Connecting Language Technologies with Rich, Diverse Data Sources Covering Thousands of Languages

Daan van Esch[1], Sandy Ritchie[1], Sebastian Ruder[2], Julia Kreutzer[3]
Clara Rivera, Ishank Saxena[1], Isaac Caswell[1]
Google Research[1], Cohere[2], Cohere For AI[3]
1600 Amphitheatre Parkway, Mountain View, CA 94043, USA[1]
171 John St, Toronto, ON M5T 1X3, Canada[2,3]
{dvanesch,sandyritchie,ishanksaxena,icaswell}@google.com {juliakreutzer,sebastianruder}@cohere.com

Abstract

Contrary to common belief, there are rich and diverse data sources available for many thousands of languages, which can be used to develop technologies for these languages. In this paper, we provide an overview of some of the major online data sources, the types of data that they provide access to, potential applications of this data, and the number of languages that they cover. Even this covers only a small fraction of the data that exists; for example, printed books are published in many languages but few online aggregators exist.

Keywords: language resources, under-resourced languages

1. Introduction

More than 7,000 languages are spoken around the world today (Eberhard et al., 2024).[1] Efforts are ongoing to extend technology to many more of the world's languages (Ruder, 2022), in many different settings: industry labs, for example Meta's No Language Left Behind (NLLB Team et al., 2022; Pratap et al., 2023) and Google's 1,000 Languages Initiative (Bapna et al., 2022); NGOs, for example SIL's work supported by AWS; universities; and grassroots research communities like Masakhane.

There is clear demand for language technologies in many communities (Littell et al., 2018; Mager et al., 2018; Soria et al., 2018; van Esch et al., 2019; Öktem et al., 2020) but much work remains to be done (Blasi et al., 2022; Ranathunga and de Silva, 2022; Simons et al., 2022).

Extending technologies like keyboards, speech recognition, speech synthesis, machine translation, and language learning applications to hundreds or even thousands of languages involves addressing a number of areas. The first and most important is to establish the needs and wishes of language communities in this regard. Depending on socio-cultural factors specific to each community, one or more language technologies may not be helpful, and some languages may not need any technology at all. Bird (2022) categorizes languages into three major groups: local, contact, and standardized languages, and questions whether most language technologies are desirable for local languages, since people who speak these languages may communicate and exchange information in very different ways than Western societies.

With the community's wishes established, the more technical and operational aspects of language technology development need to be considered.[2] All language technologies need (1) algorithms and modeling approaches, like BERT (Devlin et al., 2019) or wav2vec (Schneider et al., 2019); (2) data, e.g. a text corpus; (3) infrastructure to train and evaluate models, e.g. PyTorch or TensorFlow; (4) compute to train models; (5) distribution platforms like Android or iOS to launch and maintain new models and iterate based on user feedback, and (6) awareness campaigns and education programs to inform people about the availability, use cases, and limitations of a new language technology.

Historically, the main bottleneck has been a combination of (1) and (2). Because algorithms and modeling approaches required lots of data, and that data wasn't easily available in the quantities required, a narrative took hold that extending language technology to hundreds or even thousands of languages would be nearly impossible. In the meantime, though, keyboard apps have been able to make progress, as relatively little data is needed to develop the language technologies needed to enable smart keyboards. Many hundreds of language varieties are supported by in-

[1] This paper focuses on spoken languages. Language technology for signed languages also deserves greater attention – see e.g. Karpov et al. (2016); Papastratis et al. (2021); Yin et al. (2021) – but is beyond the scope of this paper.

[2] Here we are assuming (1) the language in question has a writing system in use, and (2) foundational technologies like Unicode, fonts and rendering are already in place for the script that the community uses to write the language.

LREC-COLING 2024, pages 3729–3746

dustry products like Gboard and SwiftKey, as well the less-centralized KeyMan. In addition to these large-scale "massively multilingual" efforts, other initiatives like FirstVoices have built keyboards tailored to one specific language or a smaller set of related languages.

Thanks to an impressive amount of research progress in algorithms and modeling techniques, it is now possible to get to acceptable levels of accuracy and usability even for more advanced technologies like machine translation, speech recognition, and speech synthesis using orders of magnitude less data (Ruder, 2022). Some recent examples of progress in this space include OpenAI's Whisper, a massively multilingual ASR model (Radford et al., 2022), HuggingFace's BLOOM, a multilingual large language model (Workshop, 2022), or the Aya model, a massively multilingual instruction-finetuned language model (Üstün et al., 2024).

Of course, some language-specific data is still required, and this remains a bottleneck (Ruder (2022) "Challenge #1: Limited Data"). Joshi et al. (2020) analyze the size of labeled data resources to assess the state of NLP for 2,485 languages, of which fully 2,191 (88.17%) are classified as having "exceptionally limited resources", meaning "it will be a monumentous, probably impossible effort to lift them up in the digital space" and noting that "88% of the world's languages [have] virtually no text data available to them while 5% of languages [have] very limited text data available."

However, we do not believe the situation is actually that bleak, as only a few data sources were taken into consideration for the study by Joshi et al. (2020). For labeled data, the LDC catalog and the LRE Map are used, and Wikipedia for unlabeled data. But aggregators of linguistic research like OLAC and Glottolog, for example, list many more resources. And for machine translation, aggregators OPUS-MT and WMT are not considered, although those mainly drive the progress in the field. There exist also data and model catalogues for specific geographic regions, like Lanfrica (Emezue and Dossou, 2020a) for NLP resources for African languages, or continually growing community-driven data collections like the Hugging Face Hub or the BigScience Catalogue of Language Data and Resources (McMillan-Major et al., 2022) or Zenodo. Of course, data could also exist beyond what is known and tracked by these online aggregators.

Our aim is to expand awareness of covered resources in the community and show that there is data available for thousands of languages, across many different data types, building on work done in Prasad et al. (2018). While there are clear differences in the amount of data available across languages, we think our findings warrant a re-evaluation of the general discourse in the field around the unavailability of any resources that could be used for developing language technology. Our findings align with assessments that "data scatteredness (rather than scarcity) is the primary obstacle" (Arora et al., 2022), and that "data is hard to come by, but the 'zero-resource' scenario is a myth" (Pine, 2022). We find the framing by Neubig et al. (2022) helpful: it's not that there aren't any resources, it's just that they need to be unlocked if there is a desire to build technology.

To be clear, we do not doubt that for many languages, very few resources exist at all – there are certainly languages where even basic linguistic research has not yet been done, and no writing system exists, and the first order of business is therefore language documentation (Gippert et al., 2006; Seifart et al., 2018; Hammarström, 2019).

With recent progress towards building high quality solutions with smaller amounts of data, we think it is time for the field of NLP to consider data types and sources that have typically been too small or too challenging to use in the development of language technologies. Smaller efforts targeting a few languages have led the way here, but massively multilingual efforts have typically been restricted to a few hundred languages at most, e.g. Leong et al. (2022).

To support this emerging trend, in this paper we provide an overview of various data types, covering the ways in which each data type can be applied to build language technologies. We then survey various data sources, showing that data is available for thousands of languages. Finally we discuss how more data can be created, and give a brief overview of available tools which can be used to develop technologies for more languages.

2. Data Types and Applications

2.1. Language Metadata

Massively multilingual approaches require an understanding of what languages should go into the models and systems. For example, what are the 1,000 most widely spoken languages? What writing systems are in use? Historically, Ethnologue has been the primary source for this kind of information. More recently, open-source efforts such as Glottolog (Hammarström et al., 2024), van Esch et al. (2022), Kargaran et al. (2023) and Ritchie et al. (2024) have been making this kind of data available to everyone. Language metadata can also be used for planning, prioritization and analyses like the present study of existing coverage in various data sources.

2.2. Language Research and Typology

Research works such as descriptive grammars, and more sophisticated analyses built on top of them, can inform language technology development. Resources like Glottolog aggregate linguistic research publications, and the information these contain can be extracted to create typological databases like WALS (Dryer and Haspelmath, 2013) and Grambank (Skirgård et al., 2023). These record properties of a language, like word formation strategies, morphological marking, word order parameters and syntactic features in machine-readable format. These can be used in multilingual modeling, for example in cases where it is desirable to group languages which have similar features, e.g. Ponti et al. (2019).

2.3. Input, Orthography, and Digitization

Another kind of language resource that is not truly in-language data is orthographic resources, like KeyMan layouts and the LDML layout repository, and exemplars in Unicode's CLDR describing the characters that make up the language's writing system. Defining a standard set of orthographic symbols (and conventions, where possible) for a language is critical for many technologies, such as building keyboards or normalizing text.

2.4. Text

2.4.1. Monolingual

Moving on to in-language resources, monolingual text is the most commonly available, meaning text which is almost entirely in a single language. Monolingual text can be found online in thousands of languages (Prasad et al., 2018), but may also exist in hand-written manuscripts, typed format, or in published print format. Another common type of text data is wordlists, which are sometimes given along with frequencies in a text corpus. Resources like Wikipedia, the Wikimedia Incubator, and StoryWeaver attach an ISO 639 language label the text, making it easy to identify as being in the target language. Even without language labels, it is relatively simple to crawl the web for machine-readable text using text language identification models, see e.g. Brown (2014); Caswell et al. (2020); Abadji et al. (2022). There are also examples of more targeted efforts, e.g., for Uralic languages (Arkhangelskiy, 2019), Peruvian languages (Bustamante et al., 2020), Icelandic (Snæbjarnarson et al., 2022), Norwegian (Kummervold et al., 2022), Latvian (Saulite et al., 2022), community-driven web crawls (Körner et al., 2022), and news corpora crawls (Palen-Michel et al., 2022).

Monolingual text enables a broad range of language technologies, e.g., spelling correction, word prediction and auto-correct in smartphone keyboards. When combined with other languages in a massively multilingual model, monolingual text can also enable machine translation (Siddhant et al., 2020; Ko et al., 2021) and through inclusion in massively multilingual language models like mBERT (Devlin et al., 2019), mT5 (Xue et al., 2021) or BLOOM (Workshop, 2022), natural-language understanding (NLU) and natural-language generation (NLG) tasks like part-of-speech tagging, intent classification, question answering, and summarization (through transfer learning where these tasks were learned from annotated data in other languages). In addition, combining text data with easily-curated grapheme-to-phoneme (G2P) mappings (Bleyan et al., 2019; Wiesner et al., 2019) enables speech recognition (Prasad et al., 2019; Li et al., 2022). For languages where a keyboard layout does not yet exist, monolingual text can enable rapid creation of a draft layout for human editing (Breiner et al., 2019).

We are not aware of any work that enables extending speech synthesis to new languages based on monolingual text alone, but otherwise it is possible to build usable technology like keyboards, speech recognition and machine translation with just this kind of data.

2.4.2. Parallel (Bilingual)

Similar to monolingual texts, a parallel text is a piece of writing in a particular language but crucially, it is paired at the sentence or document level with a translated version in another language. Parallel texts are often created in settings like the EU and the UN, but may also arise from translation of religious materials (Agić and Vulić, 2019; McCarthy et al., 2020b; Akerman et al., 2023), fiction/non-fiction books, manuals, and so on. They can also be created for language learning purposes, where a narrative is written in two (or more) languages as a learning aid. Parallel texts can also be mined from the web, though Kreutzer et al. (2022) find significant quality issues with multilingual corpora available online. The main aggregator of parallel data is OPUS, which includes data from Debian, Mozilla, and LibreOffice (Tiedemann and Thottingal, 2020).

The primary application of parallel texts is machine translation. Even for machine translation models trained primarily with monolingual text, a small parallel corpus is still required for evaluation of the model. Of course, the monolingual text parts of parallel corpora can be also used for all the purposes mentioned above.

2.5. Audio

Audio data consists of recordings of speech, either with or without accompanying transcriptions. The total percentage of transcribed audio is not known, but is likely to be orders of magnitude lower than all audio data available.

Untranscribed audio comes in the form of videos, archival material from language documentation projects and oral history archives, recordings of TV and radio broadcasts, and podcasts. Some examples of repositories include Global Recordings Network, which contains recordings of Biblical texts; VoxPopuli, which consists of recordings of European Parliament events (Wang et al., 2021); and radio broadcast collections, e.g. Danos and Turin (2021); Doumbouya et al. (2021) (see also Radio Garden).

Transcribed audio consists of audio data with either close word-for-word transcriptions, or looser types of transcription. Some examples of multilingual transcribed audio repositories include CMU Wilderness (Black, 2019), Linguistic Data Consortium, Mozilla Common Voice, LDCIL for Indian languages, SADiLaR for South African languages, and Librispeech for audiobooks.

Speech recognition can be achieved using text, untranscribed or transcribed audio — see e.g. (Chen et al., 2022) for an example of multilingual speech recognition with untranscribed audio — but in practice it is usually done using transcribed or loosely transcribed audio. In many cases, as little as an hour of transcribed audio can be sufficient when using a model pretrained on untranscribed audio (Tyers and Meyer, 2021), and like parallel text data for machine translation, more transcribed data is required to evaluate model quality.

High quality single-speaker transcribed corpora like those listed by Foong (2022) can be used for speech synthesis. There are efforts to achieve speech synthesis with smaller amounts of data, e.g., Yang and He (2020); Casanova et al. (2021); Pine et al. (2022); Meyer et al. (2022), but the bottleneck may now be text normalization (Sproat et al., 2001), the development of which can be partially automated with structured data from questionnaires and repositories like CLDR (Ritchie et al., 2019, 2020).

2.6. Multi-Modal

Multi-modal data includes video, for example signed bible translations (Gueuwou et al., 2023), combined image and text data like comic books (see VLRC), story books from StoryWeaver or the Bloom Library, or tagged image sets like Open Images. Multi-modal data can be used for a variety of NLP tasks, see Leong et al. (2022) for an overview. Social media also offers combined video, image and text data, see Cassels (2019) for discussion.

2.7. Structured Data

Structured data comes in various forms, and unlike other data types discussed above, is usually created with a specific purpose related to linguistic research, NLP or other related applications in mind. Some important types include (1) phoneme inventories e.g., PHOIBLE, which can be used to develop G2P mappings for speech recognition and speech synthesis; (2) morphological paradigms, e.g., UniMorph, used for NLP tasks like stemming and part-of-speech tagging; (3) syntactic parameters, e.g., Universal Dependencies and Grambank, which are similarly used for various NLP applications; (4) descriptive grammars and the data they contain, including interlinear glossed examples (Bender et al., 2013), which can be accessed through tools like KORP (Virk et al., 2020) and can be used in NLP applications; and (5) number names and verbalization data e.g. CLDR or UniNum which demonstrate how written numeric tokens like, '123', '$1' or '10m' are read in the spoken domain, and are used in speech synthesis and speech recognition.

3. Existing Data

In this section we document in more detail various online resources that either directly serve or reference data of the types discussed above. Some cover many hundreds or thousands of languages, while others have a more specific focus, for example a particular geographical region. Table 1 summarizes our (non-exhaustive) survey of 70+ online multilingual resources, and Table 2 focuses on bilingual and monolingual resources, which tend to be bottom-up data sources from researchers within language communities. The first column lists the names of the resources and provides a link to the resource homepage, the second column shows the types of data which the resource covers, the third column provides a brief description of the resource, the fourth column either cites the paper which introduced the resource, or cites the resource itself, and the final column lists the number of languages which each resource self-reports as covering, with a link to a list of the languages, or some other method for finding the languages covered, e.g. a search tool.

It is notable that a few resources cover more than 7,000 languages, as this is commonly cited as the total number of languages spoken today. Looking at these cases, it seems that a differentiation is not made between languages and language varieties. Where it is possible to make the distinction, we have only listed varieties which are asso-

Table 1: Online resources, data types and language coverage

Resource	Data types	Description	Reference	# Languages
Panlex	automatic translation; wordlists	Cross-cultural lexical database	Kamholz et al. (2014)	13213
Wikidata	language metadata	Knowledge base for linked data	Vrandečić and Krötzsch (2014)	10287
Glottolog	reference materials; family trees; language metadata; language varieties	Linguistic reference database	Hammarström et al. (2024)	8604
Wiktionary	language metadata	Writing system definitions	Wikimedia (2024c)	8178
ScriptSource	orthography	Writing system resources	SIL International (2024b)	8150
Hugging Face Hub	machine learning models; data	NLP models, data and documentation	Hugging Face Inc. (2024)	8132
URIEL	feature vectors	Database of language features	Littell et al. (2017)	8070
Glosbe	automatic translation; wordlists	Multilingual online dictionary	Glosbe (2024)	7930
Wikipedia	demographic information; language metadata; orthography; phoneme inventories; grammatical data; language varieties	Encyclopedia	Wikimedia (2024b)	7862
Lexvo	knowledge graph	Linked data about languages	de Melo (2015)	7772
LinguaMeta	language metadata	Language metadata database	Ritchie et al. (2024)	7511
GlotScript	language metadata	Writing systems metadata	Kargaran et al. (2023)	7479
Ethnologue	language metadata	Language metadata database	Eberhard et al. (2024)	7151
Joshua Project	demographic information; language metadata; audio data	People group information	Joshua Project (2024)	7140
Global Recordings Network	audio data; language varieties	Audio scripture	Global Recordings Network (2024)	7069
OLAC	reference materials; audio data; speech data	Language archive	Simons and Bird (2003)	6907
ASJP	comparative wordlists	Wordlist database	Wichmann et al. (2022)	5590
Zompist	number names	Number names database	Rosenfelder (2023)	5260
Lexibank	comparative wordlists	Wordlist database	List et al. (2022)	4069
DiACL	grammatical data; wordlists	Atlas of comparative linguistics	Carling et al. (2018)	3420
WALS	grammatical data	World Atlas of Language Structures	Dryer and Haspelmath (2013)	2662
Faith Comes By Hearing	speech data	Audio scripture	Faith Comes By Hearing (2024)	2469
Grambank	grammatical data	Grammatical features	Skirgård et al. (2023)	2467
Phoible	phoneme inventories	Phoneme inventory database	Moran and McCloy (2019)	2186
An Crúbadán	text corpora	Text resources	Scannell (2007)	2093
KeyMan	keyboard layouts; orthography	Keyboard input method editor	SIL International (2024a)	2000
Lanfrica	data; machine learning models; reference materials	Catalogue for African languages	Emezue and Dossou (2020a)	1940
Omniglot	language metadata; phoneme inventories; orthography	Language encyclopedia	Ager (2024)	1775
ABVD	wordlists	Vocabulary database	Greenhill et al. (2008)	1693
Rosetta Project	reference materials	Language archive	The Long Now Foundation (2024)	1300
FUN LangID	LangID model	LangID model	Caswell (2023)	1633
Unilex	words and frequencies	Word frequency database	Unicode (2024b)	1000
OPUS	parallel text corpora	Open parallel corpus	Tiedemann and Nygaard (2004)	721
CMU Wilderness	speech data	Speech datasets	Black (2019)	699
Native Languages of the Americas	language metadata; reference materials	Language materials archive	Native Languages of the Americas (2020)	615
Bloom Library	multi-modal	Educational books	Leong et al. (2022)	550
Forvo	speech data	Pronunciation examples	Pierson (2015)	453
MADLAD-400	text corpora	Common-crawl documents	Kudugunta et al. (2023)	419
Tatoeba	parallel text corpora	Sentence corpus	Tiedemann (2020)	419
Intercontinental Dictionary Series	wordlists	Wordlist database	Borin et al. (2013)	334
Wikimedia incubator	encyclopedias	New encyclopedias incubator	Wikimedia (2024a)	320
Leipzig corpora	text corpora	Text corpora database	Biemann et al. (2007)	274
Phonemica	speech data	Interactive phonetic atlas	van Dam et al. (2021)	240
CLDR	formatting; number names	Common Locale Data Repository	Unicode (2024a)	239
African Storybook	text corpora	Storybooks for African languages	Stranger-Johannessen and Norton (2017)	227
Universal Dependencies	grammatical data	Dependency treebank	Nivre et al. (2016)	194
UniMorph	grammatical data	Morphological paradigms	McCarthy et al. (2020a)	192
Indigenous Tweets	text corpora	Twitter corpus	Scannell (2022)	185
UniNum	number names	Number names database	Ritchie et al. (2019)	182
FastText	text classification	Library for text classification	Joulin et al. (2016)	176
SAILS	grammatical data	Language features database	Muysken et al. (2016)	167
GATITOS	Bilingual lexica	High-quality translations of 4000 common words and phrases	Jones et al. (2023)	173
Common Crawl	web corpora	Web crawl data repository	Common Crawl (2024)	161
Sketch Engine	text corpora; parallel text corpora	Text corpus query tool	Kilgarriff et al. (2014)	145
LDC	text corpora; parallel text corpora; speech data	Linguistic Data Consortium	Liberman and Cieri (1998)	113
Mozilla Common Voice	speech data	Crowdsourced speech datasets	Ardila et al. (2020)	106
Multilingual BERT	language model	Pre-trained transformer model	Devlin et al. (2018)	104
Pollex	wordlists	Polynesian comparative dictionary	Greenhill and Clark (2011)	67
NAILL Index	text corpora	North American indigenous literature	Gref (2016)	63
Twitterphrases	text corpora	Twitter corpus	Kreutz and Daelemans (2019)	50
World Loanword Database	wordlists	Etymology of borrowed words	Haspelmath and Tadmor (2009)	41
AustLang	language metadata	Australian language database	Obata (2009)	40
Wanca	text corpora	Corpora for Uralic languages	Jauhiainen et al. (2019)	29
LDCIL	text corpora; parallel text corpora; speech data	Indian language resources	Mohan and Choudhary (2024)	24
SADiLaR	text corpora; parallel text corpora; speech data	South African language resources	South African Centre for Digital Language Resources (2024)	15
SEAlang	wordlists; text corpora; parallel text corpora	Southeast Asian language resources	Center for Research in Computational Linguistics (2024)	14
Fonbund	speech data	Datasets for NLP and speech	Gutkin et al. (2018)	13
African Voices	speech data	African language recordings	Ogayo et al. (2022)	12
Librispeech	speech data	Audiobook corpus	Panayotov et al. (2015)	8
Корпус-параллельных-текстов	parallel text corpora	Parallel text for Central Asian languages	Shakirov (2023)	8
HornMT	parallel text corpora	Datasets for Horn of Africa	Hadgu et al. (2024)	6
SALT	parallel text corpora	Sentences from Ugandan languages	Akera et al. (2022)	6
Kencorpus	speech data; text corpora	Corpora for Kenyan languages	Wanjawa et al. (2022)	5
NaijaSenti	speech data	Nigerian Twitter Sentiment Corpus	Muhammad et al. (2022)	3
Spoken Wikipedia Corpus	monolingual sentiment; parallel text corpora	Wikipedia speech dataset	Baumann et al. (2019)	3
nicolingua-0005-nqo-nmt-resources	parallel text corpora	English, French and N'Ko translations	Doumbouya et al. (2023)	3

ciated with ISO 639 codes. For other sources, this was not possible, so we simply report the numbers from the source, with an acknowledgement that they are likely to include many language varieties alongside languages.

Given that technologies like keyboards, speech recognition, machine translation, NLP applications, and maybe even speech synthesis can be built with the types of data we have discussed, our survey seems to show that data availability for these language is no longer a major bottleneck. 26 of these resources cover 2,000 or more languages, and 20 cover 3,000 or more. These larger resources alone offer language metadata, reference materials, wordlists, text data, orthography and phoneme inventories, and even audio data for thousands of languages.

4. Standardizing Data

Of course, bringing together and standardizing all this data to make it usable for language technology development is a significant task. For simplicity, we assume here that any relevant data access

Table 2: Community-generated bilingual and monolingual resources

Resource	Data types	Description	Reference	languages
FFR: Fon-French Neural Machine Translation	parallel text corpus	Fon-French parallel sentences	Emezue and Dossou (2020b)	fon,fr
Office Public De La Langue Bretonne(OPDLLB)	parallel text corpus	Breton-French parallel resources	OPLB (2023)	br,fr
Abkhaz-Russian	parallel text corpus	Parallel data from multiple sources	Tlisha (2023)	ab,ru
English-Faroese	parallel text corpus	Translated sentences from English to Faroese	Andersen (2021)	fo,en
Chuvash-Russian	parallel text corpus	Automatically aligned sentences	Antonov (2023)	cv,en
Jojajovai Guarani-Spanish	parallel text corpus	Manually aligned sentences	Chiruzzo et al. (2022)	gn,es
common-parallel-corpora	parallel text corpus	N'Ko versions of NLLB and Flores	Doumbouya et al. (2023)	nqo,en
ENLUS	text corpora; parallel text corpus	Parallel and monolingual corpus for English and Mizo	Lalrempuii and Soni (2024)	lus,en
ABC Cantonese Parallel Corpus	parallel text corpus	Parallel sentences between Cantonese and English	Ayaka (2023)	yue, en
Digital Umuganda	speech data	ASR corpus for Kinyarwanda	Rutunda (2022)	rw
Digital Umuganda	images	Lingala Dataset	Rutunda (2023)	ln
Samromur	speech data	Icelandic Speech corpus	Mollberg et al. (2020)	is
Samromur Queries	speech data	Icelandic Speech corpus	Staffan Hedström (2021)	is
Samromur Children	speech data	Icelandic Speech corpus	Hernandez Mena et al. (2022)	is
Faroese BLARK	speech data	100 hours of transcribed Faroese speech (over 400 speakers).	Simonsen et al. (2022)	fo
Egyptian Arabic Chat Corpus	parallel text corpus	BOLT Egyptian Arabic SMS/Chat Parallel Training Data	Tracey et al. (2021)	arz
Makerere Radio Speech	speech data; text corpus	Luganda Radio Corpus	Mukiibi et al. (2022)	lg
SAHAAYAK	parallel text corpus	the Multi Domain Bilingual Parallel Corpus of Sanskrit to Hindi	Bakrola and Nasariwala (2023)	sa,hi
YouTube-ASL	video corpus	American Sign Language videos with English captions	Uthus et al. (2022)	ase
Central Kurdish dataset	parallel text corpus	200K manually-aligned translations	Amini et al. (2021)	ckb
English-Twi Parallel Corpus	parallel text corpus	50k human-translated sentences into Twi	Azunre et al. (2021)	en,tw
Crowdsourced English-Oromo Parallel Corpus	parallel text corpus	40k human-translated English sentences into Oromo	Chala et al. (2021)	en,om
Kurdish Parallel corpus	parallel text corpus	Scraped news websites	Ahmadi et al. (2020)	en,kmr,ckb
SDS-200	parallel speech and text corpus	Swiss German Speech to Standard German text corpus	Plüss et al. (2022)	gsw,de
Swiss Parliaments Corpus	parallel text corpus	Swiss German Speech to Standard German text corpus	Plüss et al. (2020)	gsw,de

discussions have already been completed (though in all transparency, this is a non-trivial task). The first task is to gather the data from the source. This can be a challenge – as noted above, many of the resources listed in Table 1 only link to or reference data rather than serving it themselves, and few have clear instructions on how to go about downloading or otherwise making copies of the data (Yi et al., 2022).

With the data in hand, the next step is to understand the format, including naming conventions, file types and encoding, and so on, and then convert the files or extract the data they contain to a standardized format (Mäkelä et al., 2020; Nordhoff, 2020). For text data, the next step is text normalization, i.e. converting various orthographic standards or non-standard features of the input data to a unified standard. This requires encoding language-specific knowledge of orthographic and formatting conventions: punctuation, capitalization, spelling conventions, and so on, in a standardized format such as a rule-based FST (Chua et al., 2018). For transcribed audio, the next steps are more complex. Typically, each audio file is associated with both a transcription and also some metadata, in particular information about the speaker(s) in the audio like their age, gender, and a unique identifier like their name or some anonymized identifier. This audio-transcript-metadata cluster can be and is stored in many heterogeneous formats – one common structure employed by the linguistic research community in particular is ELAN files, which anchor transcriptions and other annotations along with speaker metadata to the corresponding part of the audio using tiers and timestamps. There is little standardization in the way that tiers and annotation types are named and formatted, meaning custom code is required for each project to extract the relevant data (Levow et al., 2021, Section 7). An alternative approach would be to simply extract the text and audio and use the 'islands of confidence' approach to align them (Liao et al., 2013). This approach pairs up text and audio automatically, using phoneme-mediated zero-shot transfer and a custom language model trained on just the text in that data set. Such an approach might well be the most scalable for converting data sets in many formats into one consistent format, albeit likely with some quality loss.

With aligned text–audio pairs, the next steps for transcribed audio data depend on the downstream task (typically either speech recognition or speech synthesis). In both cases, normalization of the transcripts is required, and there is the additional issue of expansion of non-word tokens: numbers like '102', abbreviations like 'st.', and so on, for which there can be more than one spoken form which needs to be determined — see Pratap et al. (2020) for a practical example of how to resolve this and other issues raised here for a large multilingual dataset. Bakhturina et al. (2021) also provide a general overview of challenges with creating speech datasets.

5. Creating Data Sets

In the absence of any usable data from the sources surveyed in this paper or elsewhere, creating new machine-readable data would be the next step. For text data, if web crawls fail to produce meaningful amounts of data, the next strategy would be to look for printed materials (books, newspapers etc) and use OCR tools like Transkribus to convert them to digital format (Nockels et al., 2022). OCR-generated text has been shown to improve quality for machine translation where other data types fall short (Ignat et al., 2022). Another way to create text data is crowdsourcing, for which there are many approaches, see e.g. Bhatnagar et al. (2021) on creating parallel text data with this method.

To collect audio data, there are various open source tools and methods available. Mozilla Common Voice provides a web platform for collecting audio using text prompts (Ardila et al., 2020), and LIG-AIKUMA offers a mobile app interface which can be used in offline settings (Blachon et al., 2016). Another common technique is to use images as prompts for audio data collection. This method is being used at scale in the Vaani (through Bhashini) and Waxal projects, which aim to improve data availability for Indian and African languages respectively.

6. Model Training Tools and Techniques

When data is available in a standardized format, the next step is to use it to train models for the language technology you want to support. Historically this has been a major bottleneck. Fortunately, this has become much easier in recent years due to the development of various high quality open source tools. For NLP applications, tools like Adapter-Hub and Trankit provide a suite of tools and training techniques (Van Nguyen et al., 2021). For speech recognition, the Elpis toolkit builds on the Kaldi training pipeline, enabling training of speech recognition models with as little as an hour of transcribed audio data (Foley et al., 2018).

Initiatives like Masakhane have shown that open-source approaches that enable more people to build language technologies can help increase language coverage significantly (∀ et al., 2020). Such approaches are especially promising because bringing useful language technologies to more languages involves many more challenges beyond data scarcity alone: as shown by e.g. Joshi et al. (2019) and van Esch et al. (2019), there is much more to the problem than simply training some machine-learning models and making them available for download. Decentralized community-led efforts can play a pivotal role in addressing these challenges.

Work to scale language technologies to more languages needs to be underpinned by algorithms that are robust to languages with different structures, using approaches that require minimal amounts of data per language. Making speech and NLP technologies for a new language should be easy and accessible. This should be possible without any training or inference using a multilingual model, so the only major work required would be to convert existing data. Or with training: preprocessing should be automated as far as possible, and it should be simple to extend an existing model to a new language, either using untranscribed audio, e.g. Khare et al. (2023), or with transcribed audio data.

Of course, depending on the situation, it may be impossible to make any newly-digitized and/or newly-structured resources for a specific language available publicly for all machine-learning researchers to use. However, this should not preclude these resources from being used for expanding the coverage of language technologies: today's vibrant open-source ecosystem offers purpose-built tools like those referenced above, which were designed with user-friendliness in mind so that it is possible to extend language technologies to additional languages without the steep learning curve that tools like Kaldi or MosesMT pose for users without a thorough understanding of machine learning, computer science, or even use of a command-line interface.

7. Conclusion

Given the research advances we briefly covered, the wealth in data types and sources across thousands of languages, and the booming open-source ecosystem, we expect to see language technologies developed in many more languages over the next few years, which gives us cause for optimism. To accelerate this trend, we argue in favour of broader, more interdisciplinary approaches, with stronger connections and more exchange of knowledge between NLP and speech researchers, linguists, language communities, language archives, human-computer interaction researchers (including user experience research and design), and other interested parties.

We believe that if recent technological advances can be combined with the rich and diverse sources of data that exist in places like those surveyed in this paper, this can help bring technology to thousands more languages. We showed how the data that is available can be matched up with technological approaches. We believe that user-friendly tools will be the final piece of the puzzle: once data has been collected, curated and prepared, it is such tools that will enable scale by letting anyone who is interested apply algorithmic advances from core speech and NLP research to many more languages, in order to help meet the clear demand among many language communities for better support for their languages.

Limitations

The major limitation of this study is that this is primarily a high-level overview of various online resources and aggregators: we did not look at these resources in great detail in every language, and cannot vouch for the quality of the data in each language, and as a result even the number of languages which they self-report as covering. We present our understanding of the situation at this point to counter the common narrative around data being a major bottleneck, and hope that future work will refine and build on this analysis – both for online sources as well as for other materials, like printed books.

Ethics Statement

This is a survey of open-source research archives and tools. When considering the use of these resources, it is important to do an ethics assessment of the specific intended use case and application for the relevant language(s), and to check the compatibility of any licenses applicable to the data with the intended use case.

Acknowledgements

We'd like to thank Aidan Pine, Antonios Anastasopoulos, Graham Neubig, Aryaman Arora, Kalika Bali, and Steven Bird for discussions on this topic over the last few years, at various workshops and conferences and on Twitter. Many thanks also to Bhuvana Ramabhadran and Kartik Audhkhasi for their valuable comments and feedback, and to Emily Drummond for help with preparation of the final manuscript. Any remaining errors are our own.

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Simon Ager. 2024. Omniglot: Writing systems and languages of the world. `www.omniglot.com`.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Sina Ahmadi, Hossein Hassani, and Daban Q. Jaff. 2020. Leveraging multilingual news websites for building a Kurdish parallel corpus. *CoRR*, abs/2010.01554.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2022. Machine translation for African languages: Community creation of datasets and models in Uganda. In *3rd Workshop on African Natural Language Processing*.

Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Taeho Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The ebible corpus: Data and model benchmarks for bible translation for low-resource languages.

Zhila Amini, Mohammad Mohammadamini, Hawre Hosseini, Mehran Mansouri, and Daban Jaff. 2021. Central Kurdish machine translation: First large scale parallel corpus and experiments.

Jógvan Andersen. 2021. Sprotin Translations. `https://github.com/Sprotin/translations`.

Alexander Antonov. 2023. Chuvash-Russian parallel corpus. `https://github.com/AlAntonov/chv_corpus?tab=readme-ov-file`.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Timofey Arkhangelskiy. 2019. Corpora of social media in minority uralic languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 125–140, Tartu, Estonia. Association for Computational Linguistics.

Aryaman Arora, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2022. Computational historical linguistics and language diversity in South Asia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1396–1409, Dublin, Ireland. Association for Computational Linguistics.

Ayaka. 2023. ABC Cantonese parallel corpus. `https://github.com/ayaka14732/abc-cantonese-parallel-corpus`.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago

Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021. English-Twi parallel corpus for machine translation.

Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. 2021. A toolbox for construction and analysis of speech datasets.

Vishvajitsinh Bakrola and Jitendra Nasariwala. 2023. Sahaayak 2023 – the multi domain bilingual parallel corpus of sanskrit to hindi for machine translation.

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.

Timo Baumann, Arne Köhn, and Felix Hennig. 2019. The Spoken Wikipedia Corpus collection: Harvesting, alignment and an application to hyperlistening. *Language Resources and Evaluation*, 53:303–329.

Emily M. Bender, Michael Wayne Goodman, Joshua Crowgey, and Fei Xia. 2013. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In *LaTeCH@ACL*.

Rajat Bhatnagar, Ananya Ganesh, and Katharina Kann. 2021. Don't rule out monolingual speakers: A method for crowdsourcing machine translation data.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig corpora collection-monolingual corpora of standard size. *Proceedings of corpus linguistic*, 2007.

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.

David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. 2016. Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. *Procedia Computer Science*, 81:61–66. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

Alan W Black. 2019. CMU Wilderness multilingual speech dataset. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975. IEEE.

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.

Harry Bleyan, Sandy Ritchie, Jonas Fromseier Mortensen, and Daan van Esch. 2019. Developing pronunciation models in new languages faster by exploiting common grapheme-to-phoneme correspondences across languages. In *Proceedings of Interspeech 2019*.

Lars Borin, Bernard Comrie, and Anju Saxena. 2013. The Intercontinental Dictionary Series–a rich and principled database for language comparison. *Approaches to measuring linguistic differences*, 285:302.

Theresa Breiner, Chieu Nguyen, Daan van Esch, and Jeremy O'Brien. 2019. Automatic keyboard layout design for low-resource latin-script languages.

Ralf Brown. 2014. Non-linear mapping for improved identification of 1300+ languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–632, Doha, Qatar. Association for Computational Linguistics.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Gerd Carling, Filip Larsson, Chundra A Cathcart, Niklas Johansson, Arthur Holmer, Erich Round, and Rob Verhoeven. 2018. Diachronic Atlas of Comparative Linguistics (DiACL)—A database for ancient language typology. *PloS one*, 13(10):e0205313.

Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir Antonelli Ponti. 2021. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone.

Morgan Cassels. 2019. Indigenous languages in new media: Opportunities and challenges for language revitalization.

Isaac Caswell. 2023. Fun LangID. https://github.com/google-research/url-nlp/tree/main/fun-langid.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Center for Research in Computational Linguistics. 2024. SEAlang. http://sealang.net/.

Sisay Chala, Bekele Debisa, Amante Diriba, Silas Getachew, Chala Getu, and Solomon Shiferaw. 2021. Crowdsourcing parallel corpus for English-Oromo neural machine translation using community engagement platform. *CoRR*, abs/2102.07539.

Zhehuai Chen, Ankur Bapna, Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Pedro Moreno, and Nanxin Chen. 2022. Maestro-U: Leveraging joint speech-text representation learning for zero supervised speech ASR.

Luis Chiruzzo, Santiago Góngora, Aldo Alvarez, Gustavo Giménez-Lugo, Marvin Agüero-Torales, and Yliana Rodríguez. 2022. Jojajovai: A parallel Guarani-Spanish corpus for MT benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2098–2107, Marseille, France. European Language Resources Association.

Mason Chua, Daan van Esch, Noah Coccaro, Eunjoon Cho, Sujeet Bhandari, and Libin Jia. 2018. Text normalization infrastructure that scales to hundreds of language varieties. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Common Crawl. 2024. Common Crawl. https://commoncrawl.org/.

David Danos and Mark Turin. 2021. Living language, resurgent radio: A survey of indigenous language broadcasting initiatives. *Language Documentation & Conservation*, 15:75–152.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Moussa Doumbouya, Lisa Einstein, and Chris Piech. 2021. Using radio archives for low-resource speech recognition: Towards an intelligent virtual assistant for illiterate users.

Moussa Koulako Bala Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory 2 Condé, Kalo Mory Diané, Chris Piech, and Christopher Manning. 2023. Machine translation for nko: Tools, corpora and baseline results.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

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.

Chris C. Emezue and Bonaventure F. P. Dossou. 2020a. Lanfrica: A participatory approach to documenting machine translation research on african languages.

Chris Chinenye Emezue and Femi Pancrace Bonaventure Dossou. 2020b. FFR v1.1: Fon-French neural machine translation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 83–87, Seattle, USA. Association for Computational Linguistics.

Faith Comes By Hearing. 2024. Faith Comes By Hearing. https://www.faithcomesbyhearing.com/.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL endangered language pipeline and inference system. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018)*.

Ng Wai Foong. 2022. 20 open-source single speaker speech datasets.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, editors. 2006. *Essentials of Language Documentation*. De Gruyter Mouton, Berlin, New York.

Global Recordings Network. 2024. Global Recordings Network. `https://globalrecordings.net/`.

Glosbe. 2024. Glosbe Dictionary. `https://glosbe.com/`.

Simon J Greenhill, Robert Blust, and Russell D Gray. 2008. The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:EBO–S893.

Simon J Greenhill and Ross Clark. 2011. POLLEX-Online: The Polynesian lexicon project online. *Oceanic Linguistics*, pages 551–559.

Emily Gref. 2016. Publishing in North American Indigenous languages.

Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. 2023. JWSign: A highly multilingual corpus of Bible translations for more diversity in sign language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9907–9927, Singapore. Association for Computational Linguistics.

Alexander Gutkin, Martin Jansche, and Tatiana Merkulova. 2018. Fonbund: A library for combining cross-lingual phonological segment data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Asmelash Teka Hadgu, Gebrekirstos G. Gebremeskel, and Abel Aregawi. 2024. HornMT. `https://github.com/asmelashteka/HornMT`.

Harald Hammarström. 2019. Which language should I document? Some concrete suggestions from diversity and endangerment.

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

Martin Haspelmath and Uri Tadmor. 2009. The Loanword Typology project and the World Loanword Database. *Loanwords in the world's languages: A comparative handbook*, 1:34.

Carlos Daniel Hernandez Mena, David Erik Mollberg, Michal Borský, and Jón Guðnason. 2022. Samrómur children: An Icelandic speech corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 995–1002, Marseille, France. European Language Resources Association.

Hugging Face Inc. 2024. Hugging Face Hub. `https://huggingface.co/docs/hub/en/index`.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022. OCR improves machine translation for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Linden. 2019. Wanca in Korp: Text corpora for underresourced Uralic languages. In *Proceedings of the Research data and humanities (RDHUM) 2019 conference*, 17, pages 21–40.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Joshua Project. 2024. Joshua Project: People Groups of the World. `https://joshuaproject.net/`.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. (arXiv:1612.03651). ArXiv:1612.03651 [cs].

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

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*.

Alexey Karpov, Irina Kipyatkova, and Milos Zelezny. 2016. Automatic technologies for processing spoken sign languages. *Procedia Computer Science*, 81:201–207. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

Aparna Khare, Minhua Wu, Saurabhchand Bhati, Jasha Droppo, and Roland Maas. 2023. Guided contrastive self-supervised pre-training for automatic speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting high-resource nmt models to translate low-resource related languages without parallel data.

Erik Körner, Felix Helfer, Christopher Schröder, Thomas Eckart, and Dirk Goldhahn. 2022. Crawling under-resourced languages - a portal for community-contributed corpus collection. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 36–43, Marseille, France. European Language Resources Association.

Tim Kreutz and Walter Daelemans. 2019. How to optimize your Twitter collection: Dutch keywords for better coverage. *Computational Linguistics in the Netherlands Journal*, 9:55–66.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset.

Per Kummervold, Freddy Wetjen, and Javier de la Rosa. 2022. The Norwegian colossal corpus: A text corpus for training large Norwegian language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3852–3860, Marseille, France. European Language Resources Association.

Candy Lalrempuii and Badal Soni. 2024. Enlus: Bridging language divide with english-mizo machine translation corpus.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks.

Gina-Anne Levow, Emily Ahn, and Emily M. Bender. 2021. Developing a shared task for speech processing on endangered languages. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 96–106, Online. Association for Computational Linguistics.

Xinjian Li, Florian Metze, David R Mortensen, Alan W Black, and Shinji Watanabe. 2022. ASR2K: Speech recognition for around 2000 languages without audio.

Hank Liao, Erik McDermott, and Andrew Senior. 2013. Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 368–373.

Mark Liberman and Christopher Cieri. 1998. The creation, distribution and use of linguistic data: the case of the Linguistic Data Consortium. In *LREC*, pages 159–166.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(1):316.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Arya D McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2020a. Unimorph 3.0: Universal morphology. In *Proceedings of The 12th language resources and evaluation conference*, pages 3922–3931. European Language Resources Association.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020b. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Angelina McMillan-Major, Zaid Alyafeai, Stella Biderman, Kimbo Chen, Francesco De Toni, Gérard Dupont, Hady Elsahar, Chris Emezue, Alham Fikri Aji, Suzana Ilić, Nurulaqilla Khamis, Colin Leong, Maraim Masoud, Aitor Soroa, Pedro Ortiz Suarez, Zeerak Talat, Daniel van Strien, and Yacine Jernite. 2022. Documenting geographically and contextually diverse data sources: The bigscience catalogue of language data and resources.

Josh Meyer, David Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon KABONGO KABENAMUALU, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Chinenye Emezue, Jonathan Mukiibi, Salomey Osei, Apelete AGBOLO, Victor Akinode, Bernard Opoku, Olanrewaju Samuel, Jesujoba Alabi, and Shamsuddeen Hassan Muhammad. 2022. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus. In *Proc. Interspeech 2022*, pages 2383–2387.

Shailendra Mohan and Narayan Choudhary. 2024. Linguistic Data Consortium for Indian Languages (LDC-IL). https://www.ldcil.org/.

David Erik Mollberg, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Steinþór Steingrímsson, Eydís Huld Magnúsdóttir, and Jon Gudnason. 2020. Samrómur: Crowd-sourcing data collection for Icelandic speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3463–3467, Marseille, France. European Language Resources Association.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Jonathan Mukiibi, Andrew Katumba, Joyce Nakatumba-Nabende, Ali Hussein, and Joshua Meyer. 2022. The makerere radio speech corpus: A Luganda radio corpus for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1945–1954, Marseille, France. European Language Resources Association.

Pieter Muysken, Harald Hammarström, Olga Krasnoukhova, Neele Müller, Joshua Birchall, Simon van de Kerke, Loretta O'Connor, Swintha Danielsen, Rik van Gijn, and George Saad. 2016. South American indigenous language structures (SAILS) online. https://sails.clld.org/.

Eetu Mäkelä, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mika Hämäläinen, Samuli Kaislaniemi, and Terttu Nevalainen. 2020. Wrangling with non-standard data. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, Riga, Latvia. CEUR-WS.org.

Native Languages of the Americas. 2020. Native Languages of the Americas website. http://native-languages.org/.

Graham Neubig, Shruti Rijhwani, Xinyi Wang, Antonios Anastasopoulos, Daisy Rosenblum, and Sebastian Ruder. 2022. Unlocking resources for under-resourced languages.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Joe Nockels, Paul Gooding, Sarah Ames, and Melissa Terras. 2022. Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of transkribus in published research. *Archival Science*, 22:367–392.

Sebastian Nordhoff. 2020. From the attic to the cloud: mobilization of endangered language resources with linked data. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, pages 10–18, Marseille, France. European Language Resources Association.

Kazuko Obata. 2009. AUSTLANG - online Australian Indigenous languages database. *Incite*, 30(4):23–23.

Perez Ogayo, Graham Neubig, and Alan W Black. 2022. Building African voices. *arXiv preprint arXiv:2207.00688*.

OPLB. 2023. Office Public de la Langue Bretonne. https://www.fr.brezhoneg.bzh/212-donnees-libres-de-droits.htm.

Chester Palen-Michel, June Kim, and Constantine Lignos. 2022. Multilingual open text release 1: Public domain news in 44 languages. In *Proceedings of the Thirteenth Language*

Resources and Evaluation Conference, pages 2080–2089, Marseille, France. European Language Resources Association.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Ilias Papastratis, Christos Chatzikonstantinou, Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. 2021. Artificial intelligence technologies for sign language. *Sensors*, 21(17).

Catherine Pierson. 2015. Forvo: All the words in the world. Pronounced. *Reference Reviews*, 29(7):29–30.

Aidan Pine. 2022. Data is hard to come by, but the "zero-resource" scenario is a myth. [Twitter].

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022. Requirements and motivations of low-resource speech synthesis for language revitalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7346–7359, Dublin, Ireland. Association for Computational Linguistics.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2020. Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus. *CoRR*, abs/2010.02810.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. Sds-200: A swiss german speech to standard german text corpus.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Manasa Prasad, Theresa Breiner, and Daan van Esch. 2018. Mining training data for language modeling across the world's languages. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018)*.

Manasa Prasad, Daan van Esch, Sandy Ritchie, and Jonas Fromseier Mortensen. 2019. Building large-vocabulary ASR systems for languages without any audio training data. In *Proceedings of Interspeech 2019*.

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.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Interspeech 2020*. ISCA.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. Technical report.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world.

Sandy Ritchie, Eoin Mahon, Kim Heiligenstein, Nikos Bampounis, Daan van Esch, Christian Schallhart, Jonas Mortensen, and Benoit Brard. 2020. Data-driven parametric text normalization: Rapidly scaling finite-state transduction verbalizers to new languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 218–225, Marseille, France. European Language Resources association.

Sandy Ritchie, Richard Sproat, Kyle Gorman, Daan van Esch, Christian Schallhart, Nikos Bampounis, Benoît Brard, Jonas Fromseier Mortensen, Millie Holt, and Eoin Mahon. 2019. Unified verbalization for speech recognition & synthesis across languages. In *Proc. Interspeech 2019*, pages 3530–3534.

Sandy Ritchie, Daan van Esch, Uche Okonkwo, Shikhar Vashishth, and Emily Drummond. 2024. Linguameta: Unified metadata for thousands of languages. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING) 2024*, Turin, Italy.

Mark Rosenfelder. 2023. Zompist. `https://www.zompist.com/numbers.shtml`.

Sebastian Ruder. 2022. The State of Multilingual AI. http://ruder.io/state-of-multilingual-ai/.

Samuel Rutunda. 2022. Digital Umuganda Kinyarwanda corpus. https://huggingface.co/DigitalUmuganda.

Samuel Rutunda. 2023. Digital Umuganda Lingala corpus. https://huggingface.co/DigitalUmuganda.

Baiba Saulite, Roberts Darģis, Normunds Gruzitis, Ilze Auzina, Kristīne Levāne-Petrova, Lauma Pretkalniņa, Laura Rituma, Peteris Paikens, Arturs Znotins, Laine Strankale, Kristīne Pokratniece, Ilmārs Poikāns, Guntis Barzdins, Inguna Skadiņa, Anda Baklāne, Valdis Saulespurēns, and Jānis Ziediņš. 2022. Latvian national corpora collection – korpuss.lv. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5123–5129, Marseille, France. European Language Resources Association.

Kevin P Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora (WAC3-2007): Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval*, volume 4, page 5. Presses Univ. de Louvain.

Kevin P Scannell. 2022. Managing data from social media: The indigenous tweets project. *The Open Handbook of Linguistic Data Management*, page 481.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–345.

Iksander Shakirov. 2023. Корпус параллельных текстов. https://github.com/kod-odin/lang-tasks/wiki/3.- - - .

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

SIL International. 2024a. Keyman. https://keyman.com/.

SIL International. 2024b. ScriptSource. https://scriptsource.org/.

Gary Simons and Steven Bird. 2003. The Open Language Archives Community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18(2):117–128.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Annika Simonsen, Sandra Saxov Lamhauge, Iben Nyholm Debess, and Peter Juel Henrichsen. 2022. Creating a Basic Language Resource Kit for Faroese. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*, page 4637□4643.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L.M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R.A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O.C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W.P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith

Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye □□□, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals global patterns in the structural diversity of the world's languages. *Science Advances*, 9.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A warm start and a clean crawled corpus - a recipe for good language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Claudia Soria, Valeria Quochi, and Irene Russo. 2018. The DLDP survey on digital use and usability of EU regional and minority languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

South African Centre for Digital Language Resources. 2024. South African Centre for Digital Language Resources (SADiLaR). https://sadilar.org/.

R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 2001. Normalization of non-standard words. In *Computer Speech and Language*, volume 15, page 287–333.

Ragnheiður Þórhallsdóttir David Erik Mollberg Smári Freyr Guðmundsson Ólafur Helgi Jónsson Sunneva Þorsteinsdóttir Eydís Huld Magnúsdóttir Jon Gudnason Staffan Hedström, Judy Y. Fong. 2021. Samrómur Queries 21.12. Reykjavik University: Language and Voice Lab.

Espen Stranger-Johannessen and Bonny Norton. 2017. The African storybook and language teacher identity in digital times. *The Modern Language Journal*, 101(S1):45–60.

The Long Now Foundation. 2024. The Rosetta Project. https://rosettaproject.org/.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus-parallel and free. In *LREC*. Citeseer.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Nart Tlisha. 2023. Abkhazian-Russian Parallel data. https://huggingface.co/datasets/Nart/parallel-ab-ru.

Jennifer Tracey, Dana Delgado, Song Chen, and Stephanie Strassel. 2021. BOLT Egyptian Arabic SMS/Chat Parallel Training Data.

Francis M. Tyers and Josh Meyer. 2021. What shall we do with an hour of data? speech recognition for the un- and under-served languages of common voice.

Unicode. 2024a. Common language data repository. https://cldr.unicode.org/.

Unicode. 2024b. Unilex. https://github.com/unicode-org/unilex.

David Uthus, Maria Voitovich, and R.J. Mical. 2022. Augmenting poetry composition with Verse by Verse. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 18–26, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Kellan Parker van Dam, Steve Hansen, and Qi Jiayao. 2021. Phonemica. https://phonemica.net/.

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

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.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing.

Shafqat Mumtaz Virk, Harald Hammarström, Markus Forsberg, and Søren Wichmann. 2020.

The DReaM corpus: A multilingual annotated corpus of grammars for the world's languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 878–884, Marseille, France. European Language Resources Association.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Barack Wanjawa, Lilian Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2022. Kencorpus: A Kenyan language corpus of Swahili, Dholuo and Luhya for natural language processing tasks. *arXiv preprint arXiv:2208.12081*.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. 2022. The ASJP database (version 20).

Matthew Wiesner, Oliver Adams, David Yarowsky, Jan Trmal, and Sanjeev Khudanpur. 2019. Zero-shot pronunciation lexicons for cross-language acoustic model transfer. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1048–1054.

Wikimedia. 2024a. Wikimedia incubator. `https://incubator.wikimedia.org/wiki/Incubator:Main_Page`.

Wikimedia. 2024b. Wikipedia, the free encyclopedia. `https://www.wikipedia.org/`.

Wikimedia. 2024c. Wiktionary, the free dictionary. `https://www.wiktionary.org/`.

BigScience Workshop. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jingzhou Yang and Lei He. 2020. Towards Universal Text-to-Speech. In *Proc. Interspeech 2020*, pages 3171–3175.

Irene Yi, Amelia Lake, Juhyae Kim, Kassandra Haakman, Jeremiah Jewell, Sarah Babinski, and Claire Bowern. 2022. Accessibility, discoverability, and functionality: An audit of and recommendations for digital language archives. *Journal of Open Humanities Data*, 8(10).

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, and Grace Tang. 2020. Gamayun - language technology for humanitarian response. In *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 1–4.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model.