names, and pedagogical transcription. Japanese katakana proves that a visually distinct foreignness register can operate at the scale of a national orthography — though katakana adapts loans *into* native phonology, where a citation register must extend the inventory to *preserve* source contrasts.

Chuzhditsa supplies (i) a compositional grammar of diacritic operations deriving letters for non-Slavic phonemes from attested Cyrillic material — revivals, imports, and rule-derived residue; (ii) a prosodic layer for stress, tone, and length; and (iii) a register-marking typeface whose substitution and positioning rules (GSUB/GPOS) are the orthography's fusion and attachment rules, executable rather than merely described. We evaluate coverage against twenty languages under declared reference varieties and a three-way classification (exact, conventionally merged, declared loss), generalize the Bulgarian pilot to the Slavic Cyrillic superset with an external stress test on Albanian, and offer the laryngeal series' articulatory-to-visual gradient as a candidate heuristic for alphabet extension, not a demonstrated law.

1 Introduction

Every literate community borrows words faster than its orthography naturalizes them — an instance of the general problem of fit between a writing system and a changing language (Coulmas 2003). What most orthographies lack is a *productive* middle layer between everyday spelling and technical transcription: a conventional register for writing a foreign word so that its source pronunciation is recoverable. The layer is not empty — monolingual dictionaries fill it with pronunciation respelling, Japanese with furigana glosses, and the zonal constructed language Interslavic with an extended Cyrillic that already revives the yuses (Merunka 2018) — but no major Slavic Cyrillic orthography has an inventory-*extending*, register-marked device of this kind.

Japanese shows the layer can be institutionalized: katakana is a full parallel syllabary whose use *is* the marking of foreignness, and whose adaptation conventions ([naɪt] → ナイト) are conventionalized enough that, for a given source pronunciation, writers usually converge on one spelling.

Loanwords enter Bulgarian, Russian, Serbian or Ukrainian through ad hoc transcription: sound-based and spelling-based renderings compete (ланч ∼ лънч for *lunch*), phonemes without Cyrillic letters collapse unpredictably (/θ/ → т or с; /w/ → в or у), and nothing in the written form signals that a word is a guest rather than a native.

*Fonts, build toolchain, and the twenty-language audit are available in the accompanying repository (github.com/anastasds/chuzhditsa); the typeface used for all object-language examples in this manuscript is the system being described.

This paper describes Chuzhditsa, a designed orthographic register that supplies the missing device for Cyrillic, piloted on Bulgarian and then generalized to the Slavic superset — the union of the Slavic Cyrillic alphabets.

One distinction must be drawn before anything else, because three different tasks hide under "writing foreign words," and a system is judged unfairly when measured against a task it does not attempt. *Loanword adaptation* bends a word to the recipient phonology — katakana's main business, and the business of every integrated borrowing. *Phonemic transcription* records the source phonology in a technical alphabet — the IPA's business. Between them sits *phonological citation*: writing a foreign word, name, or example so that a native reader can recover an adequate approximation of its source pronunciation, in the native script, without technical training.

Chuzhditsa is primarily the third: unlike katakana it *extends* the recipient inventory to preserve source contrasts rather than collapsing them into native phonotactics; unlike the IPA it stays inside Cyrillic reading habits. Its use as a marking register for not-yet-naturalized loans is a secondary, optional application, and a loan that naturalizes is expected to leave the register.

The system makes three claims of general interest to writing-system design.

First, that a citation register can be built as a *compositional diacritic grammar*: rather than inventing letters ad libitum, a small set of recurrent operations — most featurally atomic, one (the hook) relational to its base letter (§4) — derives the entire extension inventory. The analogy to Hangul is one of compositional logic, not of Sampson's (1985) strict shape-iconicity; systematic diacritic derivation is what Czech, Vietnamese, and the IPA already do, and only the laryngeal gradient (§4) reaches for genuinely featural letterforms (Sampson 1985; Kim-Renaud 1997). Second, that such a register can be assembled almost entirely from attested material: dead letters of the target orthography's own history and living letters of related orthographies, with invention reserved for genuine residue. Third, that the register-marking function that Japanese distributes over a second character inventory can instead be carried by a second *typeface register* — realized here as an OpenType family whose substitution (GSUB) and positioning (GPOS) rules are the orthography's fusion and attachment rules, executable rather than merely described.

2 Background

2.1 Katakana as an adaptation layer

Katakana's success rests on three properties (Irwin 2011; Stanlaw 2004): conventionalized conversion that renders the loan's *adapted Japanese pronunciation* — an adaptation that has historically travelled by ear, by dictionary, and sometimes by spelling (Irwin 2011: 76–79) — rather than transliterating the source orthography letter for letter, the orthographic reflex of loanword adaptation as a systematic mapping onto native categories, whether modelled phonologically or perceptually (Kang 2011; Peperkamp & Dupoux 2003); graphic distinctness sufficient to mark register at a glance; and extensibility (ヴ for /v/, ファ for /fa/) when the inventory runs out. Chuzhditsa adopts all three as requirements, and hardens the first: its input is pronunciation only. The idealization should be stated as such — real katakana practice varies by borrowing period, donor language, brand convention, and register (Irwin 2011) — and Chuzhditsa aims to be what katakana is often described as being: the target is a convention strict enough to converge, not a claim that Japanese practice already converges. Katakana also carries a social record that a designed analogue cannot disclaim: loanword marking has been read as a badge of incomplete nativeness, and katakana is deployed to mark foreigners' accented Japanese as well as foreign words (Irwin 2011; Robertson 2020). We therefore state the register's semantics narrowly: it asserts that a word's phonology lies off the native grid — nothing about its right to exist in the language, and nothing about its user. Its proposed semantics is lexical rather than ethnolinguistic — it marks words, never speakers — though a designed intention is not a sociolinguistic guarantee: registers acquire social meanings their designers do not control, and whether readers would honor the

boundary is an empirical question for user studies. A naturalized loan is expected to exit the register, exactly as テンプラ became 天ぷら.

2.2 The Cyrillic commonwealth

Cyrillic today serves more than fifty languages, whose principal non-Slavic adaptations Comrie (1996) surveys, and the non-Slavic ones have already engineered letters for most sounds Slavic lacks: Kazakh ҚҚ /q/, ҢҢ /ŋ/, ҒҒ (/ʁ/ in Kazakh, reassigned here to /ɣ/: as with ХХ, the borrowing is of the grapheme and its construction); Bashkir ҫҫ /θ/, ҙҙ /ð/; Tajik ӣӣ /iː/ and ҲҲ (/h/ in Tajik, reassigned here to /ħ/: the borrowing is of the *grapheme and its compositional construction* — x plus the place hook — not of its Tajik value, and the Caucasian digraph xI is the nearer phonetic precedent); Altai өө /ø/, үү /y/; Belarusian ўў /w/; Serbian ҷҷ /dʒ/; the Caucasian palochka Ӏ. These letters carry decades of typographic and pedagogic practice. A design that imports them inherits that practice for free — an option unavailable to katakana's designers and, we will argue, the single largest economy in the present system. Modifying Cyrillic is not, however, a politically neutral act: script is a charged identity marker across the region — Serbian Latin/Cyrillic digraphia, the state-driven Latinizations of Kazakh and Moldovan, the recurring Cyrillic-versus-Latin debates — and a proposal that *adds letters* must be read in that light. Chuzhditsa mitigates the charge structurally: it alters no existing orthography, imports only attested material, and petitions no authority (§11, §Adoption). It does not pretend the act is neutral.

2.3 The Glagolitic precedent

Bulgaria is a primary locus of a script replacement: Glagolitic, the first Slavic alphabet, created — or, on Cubberley's reading, formalized — by Constantine-Cyril for the Moravian mission of 863, was gradually displaced by Cyrillic, devised by Constantine's disciples at the Preslav Literary School in the 890s on the model of Greek uncial, drawing on Glagolitic mainly for the sounds Greek lacked (Schenker 1995; Cubberley 1996). Glagolitic then survived for a millennium as a special-register script (Croatian liturgical use into the twentieth century; the last *wholly* Glagolitic missal appeared in 1905 (Parčić), and the final printed Glagolitic missal (Vajs 1927) set only the Ordinary of the Mass in Glagolitic beside Latin transliteration). Chuzhditsa knowingly reverses the vector: the visual DNA of the *oldest* Slavic letterforms — round bowls, even low-contrast proportions (the shipped family is two-case, but keeps the unicameral source's evenness) — marks the *newest* words, and the letters it revives (ѫ ѧ ѩ ѭ) inherit the sound-slots of Glagolitic ancestors (the Cyrillic shapes themselves are reworkings). The big yus ѫ in particular survived in Bulgarian spelling until the 1945 orthographic reform (Comrie & Corbett 1993), by then carrying a non-nasal value, its nasality lost by roughly the eleventh century (Cubberley 1996).

3 Design principles

1. **Determinism, relative to normalized phonological input.** One *specified input* phoneme, one rendering, within a declared mapping: given the same source pronunciation in the reference variety, two transcribers converge. The choice of source pronunciation itself (dialect, variant, spelling pronunciation) is outside the system's jurisdiction, and loss is permitted but must be declared (§9).

2. **Compositional transparency.** Every diacritic operation encodes one recurrent phonological *relation* to its base letter and applies wherever that relation appears. Some operations are featurally atomic (nasalization, aspiration, length); one is relational rather than atomic — the place hook means "place shifted off the native grid," interpreted relative to the base and *not* in one articulatory direction (retraction on кк хх, advancement on сс зз, retroflexion on тт дд рр; §4) — so the operations form a compositional diacritic grammar, featural in Sampson's extended sense but

not in the strict shape-iconic sense Hangul exemplifies; only the laryngeal gradient approaches the latter.

3. **Pair symmetry.** Obstruent letters enter the system in voiceless/voiced pairs aligned with the native series. This principle outranks heritage: the historical theta-letter fita (ѳ) was rejected because it has no voiced partner and derives from no rule; the Bashkir pair ҫ/ҙ (/θ//ð/) does its work systematically.

4. **Precedent first.** Revive native dead letters, then import living Cyrillic letters, and only then derive new combinations; never invent an unattested base shape.

5. **Pan-Cyrillic legibility.** Every base skeleton follows the reader's existing Cyrillic, and every extension should be guessable at first sight by any Cyrillic reader from its base letter plus a transparent mark — a design *goal*, aimed at rather than achieved, of the kind the orthography-development literature calls ease of transfer (Cahill & Rice 2014).

   Three properties must be kept distinct here: that a letter exists in Unicode, that it has established use in *some* Cyrillic orthography, and that it will be transparently interpretable to readers of *any* Cyrillic orthography. This paper demonstrates the first two — a Bulgarian reader has no prior acquaintance with Kazakh ҡ or Bashkir ҙ, and importing an attested letter buys typographic and pedagogic precedent, not automatic legibility. The third is the criterion the design aims at, and it remains an empirical claim for the reader study of §11.

6. **Register by typeface.** Foreignness is marked by letterform register, not by punctuation or a second inventory: text set in the Chuzhditsa face is thereby marked — though where katakana marks by a stable script-*encoding* contrast, this marks by a *rendering* contrast that a plain-text channel loses (§11); in the terms of the biscriptality literature, a *diaphasic diglyphia*: functional differentiation between two glyphic variants of one script (Bunčić, Lippert & Rabus 2016; the nearest historical parallel is the German practice of setting foreign words in roman within blackletter text; other register mechanisms — italics, small caps, Japanese ruby glosses — trade distinctiveness against availability and learning cost, and a bespoke face buys the first at the expense of the second two), and a socially meaningful orthographic choice in Sebba's (2007) sense.

4 The compositional grammar

Table 1 gives the full operation set. The place hook generalizes a precedent internal to the Slavic base inventory itself — щ *is* ш with a hook — and the glide breve generalizes и → й. Where Unicode supplies precomposed hooked letters (ҡ ӈ ҳ ҫ ҙ) they are used; elsewhere the same hook is written with combining U+0322 (ҭ ԥ ҕ).

Table 2 lays out the complete designed inventory — every letter the system adds to a Slavic Cyrillic base, set in the typeface itself.

4.1 The laryngeal gradient

The three back fricatives that plague Cyrillic transcription of Arabic, English and German — /x/, /ħ/, /h/ — receive a designed gradient: articulatory weakening maps to graphic lightening. Native х (/x/∼/χ/) is two full crossing strokes, the densest friction; pharyngeal ҳ (/ħ/) is х pulled deeper by the place hook; glottal һ (/h/∼/ɦ/) is a single open arc, literally the most open glyph of the three; and aspiration ◌ʰ, a feature rather than a segment, is һ reduced to a superscript. Arabic exhibits the full triple within ordinary vocabulary: Һилāл /hila:l/, Мухаммад /muħammad/, Хāлид /xa:lid/. We suggest the mapping *articulatory gradient* → *visual gradient* as a candidate principle for alphabet extension — a heuristic with a lineage in the featural conscripts, from Bell's *Visible Speech* (1867), where glyph shape encodes

Mark	Operation	Derivations
◌̆ breve	vowel → glide	и→й /j/ · у→ў /w/
◌̩ hook below	place shifted off the native grid	ԛ /q/ · ҳ /ħ/ · ԩ /ŋ/ · ç /θ/ · ʒ /ð/ · ţ /ʈ/ · ạ /ɖ/ · ƿ /ɽ/
◌ꙁ crossbar	lenition	г→ґ /ɣ/
◌̈ double dot	vowel fronting	ä /æ/ · ö /ø œ/ · ÿ /y ʏ/
◌̨ yus tail	nasalization	ѫ /ɔ̃ õ/ · ѧ /ɛ̃/ · ą /ã ẽ/ · ǫ /œ̃/
◌ʰ ◌ʰ raised h	aspiration; breathy voice	пʰ тʰ кʰ чʰ · бʰ дʰ гʰ цʰ
◌̄ macron	vowel length	ӣ ȳ (precomposed) · ā ē ō
ˈ palochka	laryngeal–pharyngeal action	тˈ кˈ цˈ (ejectives) · bare ˈ = /ʕ/
ь	palatalization	нь /ɲ/ · ль /ʎ/ · хь /ç/ · сь /ɕ/

Table 1: The compositional operations. Each mark encodes a stable operation; most are featurally atomic, while the hook is relational to its base (§4). Precomposed Unicode letters are used where they exist.

articulation directly, through Shavian and Quikscript, though those aimed at whole-language reform rather than a citation register.

4.2 The Indic stress test

Hindi–Urdu and Bengali cross a four-way laryngeal contrast with a dental/retroflex place contrast. The grammar covers the entire grid with zero new letters: т /t̪/, тʰ /t̪ʰ/, д /d̪/, дʰ /d̪ʰ/; retroflex ţ ţʰ ạ ạʰ, plus ԩ /ŋ/, ƿ /ɽ/, ƿʰ /ɽʰ/. A single word such as ţʰӣк (*ṭhīk*, /ʈʰiːk/) composes three features on one base — hook, aspiration, length — which is precisely the compositional-grammar claim made concrete.

5 The ligature stratum

Cyrillic has always ligated: ю is a historic ligature of і+оу (the у element later dropped; Cubberley 1996), Serbian ль нь were fused from л/н+ь by Vuk Karadžić in 1818, building on Mrkalj's digraphs (Comrie & Corbett 1993), and Old Bulгarian possessed iotated nasals. Chuzhditsa's ligature stratum is therefore conservative, and it is implemented as executable GSUB rules rather than prose: (i) нь ль fuse to нь/ль-forms automatically (нињо, фамиљя set with fused palatals); (ii) the iotated yuses ѩ /jɛ̃/ and ѭ /jɔ̃/ are revived as directly typeable letters, and the sequences ь+ѧ and ь+ѫ fuse to them automatically (French *bien* → бѩ; Polish *pięć* → пѩч); (iii) ejective digraphs тˈ кˈ пˈ цˈ чˈ fuse with the palochka *raised to the upper half*, echoing the apostrophe of scholarly romanization ([tʼ kʼ]) — a fused full-height palochka was rejected in testing because тˈ became confusable with п. The governing rule: only sequences denoting a single phoneme, with Slavic or Old Cyrillic precedent, may fuse; mechanical soft pairs (сь хь) remain visibly two-part.

6 The prosodic stratum

Stress takes the grave of Bulgarian lexicography (бȁбушкъ); Mandarin tones 2–4 take acute, caron, grave, with tone 1 unmarked — pinyin minus the macron, which Chuzhditsa reserves for length. The unmarked first tone resolves the collision by construction: мā can only mean a long vowel. Three costs are declared rather than hidden. Tone and stress marks are homoglyphic across source languages, resolved only by knowing the lexical source (katakana shares exactly this property). Diacritic stacking degrades: Vietnamese *phở* wants quality, length and tone at once (фъ̉), and the running-text rule drops length first (фъ̉). And four marks cannot represent six-tone systems without folding (Vietnamese *hỏi*/*ngã* merge on the caron; *nặng* falls together with *huyền* on the grave). The design verdict generalizes katakana's own: prosody is optional annotation — a dictionary layer, not an orthographic one — and the system claims no tonal adequacy beyond that annotative role.

Letter	Unicode	Value	Construction	Origin
Ўў	U+040E/045E	/w/	y + breve	import: Belarusian
Ԉԉ	U+040F/045F	/ʤ/	own shape	import: Serbian
Ҫҫ	U+04AA/04AB	/θ/	с + hook	import: Bashkir
Ҙҙ	U+0498/0499	/ð/	з + hook	import: Bashkir
ҚҚ	U+049A/049B	/q ɢ/	к + hook	import: Kazakh
Ҥҥ	U+04A2/04A3	/ŋ/	н + hook	import: Kazakh
Ҳҳ	U+04B2/04B3	/ɦ/	х + hook	import: Tajik
Ғғ	U+0492/0493	/ɣ/	г + crossbar	import: Kazakh (/ʁ/ there; construction borrowed)
Ԋԋ	U+04BA/04BB	/h ɦ/	own shape	import: Tatar
Ѕѕ	U+0405/0455	/ʤz/	own shape	revival; living in Macedonian (dzelo)
Ӏ	U+04C0	/ʕ/; т' к' ԥ'	palochka	import: Caucasus
Ţ Ḍ Ṇ Ꝑ	base + U+0322	/ʈ ɖ ɳ ɽ/	base + hook	derived by rule
◌ʰ	U+02B0	aspiration	raised miniature ɦ	derived
◌ʱ	U+02B1	breathy voice	raised hooked ɦ	derived
Ӓӓ	U+04D2/04D3	/æ/	a + double dot	import: Mari
Ӧӧ	U+04E6/04E7	/ø œ/	о + double dot	import: Altai
Ӱӱ	U+04F0/04F1	/y ʏ/	у + double dot	import: Altai
Ыы	U+042B/044B	/ɨ/; Korean /ɯ/	own shape	import: Russian; also Old Bulgarian
Ѹѹ	U+046A/046B	/õ õ/	nasal о	revival (shape to 1945; nasal value ∼11th c.)
Ѧѧ	U+0466/0467	/ɛ̃/	nasal е	revival: Old Bulgarian
Ѩѩ	U+0468/0469	/jɛ̃/	ligature ь+ѧ	revival: Old Bulgarian
Ѭѭ	U+046C/046D	/jõ/	ligature ь+ѫ	revival: Old Bulgarian
ą ǫ	V + U+0328	/ã ɛ̃ œ̃/	vowel + yus tail	derived by rule
Ӣӣ Ӯӯ	U+04E2/3, EE/EF	/i: u:/	vowel + macron	import: Tajik (ӯ is /ɵ/ there)
◌ ̄	U+0304	length	macron	derived
◌ ́ ◌ ̀ ◌ ̌	U+0301/0300/030C	tone, stress	dictionary layer	derived

Table 2: The complete designed inventory. The letter column is set in the Chuzhditsa face. Every letter is a standard Cyrillic-block codepoint; the combining and modifier characters are likewise standard Unicode (The Unicode Consortium 2024).

7 From Bulgarian to the Slavic superset

The pilot treated Bulgarian's thirty letters as the base. Nothing in the compositional grammar is Bulgarian-specific, however, and the generalization is administrative rather than structural. We use *superset* in its set-theoretic sense — the union of the Slavic Cyrillic character inventories — and intend no pan-Slavist claim; the older political resonance of a "unifying" Slavic frame is precisely not the argument, as the Ukrainian constraint below makes concrete. The base becomes each language's own Cyrillic alphabet (for the phoneme inventories, see Comrie & Corbett 1993), and the extension and prosodic strata apply unchanged.

Concretely, the typeface's character set was widened to the Slavic Cyrillic superset — Russian э ё ы, Ukrainian і ї є ґ, Belarusian ў, Serbian ј ѣ ѣ ѣ ѣ ѣ, Macedonian ѕ ѓ ѓ — so that any Slavic Cyrillic text sets natively. (ѕ is doubly loaded: a distinctive letter of the Macedonian alphabet and, as such, an emblem in the Bulgaria–North Macedonia language dispute. It is used here as shared Old Cyrillic heritage — current Cyrillic-wide before the nineteenth century, readopted by the 1944 Macedonian commission (Comrie & Corbett 1993) — claimed as the property of neither side.)

One design conflict surfaced and is instructive: the pilot's round e-form was, in fact, the shape of Ukrainian є; pan-Slavic coverage forced є to a closed bowl — aperture roughly half the open forms' — so its silhouette parts from the open epsilon of є, which keeps the wide aperture and a half-bar. Extending a script for one language can silently occupy a sibling's letterform space; superset-first design catches this early.

For each language the system's marginal contribution differs, and several languages turn out to need the extension layer for *native* material, not only loans: Slovak *ä* is ӓ and Slovak *h* is /ɦ/ → ɦ; Slovenian's native schwa (*polglasnik*) is exactly ъ (пъс /pəs/ 'dog'); Serbian's pitch accents are the

prosodic layer's use case; Polish *ą ę* are the yuses meeting their own descendants (мѫка, пѭч). Latin-alphabet Slavic languages connect through a deterministic bridge: Gaj's digraphs map one-to-one onto Serbian Cyrillic letters (*lj nj dž* → љ њ џ), and the residue is rule-derived — Czech *ř* /r̩/, the notorious singleton, is written р̌ with the caron it already wears; Slovak long syllabic liquids *ĺ ŕ* take the length macron (вр̄ба /vr̩:ba/).

The generalization also has ancestors to acknowledge: the determinism principle is Vuk Karadžić's *пиши као што говориш* avant la lettre, and Serbian orthography contributed џ љ њ ј — and the very idea of fused palatals — before this project existed; Ukrainian orthography constrained the design as a peer, not a client (the є letterform conflict above was resolved in Ukrainian's favor). The Bulgarian name records the pilot, not a hierarchy.

The citation-register framing predicts something checkable across the family: *integrated* loans differ by language, because each language adapted the word to its own habits, while the *citation* form is shared, because its input is the source pronunciation. Table 3 shows the pattern.

Source	Bulgarian	Russian	Ukrainian	Serbian	shared citation form
weekend	уикєнд	уик-энд	вікєнд	викєнд	ўӣкєнд
jazz	джаз	джаз	джаз	џез	џäз

Table 3: Integrated loans vary by recipient language; the citation register does not, because its input is the source pronunciation rather than any recipient's adaptation.

As an external stress test we applied the system to Albanian, a non-Slavic Balkan language: every Albanian phoneme lands on an existing letter (*ë* /ə/ → ъ, *th/dh* → ç/ȝ, *y* → ӱ, *xh* → џ, *x* /dz/ → ѕ, *q/gj* → кь/гь): Шкьипърйа, Тирàнъ. A system built for Slavic phonology plus its compositional operations appears to cover the Balkan sprachbund without amendment. The practical beneficiary is not Albanian itself, which is securely Latin-written, but Cyrillic-language text citing Albanian: the *ë* that Macedonian and Serbian renderings currently drop is exactly ъ.

8 Typeface implementation

The Chuzhditsa family (Regular, Bold, Italic, Bold Italic; TTF and CFF/OTF) is generated by a compact parametric builder: every glyph is declared as centerline strokes — lines, arcs, and the full circles that give the face its Glagolitic character — and expanded to outlines with weight and slant as parameters. The construction itself — terminals, joins, the optical-correction canon, fitting, and the small-size behavior of the marks — is the subject of the companion paper; here we state only its orthographic consequences.

Capitals stand at 700 units, lowercase at an x-height of 500 with true two-case architecture (bowl-and-stem а, ascending б, descending р). Every combining mark receives computed GPOS mark-to-base anchors (the macron of ā attaches 251 units lower than that of Ā; the retroflex hook of ṛ lands on the letter's stem), and the ligature stratum is a GSUB `liga` feature — the orthography's rules are the font's code, verified by HarfBuzz shaping tests in the build toolchain — a concrete instance of Sproat's (2000) regularity hypothesis: the mapping from linguistic representation to written form is a regular relation, computable by finite-state means.

The linguistic point of this engineering is falsifiability: an orthographic rule that exists as a GSUB lookup either fires or does not, so the specification's claims about fusion and attachment are machine-checkable properties of an artifact rather than intentions in prose. The manuscript the reader is holding sets all object-language material in this typeface.

Прекарахме ўйкєнда в планината и казахме çäнкс.
Прекарахме *weekend*а в планината и казахме *thanks*.
Прекарахме уикенда в планината и казахме тенкс.

Figure 1: The register in context, at reading size, in a Bulgarian sentence ("We spent the weekend in the mountains and said thanks"). Top: the foreign words set in the Chuzhditsa face inline in the Cyrillic body face — the proposed register. Middle: the current practice of Latin italics. Bottom: plain recipient-adapting transcription, which marks nothing. The bespoke face marks more distinctively than italics and, unlike them, preserves the source contrasts (ў /w/, ç /θ/, ä /æ/, н /ŋ/) — at the cost, stated in §11, of depending on font control. Whether readers parse the top line as "foreign" rather than "archaic" is the open perception question.

9 Evaluation: the twenty-language audit

Method. The audit set is not a demographic top-twenty: it is a high-speaker-count convenience sample adjusted for phonological diversity and Slavic relevance, comprising English, Mandarin, Hindi–Urdu, Spanish, French, Arabic (MSA), Bengali, Portuguese, Russian, Japanese, German, Korean, Vietnamese, Turkish, Italian, Persian, Swahili, Greek, Polish, and Indonesian; clicks were excluded by scope. Each language is audited under one declared reference variety (General American English; Standard Mandarin; Delhi Hindi; Castilian Spanish; Parisian French; MSA; Brazilian Portuguese; Seoul Korean; Hanoi Vietnamese; and the respective standards elsewhere), with inventories taken from the language illustrations of the *Handbook of the IPA* (International Phonetic Association 1999) and standard language descriptions (for Slavic, Comrie & Corbett 1993).

Every phoneme of each inventory is classified three ways: *exact* (a rendering that preserves all of that language's contrasts involving the phoneme), *conventional merger* (a declared many-to-one mapping, enumerated below), or *declared loss* (prosodic folding, plus the one segmental length-undertone case). A rendering that depends on position (/w/, /ɲ/) counts as exact when the positional rule is stated. The system is deterministic relative to this input: variety choice is a parameter of the audit, not of the mapping. Against existing practice the relevant baselines are the national transcription traditions — Bulgarian's prescriptive system (Данчев 1982) and its analogues — which are recipient-adapting by design; Chuzhditsa differs precisely in preserving source contrasts those systems merge, which is what the citation-register framing predicts.

One boundary of this evaluation must be stated plainly: it is an *engineering coverage audit*, internally scored. The author of the mappings selected the reference varieties, declared the acceptable mergers, assembled the examples, and judged exactness; the machine-readable dataset makes every judgment inspectable and refutable, but no independent phonologist has yet re-scored it, no blinded transcribers have been tested for convergence, and no unseen language has been audited that was not available while the system was being refined (the Albanian application is a stress test, not external validation).

The sample is moreover biased toward segmental inventories: complex lexical tone (only Mandarin and Vietnamese), ATR and vowel-harmony systems, contrastive phonation, and ejectives are under-represented or routed to loss, so the exact-rate is best read as an upper bound. And because the author defines which mappings count as conventional mergers, the audit scores representational capacity against the declared rules, not adequacy against an external standard — functional load of the merged contrast, minimal-pair recoverability, or blind-transcriber convergence, the appropriate next instruments. What the audit establishes is representational capacity under the declared rules — not communicative success, which is the next experiment's question.

The audit's full 125-item example set is printed as Appendix A and accompanies the repository in machine-readable form (`data/audit_20_languages.json`/`.csv`, generated from the same source table as the appendix and the per-language guides, so the three cannot drift), together with the phoneme-

to-grapheme mapping it instantiates (`data/chuzhditsa.json`).

Table 4 summarizes the audit per language: the reference variety, the extension letters the language actually exercises, the conventional mergers engaged, and the declared losses. The pattern worth reading off is that no language needs more than a small subset of the inventory, that the mergers cluster on rhotics and sibilant places, and that declared loss is confined to prosody and the one length-under-tone case.

Language	Reference variety	Extensions exercised	Mergers engaged	Declared losses
English	General American	ў ç ʒ ä ɳ ӣ ȳ ь ʰ ɻ	rhotic → ɹ	—
Mandarin	Standard (Beijing)	ɳ ь ◌ʰ tones	/tʂ tɕ/→ч; /ʐ/→ж	sandhi; T1 = neutral
Hindi–Urdu	Delhi	ʈ ɖ ɳ ɹ ◌ʰ ◌ʰ ӣ ь ʰ ɻ	/ɦ/→ʰ	—
Spanish	Castilian	ç ʒ ꞟ ў	/β/→в; rhotics	—
French	Parisian	ɤ ʌ ɥʌ ɑ̣ ǫ̣ ö ÿ ў ь	/ʁ/→ɹ; /ɛ ɔ œ/→є о ö; /ɥ/→ÿ	—
Arabic	MSA	ç ʒ ɤ ʰ ɤ ꞟ ɻ ˈ ◌ ¯	/χ/→ɤ; emphatics→plain	—
Bengali	Standard	ʈ ɖ ɹ ɳ ä ◌ʰ ◌ʰ	/ɔ/→о	—
Portuguese	Brazilian	ɑ̣ ɤ ę ɥ ɣ	/ʁ/→ɹ; /ɛ ɔ/→є о	—
Russian	Standard	ы (native)	/ɕː/→шч	reduction (phonemic input)
Japanese	Tokyo	◌ ¯ ў ɻ	/ɸ/→ф; /ɯ/→ɣ; /ʑ/→ɻ; /ɴ/→н	pitch accent
German	Standard	ö ÿ ʰ ɳ ◌ ¯ + хь	/ɛː//eː/→ē	—
Korean	Seoul	◌ʰ ы ь ʰ	tense = geminate	—
Vietnamese	Hanoi	ы ь ʰ ɳ ꞟ ◌ ¯ tones	/ɓ ɗ/→б д	6 tones→4; length/tone
Turkish	Istanbul	ö ÿ ы ɻ ◌ ¯	/ɫ/→л	—
Italian	Standard	ꞩ	/ɛ ɔ/→є о	—
Persian	Tehran	ɤ ʰ ä ◌ ¯	/ɡ/→ɤ	/ʔ/ dropped
Swahili	Standard	ç ʒ ꞟ ɳ ʰ	implosives → б д	—
Greek	Standard Modern	ç ʒ ꞟ	—	—
Polish	Standard	ɤ ʌ ɥʌ ў ы	/ɕ/→сь; /ʂ/→ш; ͡tʂ/͡tɕ/→ч	—
Indonesian	Standard	ɳ ь ʰ ɻ	—	—

Table 4: Per-language audit summary. "Extensions exercised" lists the letters and marks the language's inventory actually requires; palatalization via ь and length via ◌ ¯ are counted where load-bearing. Mergers and losses are the declared ones of §9.

Table 5 samples the audit; the full 125-item set is Appendix A. Declared mergers are: all rhotics → ɹ; /ʂ//ʃ/ → ш; /ɸ//f/ → ф; /ɦ//h/ → ʰ; implosives → б д; /q//ɡ/ → ɤ (so Persian ق and ڭ, one phoneme in the Tehran reference, both → ɤ); the pharyngealized emphatics /tˤ dˤ sˤ ðˤ/ → plain т д с ʒ (Arabic); the open/close mid vowels /e//ɛ/ → є, /o//ɔ/ → о, and /ø//œ/ → ö (Romance, Bengali); Korean tense (fortis) consonants written as geminates (one analysis of the contrast); /x//χ/ → ɤ; the affricates /tʃ tʂ tɕ/ → ч and /dʒ dʐ dʑ/ → ɻ, with /ʒ//ʐ/ → ж (so Polish *cz*/*ć* and Mandarin's retroflex/palatal series each fall together); /v//β/ → в; /l//ɫ/ → л; /y//ʏ/ → ÿ; French /ɥ/ → ÿ (the glide written with its vowel counterpart, licensed by the systematic /y/∼[ɥ] alternation); Japanese /ɯ/ ([ɯ̟ᵝ], compressed-rounded) → ɣ by convention, where Korean /ɯ/ → ы; German /ɛː//eː/ → ē; English input referenced to a rhotic accent (GA), with tense /i//u/ written ӣ/ȳ although General American has no phonemic length — a transcription convention imported for reader familiarity, flagged rather than hidden. /ʕ/ is written (ˈ); /ʔ/ is written with the apostrophe ʼ in Arabic contexts and dropped elsewhere, as the shipped mapping (`data/chuzhditsa.json`) specifies. Declared losses are prosodic, with two segmental cases: Mandarin Tone 1 and the neutral tone are both unmarked and so coincide; six-tone systems fold into four marks (Vietnamese *hỏi*/*ngã* merge on the caron, and *nặng* falls together with *huyền* on the grave); Vietnamese phonemic vowel length is neutralized wherever a tone mark is present (length drops under tone); and /ʔ/ is dropped outside Arabic contexts (Persian). A rendering that engages none of these is *exact* in the Method's sense.

Source		IPA	Chuzhditsa
English	*thanks*	[θæŋks]	ç äҥкс
English	*weekend*	[ˈwiːkɛnd]	ўйкєнд
Mandarin	北京 *Běijīng*	[peɪ.tɕiŋ], T3+T1	Пĕйчин
Hindi	ठीक *ṭhīk*	[tʰiːk]	ţʰйк
Arabic	محمد *Muḥammad*	[muˈħammad]	Мухаммад
French	*croissant*	[kʁwasɑ̃]	крўасą
French	*bien*	[bjɛ̃]	бьѧ
Polish	*Wrocław*	[ˈvr̥ɔ̃tswaf]	Вроцўаф
Japanese	東京 *Tōkyō*	[toːkʲoː]	Тōкьō
Korean	한글 *hangul*	[hangɯl]	Ҥангыл
Georgian	წყალი *c'qali*	[ʧʼqʼali]	цҟǀали
Albanian	*Shqipëria*	[ʃcipəˈria]	Шкьипърйа

Table 5: Sampled audit items, plus two out-of-sample demonstrations (the Georgian ejectives and the Albanian stress test); the Chuzhditsa column is set in the typeface under discussion, with live GSUB fusion and GPOS anchoring. In the Mandarin row the caron on ĕ is tone 3, attached to the first syllable's vowel; tone 1 is unmarked by rule.

10 Adoption: lessons from planned and organic systems

A designed register lives or dies socially, not technically; the sociolinguistics of script choice behaves like the sociolinguistics of language choice (Fishman 1977; Unseth 2005). Four historical cases bound the space Chuzhditsa enters. The comparison is organized by recurrent factors rather than narrative sympathy: community ownership, institutional sponsorship, literacy burden, identity payoff, compatibility with existing orthographic practice, domain of use, prestige and stigma, and the mundane infrastructure of input methods and type.

The planned failures. Volapük collapsed within a decade, less on linguistic defects than on its inventor's proprietary control of the standard (Garvía 2015). Esperanto — maximally learnable, energetically organized — plateaued as a voluntary subculture: it extended nobody's existing practice and attached to no prior community, so adopting it meant *joining a new identity* rather than exercising an old one (Forster 1982; Garvía 2015).

The Deseret alphabet is the sharpest lesson, because it controls for institutional power: commissioned by Brigham Young, taught in classes from church-printed primers, backed by church funds, it was abandoned within roughly two decades — sponsorship without a felt need or an identity payoff bought nothing (Alder, Goodfellow & Watt 1984). Learnability and patronage, the two levers a designer can pull, decided none of these cases.

Nearer to Chuzhditsa than any whole-language project is the Initial Teaching Alphabet, a transitional English orthography whose central difficulty — treated at length by its own champions (Pitman & St John 1969) — was the dual system and the transition back to standard spelling, and which was largely abandoned in the following decade: the same leave-the-register burden a citation register asks its users to carry when a loan naturalizes.

The designed success. Hangul is the one engineered script that won, and its path is instructive precisely because the design was not what won. Completed in 1443 and promulgated in 1446 (Ledyard 1998), it was opposed at once by the literati — the collective Ch'oe Malli memorial of 1444 — and spent four and a half centuries outside prestige writing, surviving in low-stakes registers: women's correspondence, vernacular fiction, private devotion (Lee & Ramsey 2011). What revalued it was not its featural elegance but nineteenth-century nationalism, culminating in the 1894 Kabo edict that made the national script the base of official documents (King 1998). The mandate arrived 448 years after the design; the niche did the surviving.

The organic success. Katakana never asked anyone to change anything. It began as monks' ab-

breviated annotation glosses on Chinese texts in the early Heian period, drifted into a division of labor with hiragana over a millennium, and was codified only in 1900 by the implementing regulations of the Elementary School Order — with the loanword-register specialization consolidating later still (Seeley 1991; Twine 1991). Codification ratified practice; it did not create it.

The pattern reduces to four conditions, against which Chuzhditsa can be scored honestly. (i) *Additive, not substitutive*: the winners changed no existing spelling; Chuzhditsa alters nothing in any native orthography by construction. (ii) *Identity attachment*: Hangul survived on Korean-ness, not on design merit; Chuzhditsa's revivals — the yuses, the Glagolitic letterforms — are the analogous move available to a designed system, heritage rather than invention. (iii) *A survivable low-stakes niche before any mandate*: Hangul had the vernacular genres, katakana the annotation margin; the dictionary pronunciation field, the language classroom, and the subtitle are proposed as the equivalent low-stakes niche. (iv) *Codification last*: every mandate that worked ratified existing practice, which is why this specification petitions no orthographic authority.

Four cases cannot license induction, and none of this predicts uptake; the comparison identifies necessary-looking conditions, not sufficient ones. One condition, moreover, cannot be engineered around: on the planned/organic axis Chuzhditsa sits with Esperanto and Deseret, not with katakana. Assembly from attested material is a mitigation — every letter arrives with a history already attached — but it is not an exemption, and the Deseret case fixes an upper bound on what design quality plus sponsorship can buy without a community that recognizes itself in the script.

11 Limitations

Register versus encoding. The register's marking lives at two levels that must not be conflated. The phonological extensions are plain-text-safe: they are standard Unicode letters that survive copy, paste, search, and font substitution. The *visual* register, however, is carried by typeface choice — a diaphasic diglyphia in Bunčić's terms, not an encoded script contrast — and it degrades exactly where font control degrades: plain-text channels keep the letters but lose the look. The register is moreover invisible to non-visual access — a screen reader or a plain-text client receives the phonological letters but no foreignness signal at all, a cost of marking by font rather than by inventory. Whether readers actually perceive the Glagolitic-derived forms as "foreign register" rather than "archaic," "ecclesiastical," or merely "a display font" is an empirical question this paper does not answer; a perception study — mixed native/Chuzhditsa text, register-identification and readability tasks with Cyrillic readers — is the single most valuable next experiment.

Input assumptions. The mapping consumes a source pronunciation in a declared reference variety. Dialect choice, phonemic versus phonetic input, and spelling-influenced variant pronunciations are parameters the user supplies, not problems the system solves; learnability of the extension letters, and the cost of input methods for them, are likewise open. The keyboard prototype in the repository addresses input mechanics but not habit.

System-internal residue. The place hook is polysemous by direction — uvular on к, pharyngeal on х, *fronting* to interdental on с з, retroflex on т д р — so "retraction" would misname it; the mark encodes "off the native grid," not one direction. No attested language contrasts two of its values on one base letter, so the ambiguity is theoretical, but it is a real weakening of the one-mark-one-meaning claim. /ɕ/ keeps two spellings (сь under contrast, ш by entrenchment) — the one point where usage beat system. Six-tone prosodies fold lossily into four marks. Click phonology is future work, as is a mark-to-mark (`mkmk`) positioning pass for the rare double-stacked vowel.

12 Conclusion

What this paper demonstrates is deliberately narrower than what it proposes. It demonstrates that a coherent, technically implementable candidate for a katakana-like citation register can be *constructed* on

an alphabet without inventing a parallel inventory: a compositional grammar over attested Cyrillic material covers twenty phonologies under internally scored audit, one historical script supplies the register aesthetics, and the whole specification compiles to a font whose OpenType rules *are* the orthography. Whether the register *communicates* — whether readers recover pronunciations, transcribers converge, and the typeface signals "foreign citation" rather than "archaic" — is not demonstrated here; those are the claims the perception study of §11 exists to test, and until it is run this paper should be read as a design proposition with an inspectable artifact, not as evidence of communicative success. The generalization from Bulgarian to the Slavic superset cost one letterform (e) and no grammar changes, which we take as evidence that the operation layer, not the letter list, is the right unit of design.

The contribution we would defend most broadly is the vertical integration. A phonological distinction here is traceable, without a gap, through a graphemic operation, a Unicode sequence, a glyph construction, an OpenType behavior, and a rendered artifact — and each level is testable at its own altitude: mappings against inventories, spellings against shaping, constructions against binaries. A companion paper documents the construction level; the same end-to-end coherence is what makes every claim in this paper falsifiable by mechanical means, and it is a property worth wanting in any designed writing system, whatever the fate of this one.

References

[1] Alder, D. D., P. J. Goodfellow & R. G. Watt. 1984. Creating a new alphabet for Zion: The origin of the Deseret alphabet. *Utah Historical Quarterly* 52(3). 275–286.

[2] Bell, A. M. 1867. *Visible Speech: The Science of Universal Alphabetics*. London: Simpkin, Marshall.

[3] Bunčić, D., S. L. Lippert & A. Rabus (eds.). 2016. *Biscriptality: A Sociolinguistic Typology*. Heidelberg: Universitätsverlag Winter.

[4] Cahill, M. & K. Rice (eds.). 2014. *Developing Orthographies for Unwritten Languages*. Dallas, TX: SIL International.

[5] Comrie, B. & G. G. Corbett (eds.). 1993. *The Slavonic Languages*. London: Routledge.

[6] Comrie, B. 1996. Adaptations of the Cyrillic alphabet. In P. Daniels & W. Bright (eds.), *The World's Writing Systems*, 700–726. New York: OUP.

[7] Coulmas, F. 2003. *Writing Systems: An Introduction to their Linguistic Analysis*. Cambridge: CUP.

[8] Cubberley, P. 1996. The Slavic alphabets. In P. Daniels & W. Bright (eds.), *The World's Writing Systems*, 346–355. New York: OUP.

[9] Данчев, А. 1982. *Българска транскрипция на английски имена*. 2. изд. София: Народна просвета.

[10] Fishman, J. A. (ed.). 1977. *Advances in the Creation and Revision of Writing Systems*. The Hague: Mouton.

[11] Forster, P. G. 1982. *The Esperanto Movement*. The Hague: Mouton.

[12] Garvía, R. 2015. *Esperanto and Its Rivals: The Struggle for an International Language*. Philadelphia: University of Pennsylvania Press.

[13] International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: CUP.

[14] Irwin, M. 2011. *Loanwords in Japanese*. Amsterdam/Philadelphia: John Benjamins.

[15] Kang, Y. 2011. Loanword phonology. In M. van Oostendorp, C. J. Ewen, E. Hume & K. Rice (eds.), *The Blackwell Companion to Phonology*, vol. 4, 2258–2282. Malden, MA: Wiley-Blackwell.

[16] Kim-Renaud, Y.-K. (ed.). 1997. *The Korean Alphabet: Its History and Structure*. Honolulu: University of Hawai‘i Press.

[17] King, R. 1998. Nationalism and language reform in Korea: The questione della lingua in precolonial Korea. In H. I. Pai & T. R. Tangherlini (eds.), *Nationalism and the Construction of Korean Identity*, 33–72. Berkeley: Institute of East Asian Studies.

[18] Ledyard, G. K. 1998. *The Korean Language Reform of 1446*. Seoul: Sin'gu Munhwasa.

[19] Lee, K.-M. & S. R. Ramsey. 2011. *A History of the Korean Language*. Cambridge: CUP.

[20] Merunka, V. 2018. *Interslavic Zonal Constructed Language: An Introduction for English-Speakers*. Prague: Slovanská unie.

[21] Parčić, D. A. (ed.). 1905. *Rimski misal slověnskim jezikom* (Missale Romanum slavonico idiomate), 3rd edn. Rome: Typis S. Congregationis de Propaganda Fide.

[22] Peperkamp, S. & E. Dupoux. 2003. Reinterpreting loanword adaptations: The role of perception. *Proceedings of the 15th International Congress of Phonetic Sciences*, 367–370.

[23] Pitman, J. & J. St John. 1969. *Alphabets and Reading: The Initial Teaching Alphabet*. London: Pitman.

[24] Robertson, W. C. 2020. *Scripting Japan: Orthography, Variation, and the Creation of Meaning in Written Japanese*. London: Routledge.

[25] Sampson, G. 1985. *Writing Systems: A Linguistic Introduction*. Stanford: Stanford University Press.

[26] Schenker, A. M. 1995. *The Dawn of Slavic: An Introduction to Slavic Philology*. New Haven: Yale University Press.

[27] Sebba, M. 2007. *Spelling and Society: The Culture and Politics of Orthography around the World*. Cambridge: CUP.

[28] Seeley, C. 1991. *A History of Writing in Japan*. Leiden: Brill.

[29] Sproat, R. 2000. *A Computational Theory of Writing Systems*. Cambridge: CUP.

[30] Stanlaw, J. 2004. *Japanese English: Language and Culture Contact*. Hong Kong: Hong Kong University Press.

[31] Twine, N. 1991. *Language and the Modern State: The Reform of Written Japanese*. London: Routledge.

[32] The Unicode Consortium. 2024. *The Unicode Standard*, Version 16.0.0. South San Francisco, CA: The Unicode Consortium.

[33] Unseth, P. 2005. Sociolinguistic parallels between choosing scripts and languages. *Written Language & Literacy* 8(1). 19–42.

[34] Vajs, J. (ed.). 1927. *Rimski misal slověnskim jezikom* (Missale Romanum slavonico idiomate). Rome: Typis Polyglottis Vaticanis.

A The twenty-language audit: full example set

The complete audit set behind Table 4 and Table 5: 125 examples across the twenty audit languages, each with source form, source-variety IPA, Chuzhditsa rendering (set live in the typeface, with GSUB fusion and GPOS anchoring), and the extension letters the rendering exercises noted where instructive. Language headers restate the declared reference variety and the declared mergers and losses of §9; every rendering not engaging a declared merger or loss is *exact* in the sense of the Method paragraph of §9. This appendix is generated from the same machine-readable table (`data/audit_20_languages.json`) distributed with the fonts.

Source	IPA	Chuzhditsa	Notes
English (General American) — mergers: rhotic → p			
weekend	/ˈwiːkɛnd/	ўйкєнд	ў /w/, й long /iː/
thanks	/θæŋks/	çäнкс	ç /θ/, ä /æ/, н /ŋ/ — three new letters
this	/ðɪs/	ȝис	ȝ /ð/
thriller	/ˈθrɪlɚ/	çрилър	ç + ъ for /ə/
birthday	/ˈbɝθdeɪ/	бърçдєй	ъ /ɝ/, ç /θ/
world	/wɝld/	ўърлд	ў + ър
squirrel	/ˈskwɝəl/	скўъръл	ў in a cluster
Harry	/ˈhæri/	Ňäри	h /h/, ä /æ/
ring	/rɪŋ/	рин	н — no fake г
cat · cut · cart	/kæt/ /kʌt/ /kɑrt/	кät · кът · карт	minimal triple preserved
Mandarin (Standard (Beijing)) — mergers: tʂ tɕ → ч; ʐ → ж; declared loss: tone sandhi; T1 = neutral tone			
北京 Běijīng	/pèɪ.tɕíŋ/	Пĕйчин	unaspirated b/j are voiceless → п/ч; tone 3 ˇ, н
茶 chá	/tʂʰǎ/	чʰá	aspiration ʰ + tone 2
谢谢 xièxie	/ɕjê.ɕje/	сьѐсьє	ɕ → сь
人 rén	/ʐǎn/	жѣ́н	ʐ → ж, ъ /ə/
上海 Shànghǎi	/ʂâŋ.xàɪ/	Шàнхǎй	ʂ → ш, х /x/
猫 māo	/máʊ/	мао	tone 1 = unmarked
气 qì	/tɕʰî/	чʰѝ	aspirated affricate + tone 4
Hindi-Urdu (Delhi) — mergers: ɦ → h			
भारत Bhārat	/ˈbʰaːrət̪/	Бʰāрът	ʰ breathy voice, ā length, ъ /ə/
धन्यवाद dhanyavād	/d̪ʰənjəˈʊaːd̪/	Дʰънйъвāд	дʰ, two schwas → ъ
ठीक ṭhīk	/ʈʰiːk/	Ţʰйк	retroflex ţ + aspiration + length
दिल्ली Dillī	/d̪ɪlliː/	Дилли	dental д, no hook; geminate
घर ghar	/gʰər/	гʰър	гʰ breathy г
झील jhīl	/dʒʰiːl/	ц̣ʰйл	ц̣ʰ — voiced affricate with breathiness
चाय chāy	/tʃaːj/	чāй	ч
पानी pānī	/paːniː/	пāнй	double length
लड़का laṛkā	/ləɽkaː/	лъркā	ṛ /ɽ/ retroflex flap
हवा havā	/ɦəˈʊaː/	Ňъвā	h /ɦ/ (merges with /h/)
Spanish (Castilian) — mergers: β → в; rhotics → p			
cerveza	/θerˈβeθa/	çєрвєçа	Castilian /θ/ → ç
jamón	/xaˈmon/	хамон	j = /x/ → native x
agua	/ˈaɣwa/	аґўа	ґ /ɣ/ + ў /w/
niño	/ˈniɲo/	ниньо	ɲ → нь (native mechanism)

Source	IPA	Chuzhditsa	Notes
Madrid	/maˈðrið/	Мазриз	/ð/ → з
paella	/paˈeʎa/	паєлья	ʎ → ль
French (Parisian) — mergers: ʁ → р; ɛ ɔ œ → е о ö			
bonjour	/bɔ̃ʒuʁ/	бѫжур	ѫ /ɔ̃/ — the resurrected yus
vin	/vɛ̃/	вѧ	ѧ /ɛ̃/
blanc	/blɑ̃/	блą	ą /ɑ̃/ — the yus tail, by rule
croissant	/kʁwasɑ̃/	крўасą	ў + ą
deux	/dø/	дö	ö /ø/
rue	/ʁy/	рў	ў /y/
parfum	/paʁfœ̃/	парфǫ̈	ǫ̈ /œ̃/ — umlaut + nasal, two marks
bien	/bjɛ̃/	бѩ	ѩ /jɛ̃/ — the iotated yus, a ligature
grenouille	/gʁənuj/	грънуй	schwa → ъ
Arabic (MSA) (Modern Standard Arabic) — mergers: χ → х; emphatics → plain			
هلال hilāl	/hiˈlaːl/	Һилāл	h /h/ — light h
محمد Muḥammad	/muˈħammad/	Муҳаммад	ҳ /ħ/ — pharyngeal h
خالد Khālid	/ˈxaːlid/	Хāлид	х /x/ — the native x: all three H's in three names
قطر Qaṭar	/ˈqɑtˤɑr/	Қатар	қ /q/ uvular
غزال ghazāl	/ɣaˈzaːl/	ғазāл	ғ /ɣ/
عمر ʿUmar	/ˈʕumar/	Ӏумар	ӏ /ʕ/ ayn
شكرا shukran	/ˈʃukran/	шукран	native letters only
Bengali (Standard (Kolkata)) — mergers: ɔ → о			
ঢাকা Ḍhākā	/ɖʱaka/	Д̣ʰака	retroflex + breathy voice at once
ঠাকুর Ṭhākur	/ʈʰakur/	Т̣ʰакур	Tagore, spelled as he sounds
কলকাতা Kolkātā	/kolkaṯa/	Колката	no new letters
বড় bôṛo	/bɔɽo/	боро	ṛ /ɽ/
মিষ্টি mishṭi	/miʃʈi/	мишт̣и	ш + т̣ in a cluster
ভাই bhāi	/bʱai/	бʰай	бʱ
মাছ māchh	/mātɕʰ/	мачʰ	final aspiration
ধন্যবাদ dhonnobād	/d̪ʱɔnnobad̪/	дʰоннобад	дʱ + geminate
Portuguese (Brazilian) — mergers: ʁ → р; ɛ ɔ → е о			
São Paulo	/sɐ̃w̃ ˈpawlu/	Сąу Паулу	ą nasal + y semivowel
pão	/pɐ̃w̃/	пąу	nasal diphthong
coração	/koraˈsɐ̃w̃/	корасąу	-ção → -cąy, systematically
obrigado	/obriˈgadu/	обригаду	no new letters
manhã	/meˈɲɐ̃/	маня̨	нь + nasal via я
Russian (Standard) — mergers: ɕː → шч; declared loss: reduction (phonemic input)			
мышь	/mɨɕ/	мыш	ы comes back home; no silent ь
Хрущёв	/xrʊˈɕːɐf/	Хрушчов	Rus. щ /ɕː/ ≠ Bul. щ /ʃt/
бабушка	/ˈbabʊɕkə/	бабушка	phonemic, not phonetic: reduction goes unwritten
ёлка	/ˈjəlkə/	йолка	ё → йо
Пётр	/pʲətr/	Пьотр	palatalization with ь
Japanese (Tokyo) — mergers: ɸ → ф; ɯ → у (conv.); ʑ → ц̣; ɴ → н; declared loss: pitch accent			
東京 Tōkyō	/toːkʲoː/	Тōкьō	double length ō
大阪 Ōsaka	/oːsaka/	Ōсака	initial long vowel

Source	IPA	Chuzhditsa	Notes
可愛い kawaii	/kaɯ̟ai:/	каўаӣ	ў + ӣ
富士 Fuji	/ɸɯʑi/	Фуцӣ	ɸ → ф, ʑ → ц (merged)
ラーメン rāmen	/ra:meɴ/	рāмен	chōonpu → the length bar
寿司 sushi	/sɯɕi/	суши	ɯ → у (merged)
German (Standard)			
München	/ˈmʏnçən/	Мўнхьън	ў /ʏ/ + хь /ç/ ich-Laut
Goethe	/ˈgø:tə/	Гӧте	long ō /ø:/
Zürich	/ˈtsy:ʁɪç/	Цӯрихь	ц + ӯ
Straße	/ˈʃtʁa:sə/	штрāсе	шт cluster — native to Bulgarian
Bach	/bax/	Бах	ach-Laut = native x
schön	/ʃø:n/	шӧн	ō
Korean (Seoul) — mergers: tense = geminate			
서울 Seoul	/sʌ.ul/	Съул	ㅓ /ʌ/ → ъ
김치 kimchi	/kimtɕʰi/	кимчʰи	aspirated чʰ
한글 hangul	/hangɯl/	Һангыл	h + ы /ɯ/
태권도 taekwondo	/tʰɛk̟ʷʌndo/	тʰеккўъндо	tʰ aspiration; tense k̟ʷ = кк
부산 Busan	/pusan/	Пусан	lax п
Vietnamese (Hanoi) — mergers: ɓ ɗ → б д; declared loss: 6 tones → 4 marks			
phở	/fəː˧˩˧/	фъ̆	ъ /əː/ + tone ̆ (length drops under tone)
Việt Nam	/vjə̌t̚ naːm/	Вьѐт Нāм	ье
Hà Nội	/haː nôj/	Ҥà Нòй	h + tones
nước	/nɨ́ək/	ны́ък	ы /ɨ/ + rising tone
bánh mì	/ɓáɲ mì/	бáнь мѝ	implosive ɓ → б (merged)
Turkish (Istanbul) — mergers: ɫ → л			
Atatürk	/aˈtatyɾc/	Ататўрк	ў /y/
ırmak	/ɯɾˈmak/	ырмак	ı /ɯ/ → ы
göl	/ɟøl/	гӧл	ö /ø/
dağ	/daː/	дā	ğ = pure length → ̄
teşekkür	/teʃekˈkyɾ/	тешеккўр	geminate + ў
İstanbul	/isˈtanbuɫ/	Истанбул	ɫ → л (merged)
Italian (Standard) — mergers: ɛ ɔ → е о; ʎʎ → ль			
pizza	/ˈpittsa/	пицца	the geminate is written double
zero	/ˈdzɛːro/	зеро	s /dz/ — alive in Macedonian, borrowed here
gnocchi	/ˈɲɔkki/	ньоккки	нь + кк
famiglia	/faˈmiʎʎa/	фамилья	ль — palatal geminates degeminate (declared)
spaghetti	/spaˈgetti/	спагетти	тт
Persian (Tehran) — mergers: ɢ → ӄ			
تهران Tehrān	/tehˈrɒːn/	Теһрāн	h between vowels
قم Qom	/ɢom/	Ӄом	ӄ /ɢ/
خدا khodā	/xoˈdɒː/	ходā	х /x/
شیراز Shirāz	/ʃiːˈrɒːz/	Шӣрāз	ӣ + ā
دوغ dugh	/duːɢ/	дӯӄ	ġ/ق merged in Tehran → ӄ
Swahili (Standard) — mergers: implosives → б д			
dhahabu	/ðaˈhabu/	ʒаһабу	ʒ + h (Arabic loans in Swahili)

Source	IPA	Chuzhditsa	Notes
thelathini	/θelaˈθini/	ҫелаҫини	double ҫ
ng'ombe	/ŋombe/	ӈомбе	initial ӈ — impossible today
ghali	/ɣali/	ғали	ғ
Kenya	/keɲa/	Кеня	/ɲa/ = ня, the native mechanism
Greek (Standard Modern)			
Θεσσαλονίκη	/θesaloˈnici/	Ҫесалоники	θ → ҫ
δεν	/ðen/	ʒен	δ → ʒ
γάλα	/ˈɣala/	ғала	γ → ғ
ευχαριστώ	/efxariˈsto/	ефхаристо	χ = native х
χάος	/ˈxaos/	хаос	Greek becomes fully writable
Polish (Standard) — mergers: ɕ → сь; ʂ → ш			
Wrocław	/ˈvr͡ɔtswaf/	Вроцўаф	ł /w/ → ў
mąka	/ˈmɔ̃ka/	мѫка	ą is the old nasal: the yus meets its own descendant
pięć	/pʲɛ͡tɕ/	пѭч	ѭ — the font fuses ьѧ by itself
szczęście	/ʃ͡tʂɛ̃ɕ͡tɕɛ/	шчѧсьче	the famous cluster, tamed
Kraków	/ˈkrakuf/	Кракуф	final devoicing, honestly
Indonesian (Standard)			
orang	/ˈoraŋ/	ораӈ	ӈ
bahasa	/baˈhasa/	бахаса	h
terima kasih	/təˈrima ˈkasih/	търима касих	schwa → ъ, final h
nyanyi	/ɲaɲi/	няньи	нь
ngomong	/ŋɔmɔŋ/	ӈомоӈ	ӈ in two positions