LinguaMeta: Unified metadata for thousands of languages

Sandy Ritchie[1], Daan van Esch[1], Uche Okonkwo[1], Shikhar Vashishth[1], Emily Drummond[2]
Google Research[1], Intellipro Group[2]
1600 Amphitheatre Parkway, Mountain View, CA 94043, USA[1]
3120 Scott Blvd, Santa Clara, CA 95054, USA[2]
{sandyritchie,dvanesch,uokonkwo,shikharv,emilydrummond}@google.com

Abstract

We introduce LinguaMeta, a unified resource for language metadata for thousands of languages, including language codes, names, number of speakers, writing systems, countries, official status, and geographic coordinates. The resources are drawn from various existing repositories and supplemented with our own research. Each data point is tagged for its origin, allowing us to easily trace back to and improve existing resources with more up-to-date and complete metadata. The resource is intended for use by researchers and organizations who aim to extend technology to thousands of languages.

Keywords: metadata, prioritization, planning

1. Introduction

There are over 7,000 languages in use today.[1] However, technologies such as machine translation, speech recognition and text-to-speech synthesis are only widely available in about a hundred or so languages, while for most other languages, coverage is low-quality or non-existent (Blasi et al., 2022).

Massively multilingual language technologies such as Meta's Massively Multilingual Speech project (Pratap et al., 2023), and Google's Next Thousand Languages machine translation model (Bapna et al., 2022) and Gboard keyboard app (van Esch et al., 2019) have demonstrated that it is possible to extend speech, translation and text input capabilities to thousands of languages. To help underpin these developments with proper contextualization and analysis, we believe reliable open-source statistics and information should be available for all these languages.

In this paper we introduce LinguaMeta, a unified open-source repository of language metadata.[2] The aim of LinguaMeta is to offer reliable language metadata broadly, in the hope that it will help with various scoping, planning and tracking tasks required to extend language technology to more of the world's languages. For instance, reliable information on numbers of speakers makes it easier to understand how many potential users might benefit from a new technology for a given language, which helps with prioritizing which languages to work on first. Knowledge of the writing system(s) used for a language, and which countries it is spoken in, is also critical: it enables understanding what language-script-locale combinations need to be supported (e.g. 'pa-Guru-IN'[3]); and helps to define these language-locale definitions which form the basis of many language technologies and localization systems. Metadata can also provide important basic context about a language, which can then be enhanced by deeper engagements with speakers of the language: for example, community desires may differ between larger 'institutional' languages (like English or Mandarin) on the one hand, and languages spoken by a smaller community in a more local setting with extensive multilingualism on the other (Bird, 2022). Finally, language metadata also helps with surveying and cross-comparison of large multilingual projects, making it easier to analyze differences in coverage and quality. This can help communities and researchers to make more informed decisions about which technology, product or ecosystem is best suited to their needs.

Various existing resources, notably Ethnologue (Eberhard et al., 2024) and Glottolog (Hammarström et al., 2024), aim to provide comprehensive metadata for all of the world's languages. However, Glottolog does not provide metadata on writing systems or speaker counts, and Ethnologue is not accessible without a paid subscription. van Esch et al. (2022) released writing systems,

[1]This paper focuses on spoken languages. Metadata for signed languages is also available in some of the repositories discussed in this paper, but we do not include them in the current version of LinguaMeta, as more research is needed on the deployment of language technologies for signed languages.

[2]LinguaMeta is available at https://github.com/google-research/url-nlp/tree/main/language_metadata/linguameta.

[3]Punjabi written in Gurmukhi and spoken in India.

LREC-COLING 2024, pages 10530–10538
20-25 May, 2024.

Table 1: LinguaMeta metadata categories and examples for Romanian.

Metadata type	Source(s)	Example for Romanian
ISO 639-3 code	ISO 639	ron
BCP-47 code	IETF	ro
ISO 639-2b code	ISO 639	rum
Deprecated BCP-47 codes	IETF	mo
Glottocode	Glottolog	roma1327
Wikidata code	Glottolog	Q7913
English name	CLDR, Glottolog, Google, IETF, ISO 639, Wikidata, Wiktionary	Romanian
Endonym	CLDR, Glottolog, Wikidata, Google	română
Names in other languages	CLDR, Glottolog, Google, Wikidata	roumain [fr], Rumänisch [de], román nyelv [hu]...
Estimated number of speakers	CLDR, Google, Wikipedia	21,100,000
Writing system(s)	Google, Wikidata, Wiktionary, assumed by locale, GlotScript*	Latin
Locale(s)	Glottolog, Google	Romania, Moldova
Regions	n/a, derived from locales	region: Europe subregion: Eastern Europe
Coordinates	Glottolog, Google	latitude: 46.3913 longitude: 24.2256
Official status	CLDR, Google	official in Romania, Moldova
Endangerment	Glottolog	safe
Scope	ISO 639	individual language
Macrolanguage BCP-47 code	ISO 639	
Individual language BCP-47 codes	ISO 639	
Description	Wikidata	Eastern Romance language

speaker counts and other metadata for approximately 2,800 languages, but there are more metadata categories and many more languages that can be covered.

Currently, open-source language metadata is scattered across these and other sources in various different formats. The aim of LinguaMeta is to unify this metadata and make it available in a simple machine-readable format. Crucially, we link every data point back to its original source. In this way, we also document and identify gaps in our current state of knowledge about world languages across these metadata categories. This paper provides an overview of the various categories and sources of metadata, and discusses various issues uncovered in the process of bringing them together. It also provides an analysis of the metadata in terms of its language coverage.

2. Metadata categories and sources

The various metadata categories and their primary sources are given in Table 1 with examples for Romanian (ron). Information about the sources themselves is provided in Table 2.[4] Here we examine each category and its source(s) in more detail, taking into account the reliability and comprehensiveness of the various sources of metadata.

In cases where the sources in Table 2 provided conflicting information, we reconciled these differences in one of two ways. If one source was more authoritative than another, we systematically privileged the more authoritative source; we discuss these cases in their individual sections below. In other instances, we conducted our own research, choosing the data point that was most consistent across our integrated sources as well as more detailed materials, such as published academic work, census data, and constitutions.

2.1. Language codes

Language codes are intended to uniquely identify languages, in order to dispel the ambiguity that can arise, for example, when several languages happen to have the same name, e.g. Lele (lel, DRC), Lele (llc, Guinea), Lele (lle, Papua New Guinea) and Lele (lln, Chad). However, there is only partial general agreement about what types or varieties of languages should be designated with these codes – see e.g. Good and Hendryx-Parker (2006); Good and Cysouw (2013) among others on the inherent issues in defining what 'languages' are. In the early 2020s, there is general agreement that there are around 7,000 living languages, though it is notable that this figure has risen from the commonly cited 6,000 in the early 2000s and late 20th century. Definitions of what should be designated as a language are clearly evolving. In general, the trend is towards 'splitting', i.e., acknowledging that some languages which were previously considered varieties of other languages are in fact languages in their own right. An example is the 43 varieties of Quechua (que)

[4] One resource, GlotScript, is included but is not yet incorporated in the current version of LinguaMeta; see Section 2.4 for discussion of this resource.

Table 2: Metadata sources incorporated into LinguaMeta.

Source	Description
CLDR (Common Locale Data Repository)	Repository of language and locale metadata maintained by the Unicode Consortium
GlotScript	Repository of writing system information for 7000+ languages developed by Kargaran et al. (2023)
Glottolog	Catalogue of the world's languages, language families, and varieties developed by Hammarström et al. (2024)
Google Research	Research conducted at Google, including metadata for 2800+ languages developed by van Esch et al. (2022) and additional research and refinements of that dataset carried out by the authors
IETF (Internet Engineering Task Force)	BCP-47 language codes maintained by IETF, an Internet standards organization
ISO 639-3 Registration Authority	ISO 639-3 codes and metadata maintained by SIL International, a non-profit Christian organization
Wikidata	Open-source knowledge base of structured data maintained by Wikimedia
Wikipedia	Open-source online encyclopedia maintained by Wikimedia
Wiktionary	List of languages and scripts supported by Wiktionary, a free dictionary maintained by Wikimedia

which have been designated with separate active ISO 639 codes.

Another related issue is macrolanguages. Languages such as Quechua, Chinese (zho) and Arabic (ara) function like languages in some ways, e.g. they have writing systems, vocabularies and enjoy considerable cultural life. However, they are not fully-fledged spoken varieties in the same way as e.g. Mandarin (cmn) or Cantonese (yue); instead they function more like written lingua francas, or like Arabic, where the macrolanguage is only spoken in certain formal social contexts. As well as being grouped under macrolanguages, languages can also fall into broader families according to proposed genealogical lines of descent from historical predecessors. For example, the Chinese languages fall in the broader Sino-Tibetan family.

Language families, macrolanguages, spoken languages, language varieties and also ancient, historical, constructed and signed languages have all been designated with language codes in different systems. In some cases like the ISO 639 standard, distinctions are made between these categories, while others e.g. Glottolog make no such distinction (Forkel and Hammarström, 2022). Since the aim of LinguaMeta is to provide unified metadata for researchers and organizations who want to advance language technology for living spoken languages, we limit its scope by excluding language families and ancient, historical, constructed and signed languages. We do include macrolanguages and map the languages they encompass to their macrolanguage code, since the distinction is not always very clear cut, e.g. Uzbek (uzb) and Persian (fas) are classified as macrolanguages, but their codes are commonly used to refer to the predominant variety among the languages they encompass. We also include languages classified as extinct, as some have been shown to still have communities of speakers, e.g. Diyari (dif) (Austin, 1978).

Despite the issues outlined in this section, language codes are one of the most comprehensive and reliable metadata categories available. There are several efforts to designate all languages with language codes, and mapping between the various standards is made possible by the efforts of the various organizations, in particular Glottolog, ISO 639 and IETF, which develop and maintain this kind of metadata.

LinguaMeta is organized by BCP-47 code, and also includes ISO 639-3 and 639-2b codes, deprecated BCP-47 codes, and codes used to identify languages on Glottolog, Wikidata, and GRN. For macrolanguages, we provide the BCP-47 codes for their individual languages, and for individual languages, we provide the BCP-47 code for their corresponding macrolanguage (if any).

2.2. Language names

Language names exhibit significant variation, with some languages such as Greek (ell) having a clear and singular English name, while others such as Persian/Farsi have two or more names. Haspelmath (2017) introduces some principles for defining or selecting language names in English. Beyond English, each language also has their own names for other languages, and these can also vary widely from the endonym[5] - a good example of this is the various names for German: cf. *Deutsch* (German, deu), *allemand* (French, fra), *tedesco* (Italian, ita), *němčina* (Czech, ces), *saksa* (Estonian, est) and so on. Exonyms such as these can be neutral, but in some cases, they can also carry negative connotations, and there is a growing trend towards favouring language names which are closer to the endonym - see Dryer (2019)'s response to Haspelmath (2017)

[5]An *endonym* is a name for a language/community used within the community, while an *exonym* refers to a name for a language/community used by those outside of the community.

for discussion. An example is the English name Berber (ber) and its cognates in many European languages. This exonym is perceived negatively by some community members and researchers, who prefer Tamazight (derived from the endonym).

Complete lists of English names for all languages in the group targeted by LinguaMeta are available in the ISO 639 standard and in Glottolog. The latter also contains a wealth of alternative names, though it is not always clear in the source whether the alternative names are in English or in other languages. One subset provided through Glottolog comes from Lexvo (de Melo, 2015), which are tagged with the BCP 47 code for the language that the name is written in, e.g. 'anglais [fr]'. Similar tagging of language names can also be found in the CLDR standard and in Wikidata, and Google has translated language names into around 80 languages in order to provide localized user interfaces in its products.

LinguaMeta unifies these various sources of language names[6] and provides more detailed metadata tags; in particular we have attempted to add the script that the name is written in where that would otherwise be ambiguous. For example, Punjabi (pan) is written in both Gurmukhi and Shahmukhi script, so for each language name in Punjabi, we provide the name itself, a BCP 47 code indicating that the name is written in Punjabi, a script code indicating which script the name is written in, and the source of the name.

2.3. Speaker numbers

As noted by van Esch et al. (2022), there is some inherent uncertainty when it comes to reporting numbers of speakers of different languages, as populations are in constant flux, and in the case of large widely spoken languages, there are difficulties in distinguishing between L1 and L2 speakers. For example, if we count only L1 speakers, Mandarin is the most widely spoken language in the world. However, if we include competent L2 speakers, English and Spanish are likely spoken by much larger numbers of people. Reliable statistics on numbers of L2 speakers are hard to come by, however, so we do not include those in the current version of the metadata, leaving this set of statistics to future research.

LinguaMeta reports total speaker populations as well as populations broken down by locale; this data is readily available for some countries and languages, while for others there is no reliable information.[7] The primary source of speaker numbers is research carried out by Google. A secondary resource is CLDR, which provides population percentage breakdowns for the most widely spoken language(s) in each locale. For languages spoken primarily in a single locale, we have used population statistics reported on individual language Wikipedia pages as well, if population statistics were otherwise unavailable. Since Wikipedia is community-edited, these statistics may be less reliable than data from other sources; however, we note that many (if not all) of the population statistics mentioned on Wikipedia are often informed by other reputable sources, such as academic research, directly or through Ethnologue.

A comparison of macrolanguages and their individual languages in our repository reveals the extent of the inconsistencies with reported population statistics. In theory, the population count for a macrolanguage should equal the sum of the populations for each language it encompasses; however, in practice, this is not always the case. For example, CLDR reports a population of 3.3 million for the macrolanguage Bikol (bik), while the populations of its component languages (bcl, bln, bto, cts, fbl, lbl, rbl, and ubl), sourced from CLDR and Wikipedia, sum to almost 6 million. This discrepancy shows that population statistics require particular attention in the metadata literature to ensure their quality, and need to be improved by community members and organizations that are familiar with language taxonomies (e.g. macro and individual languages, language names). Larger-scale improvement of speaker population data would also be possible if international census databases expand to include L1 speaker counts.

2.4. Writing systems

Metadata on writing systems in LinguaMeta comes primarily from Google's program to develop smartphone keyboards for 1,000+ languages. Another significant resource for this kind of metadata is Wiktionary,[8] which lists writing systems for 5,600+ languages. More recently, Kargaran et al. (2023) released GlotScript, which provides writing system metadata and text language identification tools for 7,000+ languages, though at the time of writing we are not able to evaluate the quality and accuracy of this new resource. We took Google research as our most authoritative source on script information, as it was confirmed and refined by web crawls,

[6] LinguaMeta formats language names according to the capitalization conventions reported by Wikimedia for 60 languages: https://meta.wikimedia.org/wiki/Capitalization_of_Wiktionary_pages

[7] Some reported speaker populations in a particular locale are larger than the total population for that locale. This situation arises when there are a number of speakers outside of that locale, but speaker population breakdowns for those locales are not available.

[8] https://en.wiktionary.org/wiki/Wiktionary:List_of_languages

native speaker input, and user feedback. However, in the absence of any other information, we include the Wiktionary metadata, and we may include GlotScript in a future version of LinguaMeta.

One final method which we have used for languages where we couldn't find any relevant information is to predict what the writing system will be based on the primary locale or region where the language is spoken. For example, as far as we know, all Australian Aboriginal languages are written in Latin script (if they are written at all). The same technique can probably also be applied to languages of Latin America, some African countries, and other parts of Oceania, though the existence of other scripts in these regions makes this technique less reliable as a predictor of the script used to write the language.

Typically, all writing systems in our sources are provided for a particular language with no internal structure or hierarchy. However, as Kargaran et al. (2023) note, there are important distinctions to be made between scripts that are widely used and conventionalized for a language and scripts that are rarer or have more specialized uses. To address these distinctions, we developed a set of tags that capture the range of uses that we encountered in our research:

- *Canonical use*: Scripts that have a current, significant cultural life for the language. To facilitate applications, exactly one script is designated as the canonical script for each language-locale combination in our database.
- *Has official status*: Scripts that are officially used to write the language, as designated by a governing body.
- *Has symbolic value*: Scripts that have particular symbolic value for a language community (e.g. as a marker of community identity).
- *Widespread use*: Scripts that are not designated as "canonical", but are in widespread community use.
- *Accessibility use*: Scripts that are used to accommodate disabled users (e.g. Braille)
- *Transliteration use*: Scripts that are not canonical for a language, but are used in certain contexts (e.g. in smartphone keyboards) for transliteration or text input.
- *Minority use*: Scripts that are used by a minority of speakers of the language.
- *Historical use*: Scripts that were previously, but are no longer, used for the language.
- *Religious use*: Scripts that are used primarily in religious or liturgical contexts.

We have applied this richer tag system primarily to languages that use multiple scripts; in the future, we hope to extend this taxonomy to categorize all script-language pairs in LinguaMeta.

Another related issue is the use of different scripts for the same language in different locales: for example, the prescribed use of Devanagari script in India versus Perso-Arabic script in Pakistan for languages like Kashmiri (kas). Here the definition of language-script-locale combinations, cf. the notion of language-locales discussed in Section 2.8, becomes particularly important for the development of appropriately localized language technologies.

2.5. Locales, regions and official status

Metadata about the locales or regions where a language is spoken, and its official status (if any), can help with more targeted localization programs. For example, when an organization wants to extend their services in a new region, understanding which are the most widely spoken languages, and which have governmental support, can help with planning and prioritization for these kinds of projects, particularly in regions of the world where this kind of information is less widely known.

Locale metadata was also collected for the keyboard development program at Google, and has been confirmed in the same way as speaker counts. A secondary source for this metadata is Glottolog, which hosts this information for all languages. Metadata about official status is available for around 500 language-locale pairs from CLDR, which divides the status into three categories: official, regional official, and de facto official. There may be more languages which have official status, but it's likely that nearly all languages which have one of these statuses are covered by CLDR, since typically only a few languages in each locale are afforded this status. Notable exceptions are Bolivia with 37, India with 22, Zimbabwe with 16, Mali with 13, and South Africa with 12 official languages.

Another related type of metadata is geographical coordinates, which are provided by Glottolog and supplemented by our own research. One specific large-scale change which we have implemented is to shift coordinates which are outside the geographical borders of a locale to a location within them. It should be noted, however, that languages are typically spoken across broad and not necessarily contiguous geographical areas, and sometimes straddle political borders, so deciding on a geographical center for a given language or language-locale is often somewhat arbitrary and could be perceived as favouring certain dominant groups.

Despite these issues, coordinates can be useful for creating visualizations of language geography,

so we include them in LinguaMeta with the caveats outlined here.[9]

2.6. Endangerment status

Metadata about language endangerment indicates the level of risk that a language will no longer be spoken, typically evaluated in terms of intergenerational transfer (e.g. Hale et al. 1992). Along with population size and locale, a language's endangerment status may provide useful context about the status of a language, and help inform what engagement model should be chosen to understand community needs and desires around language technology.

Endangerment statuses reported in LinguaMeta are sourced from Glottolog and follow the endangerment scale developed by the UNESCO Atlas of the World's Languages in Danger (Moseley 2010), whose possible values are: safe/not endangered, vulnerable, definitely endangered, severely endangered, critically endangered, and extinct.[10] Descriptions of these categories and how they are determined can be found in Moseley (2010); however, we note that there is no absolute threshold for classifying a language as endangered.

2.7. Scope and type

LinguaMeta includes one additional category provided by ISO 639, namely scope. The Scope category distinguishes between individual languages and macrolanguages, which have their own language codes and encompass several individual languages. See §2.1 for more discussion of macrolanguages.

ISO 639 also includes information about language type, which identifies five types of languages: *living* languages; *extinct* languages, whose last native speaker died within the last few centuries; *ancient* languages, whose last native speaker died more than a millennium ago, e.g. Latin (lat); *historical* languages, which are older forms of living languages, e.g. Old English (ang); and *constructed* languages. We do not include the type category in LinguaMeta because extinct languages are identified under endangerment status, and as mentioned in §2.1, we exclude ancient, historical, and constructed languages from the repository. One notable exception is Esperanto (epo), a constructed language, which we include because it is already supported in a number of language technologies.

[9] https://glottolog.org/glottolog/glottologinformation#coordinates

[10] Note that some so-called "extinct" languages are spoken today as a result of language reclamation efforts (see e.g. Leonard 2023).

2.8. Language-locales and varieties

LinguaMeta organizes much of its metadata by language-locale, a concept which is prevalent in language technology development. Unlike language varieties, which are primarily defined in terms of their linguistic features (e.g. variation in pronunciation, vocabulary, or grammatical structure), language-locales typically define a set of parameters that are useful in creating localized language technologies: for example, the combination of a language code, a script, a locale or region, a set of spelling conventions, special symbols like localized currency symbols, and even the position of keys on a keyboard layout. Language-locales are similar to varieties in that they define a specific type or usage of a language, but they typically cross-cut languages in completely different ways from varieties. We include some language-locale-type metadata in LinguaMeta, including number of speakers broken down by locale and the script(s) used in each locale, to aid in the development of localized language technologies.

In addition to language-locale definitions, information about linguistically-defined language varieties can also be helpful for diversity and inclusion initiatives. For example, when we develop speech technologies, in particular speech recognition and language identification technologies, ideally the system should be able to identify and transcribe the speech of everyone in the community equally well, no matter which variety of the language they speak. The reality is quite far from this ideal; social inequities have found their way into technology development even for the most highly-resourced languages - see e.g. Koenecke et al. (2020) for a discussion of racial inequalities in speech recognition systems. Ideally, this issue could be addressed by incorporating information on language varieties to design more inclusive data collection and testing programs which include speakers of all varieties and from all sectors of the community.

Unfortunately, language varieties have not been subject to the same level of codification as languages and language families. The ISO 639-6 standard (ISO, 2009) attempted to define language varieties with four-letter codes, but it was withdrawn in 2014 due to issues with reliability. Glottolog also contains some entries for language varieties, which they designate with the same alphanumeric codes as other language entries. This data primarily comes from Multitree, which the Glottolog authors acknowledge as being incomplete and unreliable.[11] One resource which contains definitions of over 12,000 language varieties is Global Recordings Network (GRN), which des-

[11] https://glottolog.org/glottolog/glottologinformation#dialects

Table 3: Coverage breakdown by category

Metadata type	Number of languages	%
BCP-47 code	7511	100
ISO 639-3 code	7511	100
English name(s)	7511	100
Locale(s)	7477	99
Endangerment status	7335	98
Writing systems	6498	87
Coordinates	6482	86
Number of speakers	6253	83
Endonyms	1136	15

ignates each variety with a unique numeric code. For languages about which we have some professional knowledge, the varieties reported by GRN do seem to be fairly exhaustive and reliable; however, GRN data is not licensed for commercial use, which prevents us from including this data in the current version of LinguaMeta. We hope to include language variety information in future versions of LinguaMeta.

3. Implementation and analysis

LinguaMeta is available on GitHub in JSON format. The hierarchical nature of JSON allows each data point to be stored along with its source, and represents complex data points (e.g. language names) in an accessible way. We have also generated a TSV that provides a summary of key information organized by language; due to formatting limitations, the TSV file does not include source information, extended language name data, or detailed writing system information. Finally, the GitHub repository contains documentation to facilitate the use of the database.

In total, LinguaMeta contains 7,511 languages with BCP-47 and ISO 639-3 codes. While some metadata categories such as official status are expected to be incomplete, as only some languages have such a status, it is desirable for most of the other categories to have some value. We have identified 9 such categories, namely: BCP-47 code, ISO 639-3 code, English names, locales, endangerment status, writing systems, coordinates, number of speakers, and endonyms. Ideally, all of these categories should have some value for each language, so the first indication of completeness of the resource that we can provide is the possible number of values less the actual number of values. The possible number of values is 67,599, and the actual number is 57,714, or around 85%. We provide a more detailed breakdown by category in Table 3.

As Table 3 shows, language codes, English language names, locales, and endangerment status are more or less feature-complete in the metadata. Writing system metadata is approaching completeness, though it should be noted that not all languages are written, so we should not necessarily expect to have a value here for all languages. Similar coverage is found for coordinates and number of speakers, the latter often being unavailable for less widely-spoken languages. Endonyms are also much less widely available in existing resources, explaining its relatively low coverage.

In addition to a coverage analysis by language, we also provide an analysis of individual data points and their sources in Table 4. Here, we have only analyzed categories where data points come from multiple sources (see Table 1 for a full list of categories and their sources). For each category, we have bolded the source that provided the most data points in that category: notably, Google research provides the most endonym and writing system data, while Glottolog provides the majority of locale and geolocation data. The category that includes the widest variety of sources is English names, which also has the highest number of total entries. Note that a single language often has multiple data points in each category, so the total number of data points in a category may exceed the total number of languages.

Of course, these analyses do not offer any insight into the quality or veracity of this metadata. For that we need input from researchers and language communities; we encourage users to provide feedback on LinguaMeta via GitHub issues. By interacting with this resource, we hope that issues present can be identified and improved over time. This is partially mitigated by the fact that the metadata is largely sourced from existing open repositories with active contributor communities, meaning that improvements in these repositories should find their way into LinguaMeta over time.

4. Conclusion

We have introduced LinguaMeta, a unified repository of language metadata across various categories. We hope that this resource will support researchers and organizations in their efforts to extend language technology to many more languages than are currently supported. A possible extension which might offer even more useful insights would be to integrate this resource with locale-level statistics on internet availability and smartphone usage, in order to help understand where language technologies would be most useful now and in the future. Other potential extensions include language family information, and samples of writing and audio recordings, though the latter two begin to fall more into the domain of language data rather than metadata, for which many cross-resource aggregators already exist

Table 4: Number of data points from each source in LinguaMeta.

Metadata type	Total entries	CLDR	Glottolog	Google	IETF	ISO 639	Wikidata	Wikipedia	Wiktionary	Assumed*
English name	13599	35	834	251	268	**7409**	4263		468	
Endonym	2266	215		**1209**			808			
Writing system	7931			**3751**			50		3143	987
Locale	9147		**5501**	3646						
Number of speakers	6498	1090		2355				**3053**		
Official status	576	**419**		157						
Geolocation	6757		**5951**	806						

(van Esch et al., 2024). We actively encourage feedback and suggestions on the current version of the data set, so that the coverage and quality of LinguaMeta may be further improved.

Peter Kenneth Austin. 1978. *A grammar of the Diyari language of north-east South Australia*. Ph.D. thesis.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Gerard de Melo. 2015. Lexvo.org: Language-related information for the Linguistic Linked Data cloud. *Semantic Web*, 6(4):393–400.

Matthew S Dryer. 2019. Language names and nonlinguists: A response to Haspelmath. *Language Documentation and Conservation*, 13:580–585.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, Twenty-seventh edition. SIL International, Dallas, TX, USA.

Robert Forkel and Harald Hammarström. 2022. Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web*, 13(6):917–924.

Jeff Good and Michael Cysouw. 2013. Languoid, doculect, and glossonym: Formalizing the notion 'language'.

Jeff Good and Calvin Hendryx-Parker. 2006. Modeling contested categorization in linguistic databases. In *Proceedings of the EMELD 2006 Workshop on Digital Language Documentation: Tools and standards: The state of the art*, pages 20–22.

Ken Hale, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, and Nora C. England. 1992. Endangered languages. *Language*, 68(1):1–42.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. *Glottolog 4.8*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Martin Haspelmath. 2017. Some principles for language names. *Language Documentation and Conservation*, 11:81–93.

ISO. 2009. ISO 639-6:2009. Codes for the representation of names of languages — Part 6: Alpha-4 code for comprehensive coverage of language variants. Technical report, International Organization for Standardization.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2023. Glotscript: A resource and tool for low resource writing system identification. *arXiv preprint arXiv:2309.13320*.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Wesley Leonard. 2023. Refusing "endangered languages" narratives. *Daedalus*, 152(3):69–83.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*, 3rd edition edition. UNESCO, Paris.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Wei-Ning Hsu Xiaohui Zhang, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages.

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera. 2022. Writing system and speaker metadata for 2,800+ language varieties. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5035–5046, Marseille, France. European Language Resources Association.

Daan van Esch, Sandy Ritchie, Sebastian Ruder, Julia Kreutzer, Clara Rivera, Ishank Saxena, and Isaac Caswell. 2024. Connecting language technologies with rich, diverse data sources covering thousands of languages. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING) 2024*, Turin, Italy.

Daan van Esch, Elnaz Sarbar, Tamar Lucassen, Jeremy O'Brien, Theresa Breiner, Manasa Prasad, Evan Elizabeth Crew, Chieu Nguyen, and Francoise Beaufays. 2019. Writing across the world's languages: Deep internationalization for Gboard, the Google keyboard. Technical report.