Abbreviation across the world's languages and scripts

Kyle Gorman & Brian Roark
Google Research
kbg@google.com, roark@google.com

Abstract

Detailed taxonomies for *non-standard words*, including abbreviations, have been developed for speech and language processing, though mostly with reference to English. In this paper, we examine abbreviation formation strategies in a diverse sample of more than 50 languages, dialects and scripts. The resulting taxonomy—and data about which strategies are attested in which languages—provides key information needed to create multilingual systems for abbreviation expansion, an essential component for speech processing and text understanding.

Keywords: abbreviation, abbreviation expansion, text normalization, writing systems

1. Introduction

One of the oldest and most entrenched features of written language, dating back to the dawn of history, is the use of abbreviatory devices. Gorman et al. (2021) point out that nearly a third of the Latin words inscribed at the base of Trajan's Column in the Roman forum—completed early in the second century CE—are abbreviations, such as ⟨TRIB POT⟩ for *tribunicia potestate* 'power of the tribune'. Though abbreviation seems to be general feature of written language, most discussion of this property in the speech and language processing community has focused on English, and to a lesser degree other languages written with the Latin alphabet. Much less work has focused on abbreviation formation in other languages or writing systems. In this work we take first steps towards a more inclusive documentation of abbreviation in the world's languages and scripts.

Detailed—albeit dated—treatments of English abbreviation formation are provided by Marchand (1969, §9) and Cannon (1989). As Cannon's extensive bibliography attests, there is a long and robust literature documenting English acronyms.

More recently, speech and language processing specialists have developed data-driven *abbreviation expansion* engines for converting abbreviations to fully-expanded words, usually using nearby linguistic context to resolve ambiguities (e.g., Baldwin et al., 2015; Chrupała, 2014; Roark and Sproat, 2014; Gorman et al., 2021). Unfortunately, this literature is also dominated by work solely considering English.[1] Of course, key to these sorts of abbreviation expansion methods is the availability of data, and unfortunately there is little of that, even for the best resourced languages. Several of the above cited works make use of ad hoc methods to detect or synthesize abbreviations in context (Roark and Sproat, 2014; Żelasko, 2018), whereas others rely on social media data from shared tasks (e.g., Baldwin et al., 2015; Chrupała, 2014), data that is unfortunately no longer available. To the authors' knowledge, the data set from Gorman et al. (2021),[2] produced by asking annotators to shorten selected sentences from English-language Wikipedia, is the only large-scale abbreviation data set available to the public.

One limitation of previous computational work on abbreviation expansion—reflecting its narrow focus on a relatively small number of languages and scripts—is the simplifying assumption that all abbreviations are formed by deleting one or more characters (i.e., they are *deletion-based* in the sense of Pennell and Liu 2010). Formally, then, this assumes that abbreviations are proper subsequences of their corresponding full forms, with no further augmentations or changes to spelling. Furthermore, some of this prior work also focuses on abbreviations of a single word and ignores abbreviations formed from phrases. As will be seen below, neither of these assumptions is generally valid for the world's languages and scripts.

In this study, we describe the collection of a survey of abbreviatory mechanisms used in a diverse sample of just over 50 languages, dialects, and scripts. From the results of the survey, we provide a taxonomy of abbreviation expansion strategies, and provide information about which abbreviation strategies, if any, are attested in each of the surveyed languages, dialects, and scripts. We thus document the kinds of phenomena that can be found in this diverse sample. We conjecture that it will always be necessary to impose constraints on what abbreviatory strategies are considered, so this information is a prerequisite for building and validating the data collection and annotation processes needed to build truly multilingual abbrevia-

[1] One notable exception is Żelasko (2018), who studies the particular challenges of abbreviation expansion in Polish, a richly inflected language.

[2] `https://github.com/google-research-datasets/WikipediaAbbreviationData`

tion expansions systems. All survey data, including examples elicited from the language consultants, is publicly available under a Creative Commons CC-BY 4.0 license.[3]

2. Background

Abbreviations are a class of what Sproat et al. (2001) call *non-standard words*, forms found in written text which are generally not pronounced according to the ordinary letter-to-sound rules of the language. Non-standard words, henceforth NSWs, pose difficulties for speech and text processing applications. In particular, text-to-speech (TTS) systems require NSWs to be converted to "spoken form" (e.g., Ebden and Sproat, 2015), and automatic speech recognition (ASR) systems must invert this transduction so that the resulting transcripts can be displayed to users in a readable format (e.g., Pusateri et al., 2017).[4] Sproat et al. (2001) provide a taxonomy of NSWs, and this is further enriched by van Esch and Sproat (2017). These taxonomies provide useful information for the linguists and engineers who design the many computer systems that interact with NSWs. Sproat et al. propose three broad categories of abbreviation, focusing on the pronunciation of the NSW token: ASWDs: those read as a word (e.g., *NATO*); LSEQs: those read as a letter sequence (e.g., *CIA*); and EXPNs: general abbreviations (e.g., *Blvd.*). In this work, however, we propose a taxonomy that goes beyond mere pronunciation.

Gorman et al. (2021) discern two broad classes of abbreviations. First are *conventionalized* abbreviations, which are of high enough frequency that both their pronunciation and denotation are known to most readers, at least in certain speech communities or text genres. These include SI[5] units (e.g., *Hz* read as *Hertz*) and abbreviations for certain geographic entities (e.g., *OH* read as *Ohio*). The second type are *ad-hoc* abbreviations, those coined as needed. These are particularly common on digital communications channels which prefer brevity, such as text messaging (e.g., Crystal, 2001, 2008). These ad-hoc abbreviations are rarely recorded by lexicographers.

Like other NSWs, it is not always obvious how one might read an abbreviation. For instance, one might read *NATO*—a conventionalized abbreviation—as a word rhyming with *Cato*, or possibly expand it to its full form, the *North Atlantic Treaty Organization*, but it is not ordinarily read letter by letter. In contrast, *CIA* is an *initialism*, an abbreviation which is read letter by letter, but never as a two-syllable word rhyming with *Garcia*. Such language- and word-specific facts are crucial for building an ASR verbalizer or a TTS front-end.

Abbreviations, particularly ad-hoc abbreviations, pose an additional difficulty not common to other NSWs: it is not always obvious what they denote. For instance, *AMA* could represent: *American Medical Association*, a professional organization and lobbying group for doctors; *against medical advice*, jargon referring to a patient leaving the hospital before being cleared for discharge; or *ask me anything*, a prompt used on various online forum to solicit questions; and this does not exhaust the possibilities. Similarly, the ad-hoc abbreviation *brd* might denote *bread*, *broad*, or *bird*, and context is required to determine how to pronounce or interpret it. For this reason, methods for abbreviation expansion must take context into account.

3. Methods

We conducted a survey of abbreviatory methods used in diverse languages and scripts. Adapting pre-existing practices at Google for internationalizing text normalization systems, we hypothesized that literate, linguistically sophisticated language consultants would have reasonably reliable judgments about whether or not a particular type of abbreviation formation process is present in their language when provided with examples of that process. These examples were in English—the language used for the survey instrument itself—when relevant examples exist in English, and glosses were provided for examples from other languages when relevant examples do not exist in English. Table 1 lists the seven abbreviation strategies targeted and the examples provided; the full text of the survey can be found in Appendix A.

To strengthen conclusions about which types of abbreviations are present in which language, we also asked consultants to provide two or three examples of each process they claimed for their language.[6] Elicitation of examples allowed us to discern whether a consultant understood the strategies during the consensus procedure.

[3] https://github.com/google-research/google-research/tree/master/multilingual_abbreviation_survey

[4] Such transductions are traditionally referred to as *text normalization*, though this term has taken on a broader sense—since it was first popularized by Sproat et al., 2001—which includes the conversion of noisy user-generated text to conventional spelling for purposes unrelated to speech processing.

[5] 'International System of Units', whose abbreviation comes from the French *Système international d'unités*.

[6] We note in passing that requiring an additional step when answering "yes" might cause consultants to have a response bias in favor of "no". While the consensus procedure described in subsection 3.4 is intended to avoid errors due to this bias, we have no straightforward way to detect the presence of such a bias.

Abbreviation class	Example
First-character abbreviations	*NATO* (< *North Atlantic Treaty Organization*)
Stump compounds	*FiDi* (< *Financial District*)
Truncations	*Col.* (< *Colonel*)
Augmented truncations	Australian English *footie* (< *football* plus augment *-ie*)
Word-internal deletions	*Blvd.* (< *Boulevard*)
Inflection strategies	Spanish *EE UU* (< *Estados Unidos* 'United States')
Reduplication strategies	Indonesian *orang2* (< *orangorang* 'people')
Other strategies	(none given)

Table 1: Abbreviation strategies queried in the survey, with characteristic examples.

3.1. Languages sampled

The languages, dialects, and scripts were selected to cover many language families and script types but also in support of internationalization and quality assurance efforts at Google. We decided to focus in some cases on multiple dialects of a particular "macrolanguage", or to target the multiple scripts used to write a certain language. In a slight abuse of terminology, we refer to the entries in our survey—a language or dialect, and the associated script—as *locales*. See Table 3 for a full list of locales. Some details of how locales were defined are discussed below.

For Arabic, which is both diglossic and pluricentric, the sample targeted both the Modern Standard literary standard as well as Egyptian, Gulf, and Magrebi dialects. For Gulf and Magrebi dialects, we target abbreviations in non-standard romanization. Hindi-Urdu was treated as two separate locales, as Hindi is written with a Brahmic alphasyllabary, and Urdu with a Perso-Arabic consonantal alphabet. Separate locales were used for the Brahmic (*Gurmukhi*) and Perso-Arabic (*Shahmukhi*) scripts used to write Punjabi in India and Pakistan, respectively, and for the non-standard romanizations of Bengali, Hindi, Marathi, and Urdu. Finally, European and Brazilian dialects of Portuguese were treated as separate locales.

3.2. Participants

At least three—though occasionally as many as nine—consultants gave judgments and examples for each locale. Consultants were recruited from a pool of professional annotators and were compensated for their time.

3.3. Instrument

The survey itself was conducted using Google Forms. Before the survey began, consultants were prompted to rate their proficiency with the target locale on a seven-point scale where 1 was labeled "limited proficiency" and 7 was labeled "native fluency". The median value for this proficiency score was 5, and no consultant scored their proficiency lower than 3. The survey consists of seven main questions, each asking whether the target locale uses a particular style of abbreviation. These strategies are listed in Table 1.

If the consultant answered yes, they were then asked to provide two or three relevant examples of that style (where each example includes the abbreviated form, the expanded form, and an English gloss). This initial taxonomy of style is based on the authors' own linguistic background and is not intended to exhaust the possibilities. Thus, in a final question, the consultant was asked whether they were aware of any styles of abbreviation not yet covered in the locale, and if so, were asked to provide examples of these styles. The text of the survey is reprinted in Appendix A.

3.4. Consensus

When consultants for a given locale disagreed as to whether a given abbreviation formation strategy was present in that locale, the first author manually enforced consensus across the responses for that locale. This was done by consulting the examples provided. If, for instance, only one of the three consultants provided an example of a given strategy, but the examples clearly illustrate the strategy, the omission by the other consultants was assumed to be accidental and the strategy is coded as present in the locale. However, if the examples provided were not of the relevant strategy, or were judged uninterpretable, they were discarded and the strategy was coded as absent.

4. Results

The survey received roughly 200 responses over 55 locales corresponding to 46 unique ISO 639-1 languages. Roughly half of these locales use non-Latin scripts. The associated language codes are adapted from ISO-639-1 with additional disambiguating information where necessary.

Disagreements between annotators were somewhat common, accounting for 39% of the re-

Abbreviation class	% attested
First-character abbreviations	90.6
Stump compounds	45.3
Truncations	56.6
Augmented truncations	30.2
Word-internal deletions	45.3
Inflection strategies	15.1
Reduplication strategies	3.8
Other strategies	17.0

Table 2: The percentage of locales attesting each of the eight abbreviation strategies.

sponses (not including "Other strategies") aggregated across locale. Table 2 provides the percentages of locales that attest (following the consensus procedure in subsection 3.4) each of our abbreviation classes. Table 3 in Appendix B provides the full post-consensus per-locale results. We give an impressionistic summary of the findings below.

As can be seen from Table 2, first-character abbreviations are present for most locales, but the other strategies in English are less commonly found. Truncations are the next most common strategy; these are attested in just over half of the locales. Inflection and reduplication marking are far less common than the other strategies. Roughly one out of six locales attest other strategies beyond the seven provided.

The locales for languages spoken in the Indian subcontinent—Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu, and Urdu—make little use of abbreviation beyond first-character abbreviations (e.g., Kannada ಬಿಎಸ್‌ಎನ್‌ಎಲ್ ⟨bi-es-en-el⟩ 'Bharat Sanchar Nigam Limited'; note the failure to encode aspiration of the initial consonant), and some of these may be borrowed from English. This is marked insofar as these languages belong to two distinct language families (Indo-Aryan and Dravidian) and which may be written in Brahmic alphasyllabaries, Perso-Arabic consonantal alphabets, or Latin alphabetic transliterations.

Outside of the subcontinent, two other languages which make limited use of abbreviations are Arabic and Farsi. While these languages are unrelated, this may reflect a structural incompatibility between abbreviation formation and the Perso-Arabic script they share (also with Urdu, as mentioned above); this script is *defective* in the sense that short vowels are ordinarily omitted, and this in turn might limit the ability—or need—to form abbreviations via truncation or word-internal deletion. Something similar might be true of Korean; its *Hangul* writing system is roughly alphabetic, but symbols are organized into syllable-sized blocks called *jamo*, which might make it difficult to indicate abbreviations—particularly those which involve deletion of vowels—orthographically. Similarly, Japanese, which is written in a mixed—but predominantly syllabic—writing system, forms novel stump compounds from English phrases, as in セクハラ ⟨se-ku-ha-ra⟩ 'sexual harassment', but its writing system lacks any obvious way to represent the deletion of vowels. Amharic and Tigrinya, closely related Ethiopic languages written with the Ge'ez alphasyllabary, also make limited used of abbreviation formation. Among Latin-script locales, Yorùbá and Zulu stand out for their limited use of abbreviations.

Relatively few locales surveyed make use of augmented truncations. Both dialects of Portuguese use *-ão* as an augment, as in *burgão* (< *hambúrguer* 'hamburger'). The Russian abbreviation презик 'condom' is formed from a truncation of презерватив suffixed with a -ик augment. Turkish uses truncation with an *-o* augment to form familiar forms of proper names. While it is not abbreviation per se, Indonesian uses truncation and infixation to derive informal forms of words. For example, *sepokat* is formed via truncation of word-final *u* in *sepatu* 'shoe' and infixation of *-ok-*.[7]

Inflection-marking strategies are overall rare. Belarusian, Polish, and Romanian generalize the character-doubling strategy found in Portuguese, Serbian and Spanish by reduplicating the entire abbreviation. In Polish, for example, the plural forms of *o.* (< *ojciec* 'father') and *prof.* (< *profesor* 'professor') are *oo.* and *prof. prof.*, respectively. In Slovenian, the title of someone with two doctoral degrees is abbreviated either as *ddr.* or *dr. dr.* Specific strategies for abbreviating reduplicated words are even less common. Indonesian and Vietnamese use the Arabic numeral *2* to indicate reduplication, and the 々 character serves the same purpose in Japanese.

Many consultants reported the presence of other strategies for abbreviation formation. Many locales report the use of a "mixed" strategies. For example, the derivation of Belarusian БелТА (< Беларускае Тэлеграфнае Агенцтва 'Belarusian Telegraph Agency'), Polish *PZMot* (< *Polski Związek Motorowy* 'Polish Automobile and Motorcycle Federation'), and Uzbek *SamDU* (< *Samarqand Davlat Universiteti* 'Samarkand State University') seems to combine truncation and first-character abbreviation. Bulgarian forms abbreviations by deleting a contiguous sequence of word-internal segments, marking the deleted span with a hyphen—e.g., у-ще (< училище 'school')—and a similar strategy is found in Hebrew.

[7] This process is apparently borrowed from a thieves' argot. A similar process, the infixation of *-iz-* in African-American Vernacular English, is described by Gil Scott-Heron in his spoken-word piece "The Ghetto Code".

5. Conclusions

We have presented (and publicly released) results from a survey over diverse languages and scripts regarding abbreviatory devices in the writing systems. Explicit examples in a language of specific attested abbreviation strategies can help in developing covering grammars or similar pattern-matching methods (e.g., Gorman and Sproat, 2016; Sproat and Jaitly, 2017; Zhang et al., 2019) to find possible abbreviations and candidate expansions in raw text, en route to building abbreviation expansion engines. In future work, we intend to mine web text to identify additional examples of attested patterns.

Acknowledgments

Thanks to Alexander Gutkin, Christo Kirov, and Richard Sproat; and to the many annotators upon whom this project depended.

References

Timothy Baldwin, Young-Bum Kim, Marie Catherine de Marneffe, Alan Ritter, Bo Han, and Wei Xu. 2015. Shared tasks of the 2015 Workshop on Noisy User-generated Text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–136.

Garland Cannon. 1989. Abbreviations and acronyms in English word-formation. *American Speech*, 64(2):99–127.

Grzegorz Chrupała. 2014. Normalizing tweets with edit scripts and recurrent neural embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 680–686.

David Crystal. 2001. *Language and the Internet*. Cambridge University Press.

David Crystal. 2008. *Txtng: The Gr8 Db8*. Oxford University Press.

Peter Ebden and Richard Sproat. 2015. The Kestrel TTS text normalization system. *Natural Language Engineering*, 21(3):1–21.

Daan van Esch and Richard Sproat. 2017. An expanded taxonomy of semiotic classes for text normalization. In *Proceedings of INTERSPEECH*, pages 4016–4020.

Kyle Gorman, Christo Kirov, Brian Roark, and Richard Sproat. 2021. Structured abbreviation expansion in context. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 995–1005.

Kyle Gorman and Richard Sproat. 2016. Minimally supervised number normalization. *Transactions of the Association for Computational Linguistics*, 4:507–519.

Hans Marchand. 1969. *The Categories and Types of Present-Day English Word-Formation*, 2nd edition. Beck.

Deana Pennell and Yang Liu. 2010. Normalization of text messages for text-to-speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4842–4845.

Ernest Pusateri, Bharat Ram Ambati, Elizabeth Brooks, Ondrej Platek, Donald McAllaster, and Venki Nagesha. 2017. A mostly data-driven approach to inverse text normalization. In *Proceedings of INTERSPEECH*, pages 2784–2788.

Brian Roark and Richard Sproat. 2014. Hippocratic abbreviation expansion. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 364–369.

Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333.

Richard Sproat and Navdeep Jaitly. 2017. An RNN model of text normalization. In *Proceedings of INTERSPEECH*, pages 754–758.

Hao Zhang, Richard Sproat, Axel H. Ng, Felix Stahlberg, Xiaochang Peng, Kyle Gorman, and Brian Roark. 2019. Neural models of text normalization for speech applications. *Computational Linguistics*, 45(2):293–337.

Piotr Żelasko. 2018. Expanding abbreviations in a strongly-inflected language: are morphosyntactic tags sufficient? In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1880–1884.

A. Survey questions

1. Does the target language use abbreviations formed from the first character of each word in a phrase (e.g., “NATO” < “North Atlantic

Treaty Organization", "CIA" < "Central Intelligence Agency")? [**If yes:** Please provide 2–3 examples of abbreviations formed from the first character of each word in a phrase in the target language, giving the abbreviated form, the full/expanded form, and an English gloss. (You are also welcome to link to any relevant discussions of this phenomenon.)]

2. Does the target language use abbreviations formed from the first syllable of each word in a phrase (e.g., "FiDi" < "Financial District", "ForEx" < "foreign exchange")? [**If yes:** Please provide 2–3 examples of abbreviations formed from the first syllable of each word in a phrase in the target language, giving the abbreviated form, the full/expanded form, and an English gloss. (You are also welcome to link to any relevant discussions of this phenomenon.)]

3. Does the target language use abbreviations formed by truncating characters at the ends of words (e.g., "Col." < "Colonel", "Ave." < "Avenue")? [**If yes:** Please provide 2–3 examples of abbreviations formed by truncating characters at the ends of words in the target language, giving the abbreviated form, the full/expanded form, and an English gloss. (You are also welcome to link to any relevant discussions of this phenomenon.)]

4. Does the target language use abbreviations formed by truncating characters at the ends of words and then adding "augment" suffixes (e.g., Australian English "footie" < "football", with an -ie augment, "ambo" < "ambulance" with an -o augment)? [**If yes:** Please provide 2–3 examples of abbreviations formed by truncating characters at the ends of words and then adding "augment" suffixes in the target language, giving the abbreviated form, the full/expanded form, and an English gloss. (You are also welcome to link to any relevant discussions of this phenomenon.)]

5. Does the target language use abbreviations formed by deleting characters (e.g., particularly vowels) from the middle of words (e.g., "Blvd." < "Boulevard", "Sgt." < "Sergeant")? [**If yes:** Please provide 2–3 examples of abbreviations formed by deleting characters from the middle of words in the target language, giving the abbreviated form, the full/expanded form, and an English gloss. (You are also welcome to link to any relevant discussions of this phenomenon.)]

6. Does the target language use any orthographic tricks to mark the inflection of abbreviations (e.g., Spanish "EE UU." < "Estados Unidos" 'United States', with the doubling used to indicate that the abbreviation is plural)? [**If yes:** Please provide 2–3 examples of inflected abbreviations in the target language, giving the abbreviated form, the full/expanded form, and an English gloss. (You are also welcome to link to any relevant discussions of this phenomenon.)]

7. Does the target language use any particular orthographic tricks to abbreviate reduplication (e.g., Indonesian "orang2" < "orangorang" 'people', with "2" used to indicate reduplication)? Answer "No" if the target language does not have productive reduplication. [**If yes:** Please provide 2–3 examples of abbreviations for reduplication in the target language, giving the abbreviated form, the full/expanded form, and an English gloss. (You are also welcome to link to any relevant discussions of this phenomenon.)]

8. Does the target language use any other style of abbreviation not yet covered? [**If yes:** Please provide 2–3 examples of these other style or styles, giving the abbreviated form, the full/expanded form, and an English gloss. (You are also welcome to link to discussions of other styles of abbreviation in the target language.)]

B. Consensus locale results

Table 3 presents the consensus results for each of the locales surveyed.

Code	Locale	*NATO*	*FiDi*	*Col.*	*footie*	*Blvd.*	*EE UU*	*orang2*	Other
am	Amharic	✓	✗	✗	✗	✗	✗	✗	✗
ar-eg	Arabic (Egyptian)	✓	✗	✗	✗	✗	✗	✗	✗
ar-gu	Arabic (Gulf, Latin)	✓	✗	✗	✗	✗	✗	✗	✗
ar-ml	Arabic (Magrebi, Latin)	✓	✗	✗	✗	✗	✗	✗	✗
ar-ms	Arabic (Modern Standard)	✓	✗	✗	✗	✗	✗	✗	✗
as	Assamese	✓	✗	✗	✗	✗	✗	✗	✗
be	Belarusian	✓	✓	✓	✗	✓	✓	✗	✓
bg	Bulgarian	✓	✓	✓	✗	✗	✗	✗	✗
bn	Bengali	✓	✓	✓	✗	✓	✗	✗	✗
bn-la	Bengali (Latin)	✓	✓	✓	✓	✓	✗	✗	✗
en	English	✓	✓	✓	✓	✓	✗	✗	✗
es	Spanish	✓	✗	✗	✗	✓	✓	✗	✓
et	Estonian	✓	✓	✓	✗	✓	✗	✗	✗
fa	Farsi	✓	✗	✗	✗	✗	✗	✗	✗
fr	French	✓	✓	✓	✗	✗	✗	✗	✓
gu	Gujarati	✓	✗	✗	✓	✗	✗	✗	✗
ha	Hausa	✓	✗	✓	✓	✓	✗	✗	✓
he	Hebrew	✓	✓	✓	✓	✗	✗	✗	✓
hi	Hindi	✓	✗	✗	✓	✗	✗	✗	✗
hi-la	Hindi (Latin)	✓	✗	✓	✓	✗	✗	✗	✗
hy	Armenian	✓	✓	✓	✓	✓	✗	✗	✓
id	Indonesian	✓	✓	✓	✓	✓	✗	✓	✓
it	Italian	✓	✗	✓	✓	✓	✗	✗	✗
ja	Japanese	✓	✗	✗	✗	✗	✗	✓	✗
ka	Georgian	✓	✓	✓	✓	✓	✗	✗	✗
kn	Kannada	✓	✗	✗	✗	✓	✗	✗	✗
ko	Korean	✗	✓	✗	✗	✗	✗	✗	✗
lt	Lithuanian	✓	✓	✗	✗	✓	✗	✗	✗
lv	Latvian	✓	✗	✓	✗	✓	✗	✗	✗
ml	Malayalam	✓	✗	✗	✗	✗	✗	✗	✗
mr	Marathi	✓	✗	✗	✗	✗	✗	✗	✗
mr-la	Marathi (Latin)	✗	✗	✗	✗	✗	✗	✗	✗
nl	Dutch	✓	✓	✓	✗	✓	✗	✗	✗
or	Odia	✓	✓	✗	✗	✗	✗	✗	✗
pa-gu	Punjabi (Gurmukhi)	✗	✗	✗	✗	✗	✗	✗	✗
pa-sh	Punjabi (Shahmukhi)	✓	✗	✗	✗	✗	✗	✗	✗
pl	Polish	✓	✓	✓	✗	✓	✓	✗	✓
pt-br	Portuguese (Brazilian)	✓	✓	✓	✓	✓	✗	✗	✗
pt-pt	Portuguese (European)	✓	✓	✓	✓	✓	✓	✗	✗
ro	Romanian	✓	✓	✓	✗	✗	✓	✗	✗
ru	Russian	✓	✓	✓	✓	✓	✗	✗	✓
sl	Slovenian	✓	✓	✓	✓	✓	✗	✗	✗
sq	Albanian	✓	✗	✓	✗	✓	✓	✗	✗
sr	Serbian	✓	✓	✓	✗	✗	✓	✗	✗
sw	Swahili	✓	✗	✓	✗	✓	✗	✗	✗
ta	Tamil	✓	✗	✓	✗	✗	✗	✗	✗
te	Telugu	✓	✗	✓	✗	✗	✗	✗	✗
ti	Tigrinya	✗	✗	✓	✗	✗	✗	✗	✗
tr	Turkish	✓	✓	✓	✓	✓	✗	✗	✗
ur	Urdu	✗	✗	✗	✗	✗	✗	✗	✗
ur-la	Urdu (Latin)	✓	✗	✗	✗	✗	✗	✗	✗
uz	Uzbek	✓	✓	✓	✗	✓	✗	✗	✗
vi	Vietnamese	✓	✓	✗	✓	✓	✓	✗	✗
yo	Yorùbá	✓	✗	✗	✗	✗	✗	✗	✗
zu	Zulu	✗	✗	✓	✗	✓	✗	✗	✗

Table 3: Summary of the abbreviation strategies attested in each locale after consensus.