Building Language Technologies for Everyone

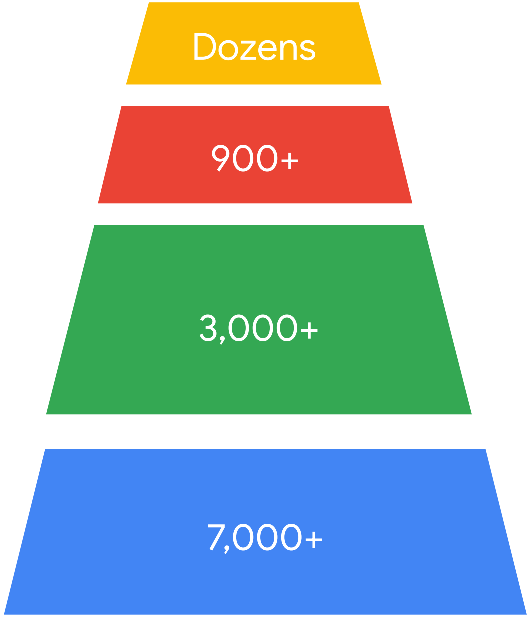
Daan van Esch (@daanvanesch)
Google Research

April 22, 2022
ContribuLing 2022, Paris, France

Language Tech

- Display
  - Unicode → unicode.org
  - Fonts → fonts.google.com/noto and github.com/googlefonts/noto-fonts
  - Rendering → github.com/harfbuzz
- Input
  - Keyboards, physical and virtual (on smartphones)
  - Handwriting recognition and optical character recognition (OCR)
  - Speech recognition (ASR)
- Understanding
  - Morphological analysis
  - Part-of-speech tagging
  - Syntactic parsing
  - Semantic/intent classification
- Generation
  - Text-to-speech
  - Natural-language generation
- Machine translation

Google   Google Assistant   Google Translate


Dozens

900+

3,000+

7,000+

➢ Google **Speech-to-Text** in **~130**, **Text-to-Speech** in **~60**

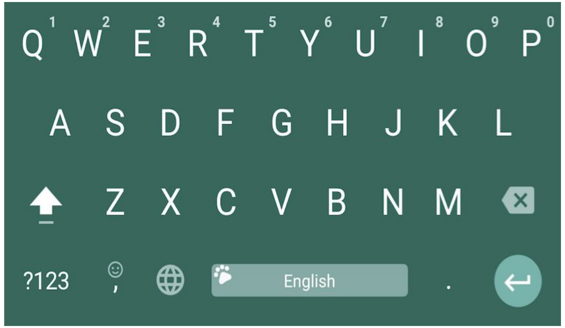
➢ **Gboard** today in **900+** language varieties (Android)
➢ Covering 95% of the world in their first language

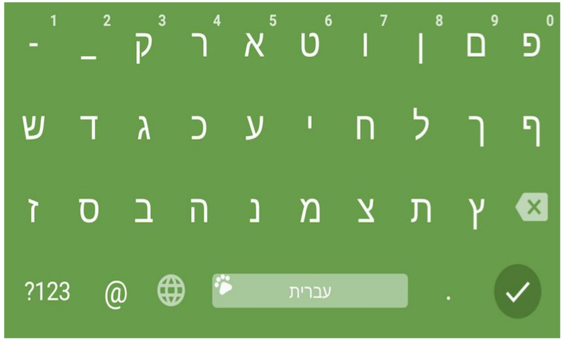
➢ At least 3,000 have some written tradition (probably more)
➢ Almost all living writing systems supported by **Unicode**
➢ Almost all of Unicode supported by Google's free open-source **Noto** fonts & the HarfBuzz renderer

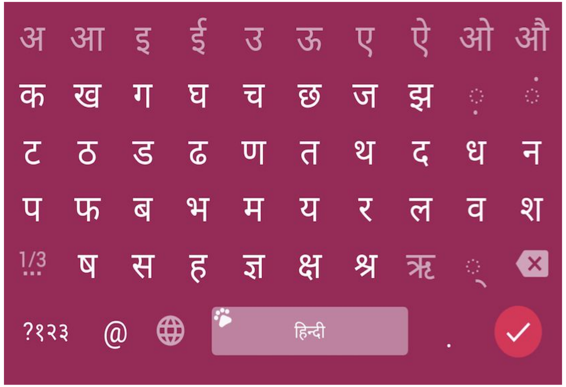
➢ 7,000+ living languages in the world

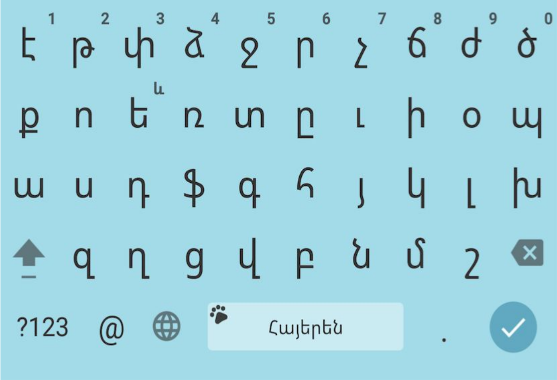
Developing Keyboards for 900+ Languages

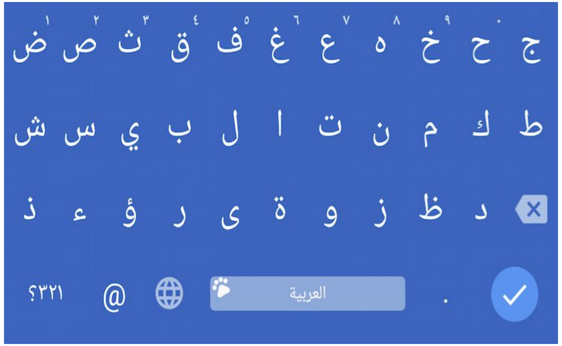
Google

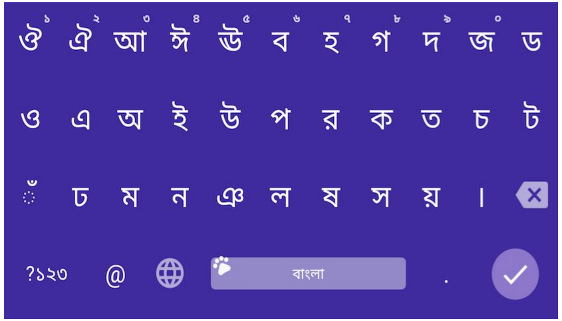

Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
?123
English

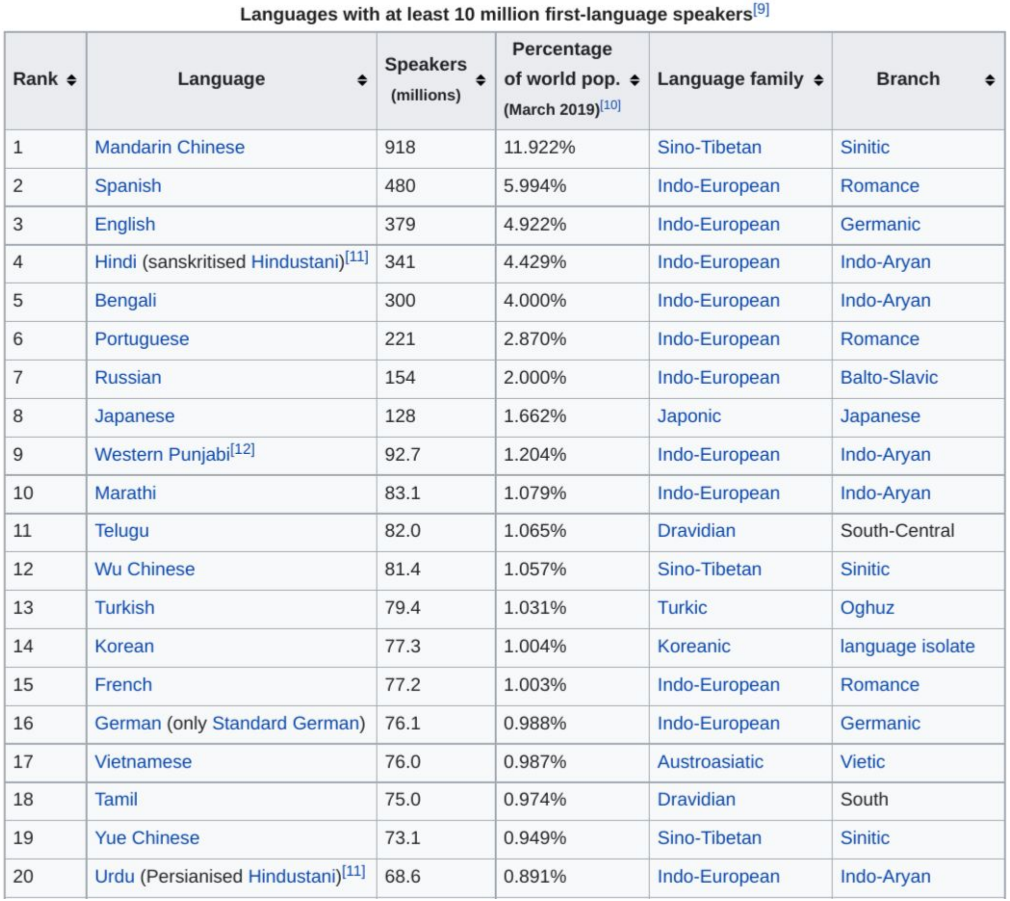

?123
@
עברית


अ आ इ ई उ ऊ ए ऐ ओ औ
क ख ग घ च छ ज झ
ट ठ ड ढ ण त थ द ध न
प फ ब भ म य र ल व श
1/3 ष स ह ज्ञ क्ष श्र ऋ
?१२३
@
हिन्दी


?123
@
Հայերեն


ض ص ث ق ف غ ع ه خ ح ج
ش س ي ب ل ا ت ن م ك ط
ذ ء ؤ ر ى ة و ز ظ د
؟٣٢١
@
العربية

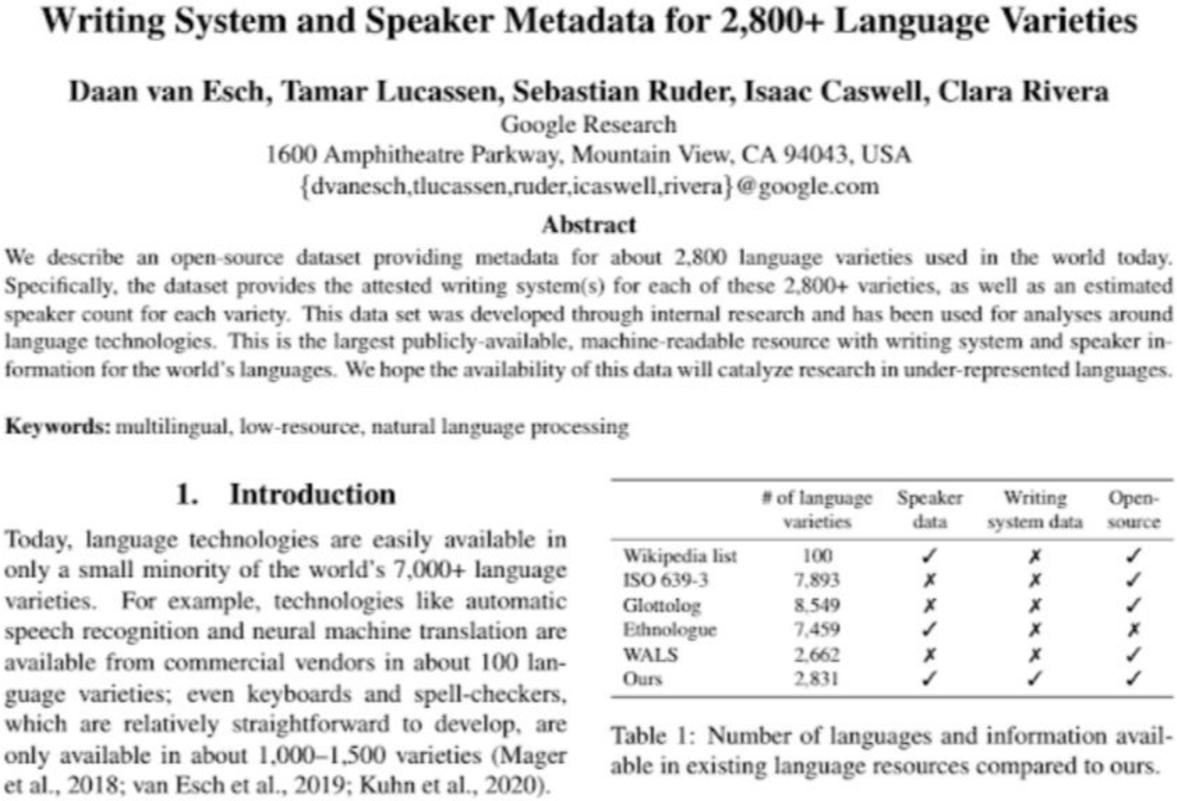

?১২৩
@
বাংলা



Languages with at least 10 million first-language speakers[9]

Rank	Language	Speakers (millions)	Percentage of world pop. (March 2019)[10]	Language family	Branch
1	Mandarin Chinese	918	11.922%	Sino-Tibetan	Sinitic
2	Spanish	480	5.994%	Indo-European	Romance
3	English	379	4.922%	Indo-European	Germanic
4	Hindi (sanskritised Hindustani)[11]	341	4.429%	Indo-European	Indo-Aryan
5	Bengali	300	4.000%	Indo-European	Indo-Aryan
6	Portuguese	221	2.870%	Indo-European	Romance
7	Russian	154	2.000%	Indo-European	Balto-Slavic
8	Japanese	128	1.662%	Japonic	Japanese
9	Western Punjabi[12]	92.7	1.204%	Indo-European	Indo-Aryan
10	Marathi	83.1	1.079%	Indo-European	Indo-Aryan
11	Telugu	82.0	1.065%	Dravidian	South-Central
12	Wu Chinese	81.4	1.057%	Sino-Tibetan	Sinitic
13	Turkish	79.4	1.031%	Turkic	Oghuz
14	Korean	77.3	1.004%	Koreanic	language isolate
15	French	77.2	1.003%	Indo-European	Romance
16	German (only Standard German)	76.1	0.988%	Indo-European	Germanic
17	Vietnamese	76.0	0.987%	Austroasiatic	Vietic
18	Tamil	75.0	0.974%	Dravidian	South
19	Yue Chinese	73.1	0.949%	Sino-Tibetan	Sinitic
20	Urdu (Persianised Hindustani)[11]	68.6	0.891%	Indo-European	Indo-Aryan

(Source: English Wikipedia, "List of languages by number of native speakers")



Writing System and Speaker Metadata for 2,800+ Language Varieties

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, Clara Rivera
Google Research
1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
{dvanesch,tlucassen,ruder,icaswell,rivera}@google.com

Abstract

We describe an open-source dataset providing metadata for about 2,800 language varieties used in the world today. Specifically, the dataset provides the attested writing system(s) for each of these 2,800+ varieties, as well as an estimated speaker count for each variety. This data set was developed through internal research and has been used for analyses around language technologies. This is the largest publicly-available, machine-readable resource with writing system and speaker information for the world's languages. We hope the availability of this data will catalyze research in under-represented languages.

Keywords: multilingual, low-resource, natural language processing

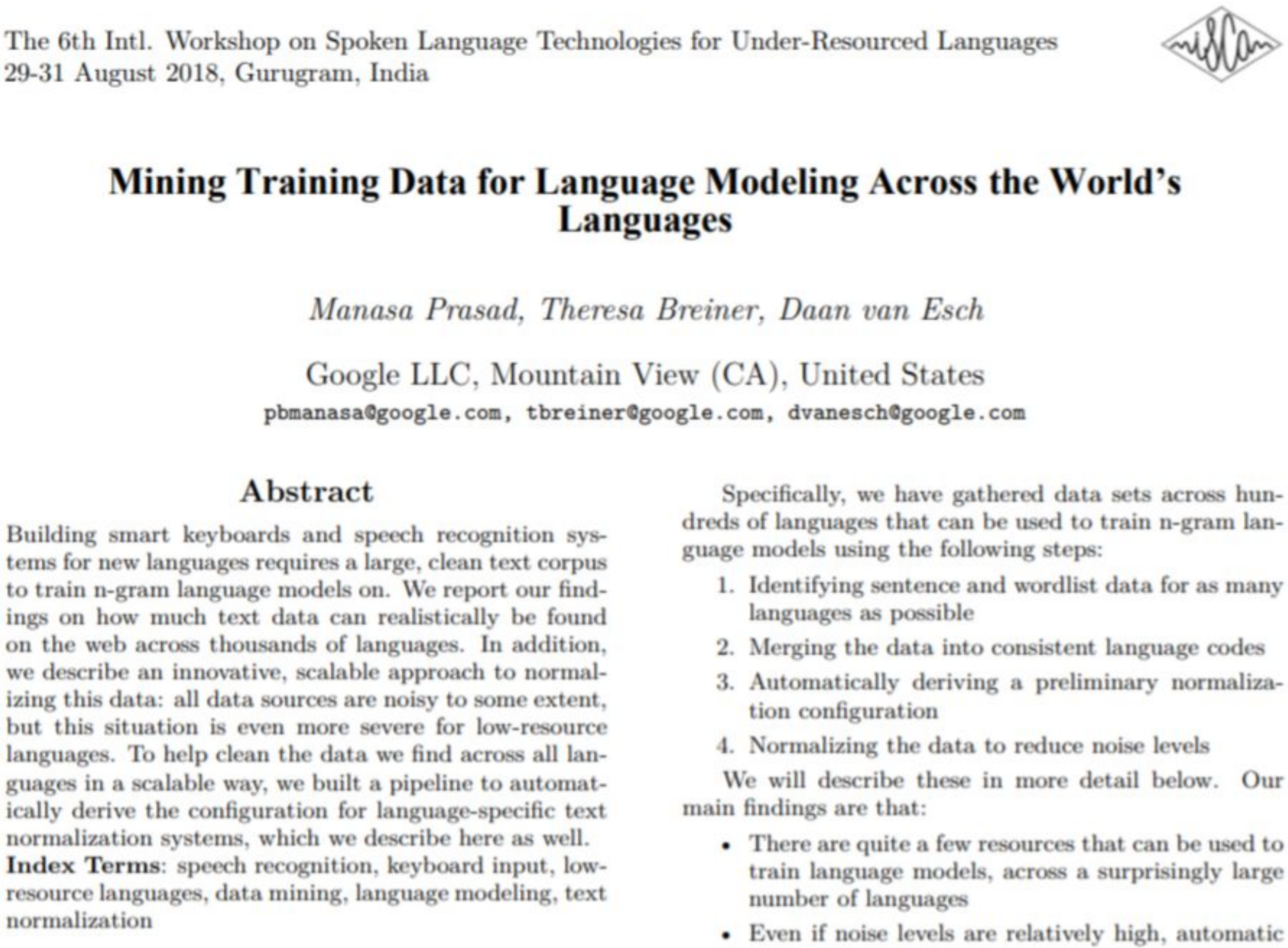
1. Introduction

Today, language technologies are easily available in only a small minority of the world's 7,000+ language varieties. For example, technologies like automatic speech recognition and neural machine translation are available from commercial vendors in about 100 language varieties; even keyboards and spell-checkers, which are relatively straightforward to develop, are only available in about 1,000–1,500 varieties (Mager et al., 2018; van Esch et al., 2019; Kuhn et al., 2020).

	# of language varieties	Speaker data	Writing system data	Open-source
Wikipedia list	100	✓	✗	✓
ISO 639-3	7,893	✗	✗	✓
Glottolog	8,549	✗	✗	✓
Ethnologue	7,459	✓	✗	✗
WALS	2,662	✗	✗	✓
Ours	2,831	✓	✓	✓

Table 1: Number of languages and information available in existing language resources compared to ours.

(Will be presented at LREC 2022 in Marseille in June and posted to GitHub)



The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Mining Training Data for Language Modeling Across the World's Languages

Manasa Prasad, Theresa Breiner, Daan van Esch

Google LLC, Mountain View (CA), United States

pbmanasa@google.com, tbreiner@google.com, dvanesch@google.com

Abstract

Building smart keyboards and speech recognition systems for new languages requires a large, clean text corpus to train n-gram language models on. We report our findings on how much text data can realistically be found on the web across thousands of languages. In addition, we describe an innovative, scalable approach to normalizing this data: all data sources are noisy to some extent, but this situation is even more severe for low-resource languages. To help clean the data we find across all languages in a scalable way, we built a pipeline to automatically derive the configuration for language-specific text normalization systems, which we describe here as well.

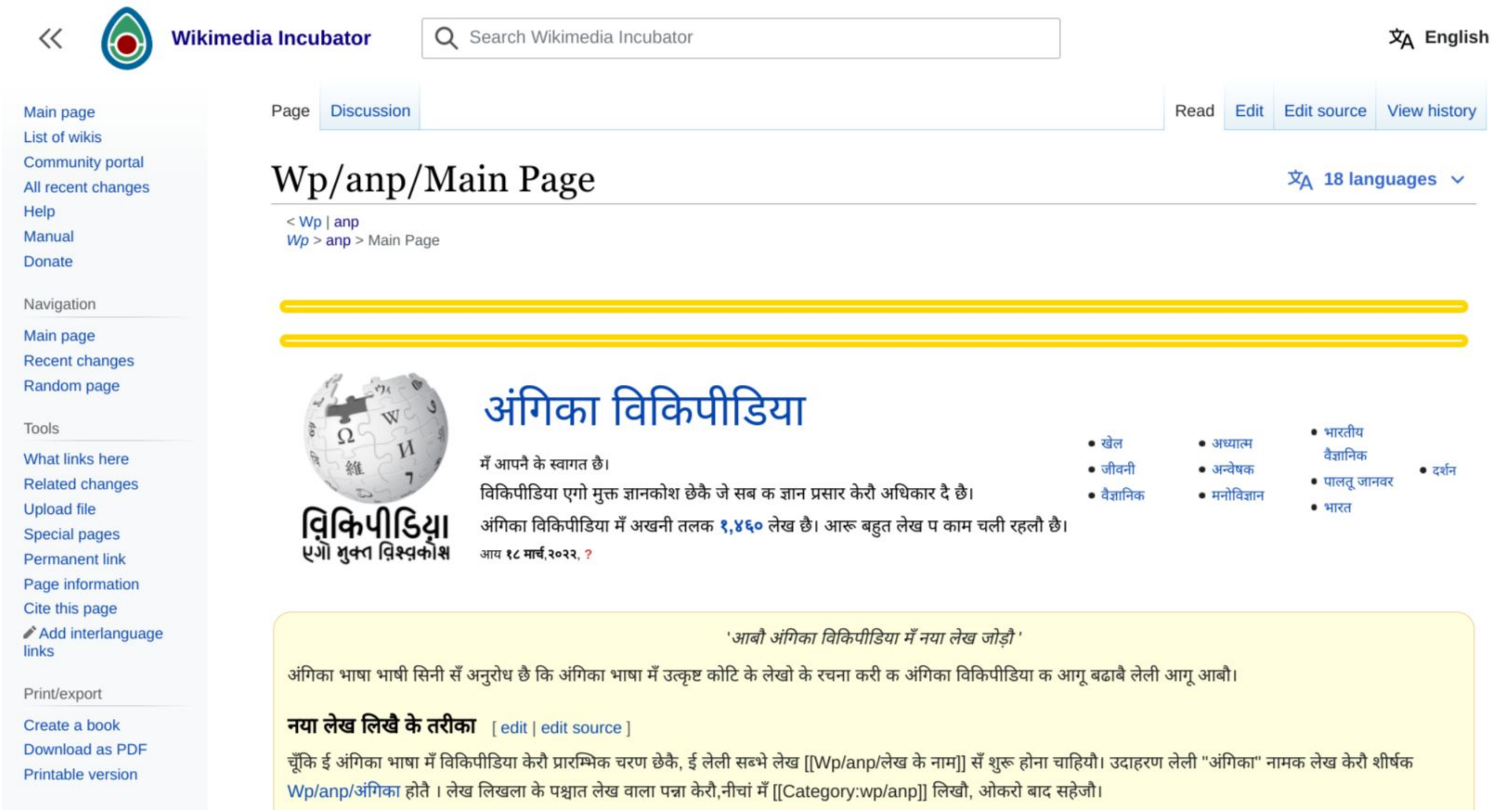
Index Terms: speech recognition, keyboard input, low-resource languages, data mining, language modeling, text normalization

Specifically, we have gathered data sets across hundreds of languages that can be used to train n-gram language models using the following steps:

1. Identifying sentence and wordlist data for as many languages as possible
2. Merging the data into consistent language codes
3. Automatically deriving a preliminary normalization configuration
4. Normalizing the data to reduce noise levels

We will describe these in more detail below. Our main findings are that:

- There are quite a few resources that can be used to train language models, across a surprisingly large number of languages
- Even if noise levels are relatively high, automatic



Wp/anp/Main Page

अंगिका विकिपीडिया

मँ आपनै के स्वागत छै।

विकिपीडिया एगो मुक्त ज्ञानकोश छेकै जे सब क ज्ञान प्रसार केरौ अधिकार दै छै।

अंगिका विकिपीडिया मँ अखनी तलक १,४६० लेख छै। आरू बहुत लेख प काम चली रहलौ छै।

आय १८ मार्च,२०२२, ?

- खेल
- जीवनी
- वैज्ञानिक
- अध्यात्म
- अन्वेषक
- मनोविज्ञान
- भारतीय वैज्ञानिक
- पालतू जानवर
- भारत
- दर्शन

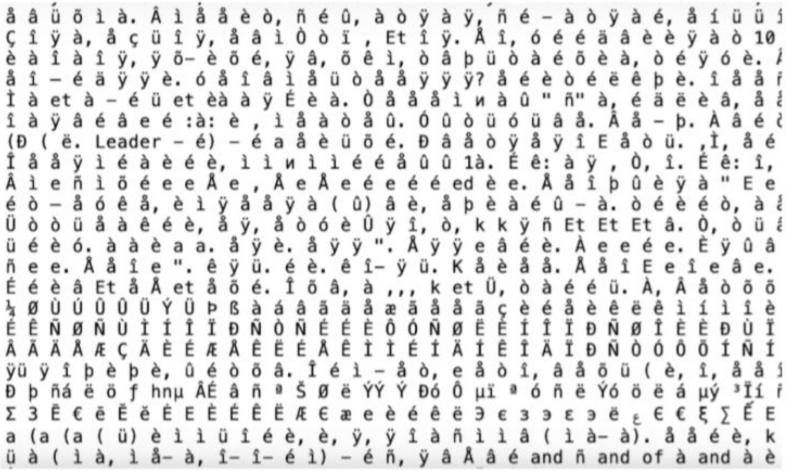
'आबौ अंगिका विकिपीडिया मँ नया लेख जोड़ौ'

अंगिका भाषा भाषी सिनी सँ अनुरोध छै कि अंगिका भाषा मँ उत्कृष्ट कोटि के लेखो के रचना करी क अंगिका विकिपीडिया क आगू बढाबै लेली आगू आबौ।

नया लेख लिखै के तरीका [edit | edit source]

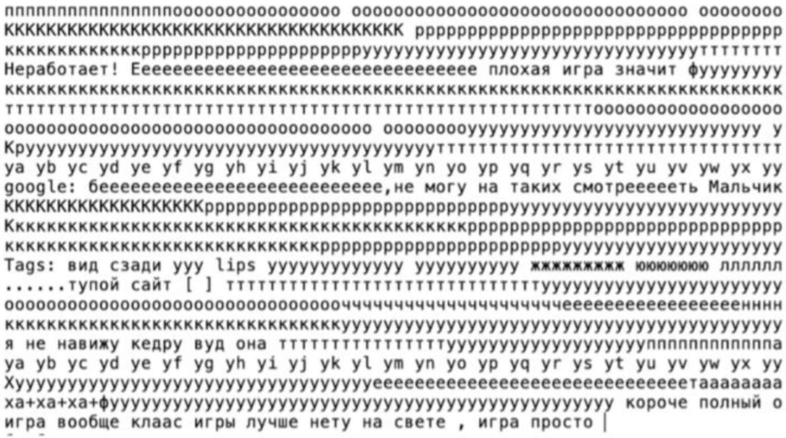
चूँकि ई अंगिका भाषा मँ विकिपीडिया केरौ प्रारम्भिक चरण छेकै, ई लेली सब्भे लेख [[Wp/anp/लेख के नाम]] सँ शुरू होना चाहियौ। उदाहरण लेली "अंगिका" नामक लेख केरौ शीर्षक Wp/anp/अंगिका होतै । लेख लिखला के पश्चात लेख वाला पन्ना केरौ,नीचां मँ [[Category:wp/anp]] लिखौ, ओकरो बाद सहेजौ।

(Source: Wikimedia Incubator for Angika, a language of India)



```
å â ü õ ì à. Â ì å å è ò, ñ é û, à ò ÿ à ÿ, ñ é – à ò ÿ à é, å í ü ü í
Ç î ÿ à, å ç ü î ÿ, å â ì Ò ò ï , Et î ÿ. Å î, ó é é ä â è è ÿ à ò 10
è à î à î ÿ, ÿ õ– è õ é, ÿ â, õ ê ì, ò â þ ü ò à é õ è à, ò é ÿ ó è. Å
å î – é ä ÿ ÿ è. ó å î â ì å ü ò å å ÿ ÿ ÿ? å é è ò é ë ê þ è. î å å ŕ
Ì à et à – é ü et èà à ÿ É è à. Ò å å å ì и à û " ñ" à, é ä ë è â, å ä
î à ÿ â é â e é :à: è , ì å à ò å û. Ó û ò ü ó ü â å. Å å – þ. À â é ¿
(Ð ( ë. Leader – é) – é a å è ü õ é. Ð â å ò ÿ å ÿ î E å ò ü. ,Ì, å é
Î å å ÿ ì é à è é è, ì ì и ì ì é é å û û 1à. É ê: à ÿ , Ò, î. É ê: î,
Â ì e ñ ì õ é e e Å e , Å e Å e é e é é ed è e. Å å î þ û è ÿ à " E e
é ò – å ó ê å, è ì ÿ å å ÿ à ( û) â è, å þ è à é û – à. ò é è é ò, à ä
Ü ò ò ü å à ê é è, å ÿ, å ò ó è Û ÿ î, ò, k k ÿ ñ Et Et Et â. Ò, ò ü ä
ü é è ó. à à è a a. å ÿ è. å ÿ ÿ ". Å ÿ ÿ e â é è. À e e é e. È ÿ û â
ñ e e. Å å î e ". ê ÿ ü. é è. ê î– ÿ ü. K å è å å. Å å î E e î e â e.
É é è â Et å Å et å õ é. Î õ â, à ,,, k et Ü, ò à é é ü. À, Å å ò õ õ
¾ Ø Ù Ú Û Û Ü Ý Ü Þ ß à á â ã ä å æ ã å å ã ç è é å è ê ë ê ì í ì î è
É Ê Ñ Ø Ñ Ù Ì Í Î Ï Ð Ñ Ò Ñ É É È Ô Ó Ñ Ø Ë Ë Í Î Ï Ð Ñ Ø Î È È Ð Ù Ï
Â Ã Ä Å Æ Ç Ä È É Æ Å Ê Ë É Å Ê Ì Ì É Í Ä Í Ê Î Ä Ï Ð Ñ Ò Ó Ô Õ Í Ñ Í
ÿü ÿ î þ è þ è, û é ò õ â. Î é ì – å ò, e å ò î, â å õ ü ( è, î, å å í
Ð þ ñá ë ö ƒ hnµ ÂÉ â ñ ª Š ø ë ÝÝ Ý Ðó Ô µï ª ó ñ ë Ýó ö ë á µý ³Ïí ŕ
Σ 3 Ē € ē Ĕ ĕ Ė E È É Ê Ë Æ Є æ e è é ê ë Э є з э ɛ э ë ͼ Є € ξ ∑ Ḗ E
a (a (a ( ü) è ì ì ü î é è, è, ÿ, ÿ î à ñ ì ì â ( ì à– à). å å é è, k
ü à ( ì à, ì å– à, î– î– é ì) – é ñ, ÿ â Å â é and ñ and of à and à è
```

(a) “Neopolitan” (actually A N T S P E A K - like content)

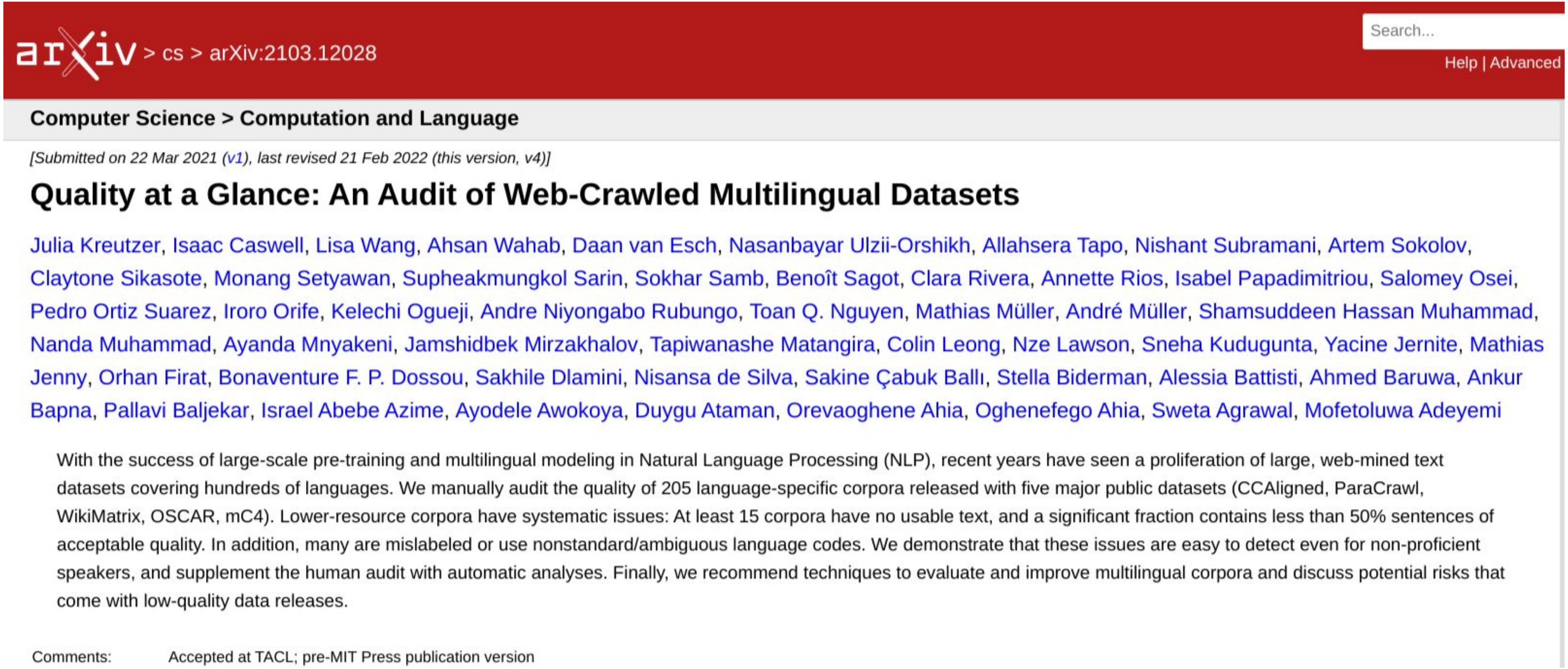


```
ппппппппппппппппoooooooooooooooo oooooooooooooooooooooooooooooooo oooooooo
KKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKK pppppppppppppppppppppppppppppppppp
kkkkkkkkkkkkpppppppppppppppppppppppyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyттттттт
Неработает! Eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee плохая игра значит фyyyyyyyy
kkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkk
ттттттттттттттттттттттттттттттттттттттттттттттттттттттттoooooooooooooooooo
ooooooooooooooooooooooooooooooooooo ooooooooyyyyyyyyyyyyyyyyyyyyyyyyyy y
Kpyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyттттттттттттттттттттттттттттттттт
ya yb yc yd ye yf yg yh yi yj yk yl ym yn yo yp yq yr ys yt yu yv yw yx yy
google: бeeeeeeeeeeeeeeeeeeeeeeeeeeee,не могу на таких смотреееееть Мальчик
KKKKKKKKKKKKKKKKKKpppppppppppppppppppppppppppppppyyyyyyyyyyyyyyyyyyyyyyyy
Kkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkppppppppppppppppppppppppppp
kkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkpppppppppppppppppppppppppyyyyyyyyyyyyyyyyyyyy
Tags: вид сзади yyy lips yyyyyyyyyyyyy yyyyyyyyyy жжжжжжжж юююююю лллллл
......тупой сайт [ ] тттттттттттттттттттттттттттттттyyyyyyyyyyyyyyyyyyyyyy
ooooooooooooooooooooooooooooooooччччччччччччччччччччeeeeeeeeeeeeeeeeeeнннн
kkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
я не навижу кедру вуд она тттттттттттттттттyyyyyyyyyyyyyyyyyyyyпппппппппппа
ya yb yc yd ye yf yg yh yi yj yk yl ym yn yo yp yq yr ys yt yu yv yw yx yy
Xyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeтaaaaaaa
xa+xa+xa+фyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy короче полный о
игра вообще клаас игры лучше нету на свете , игра просто |
```

(b) “Somali” (actually repeated ngraaaaaaaaams)

Figure 1: Representative samples from OSCAR corpora affected by two n-gram LangID error modes

Fortunately newer version of OSCAR improves the data quality quite significantly!
Many thanks to the OSCAR team :)



Computer Science > Computation and Language

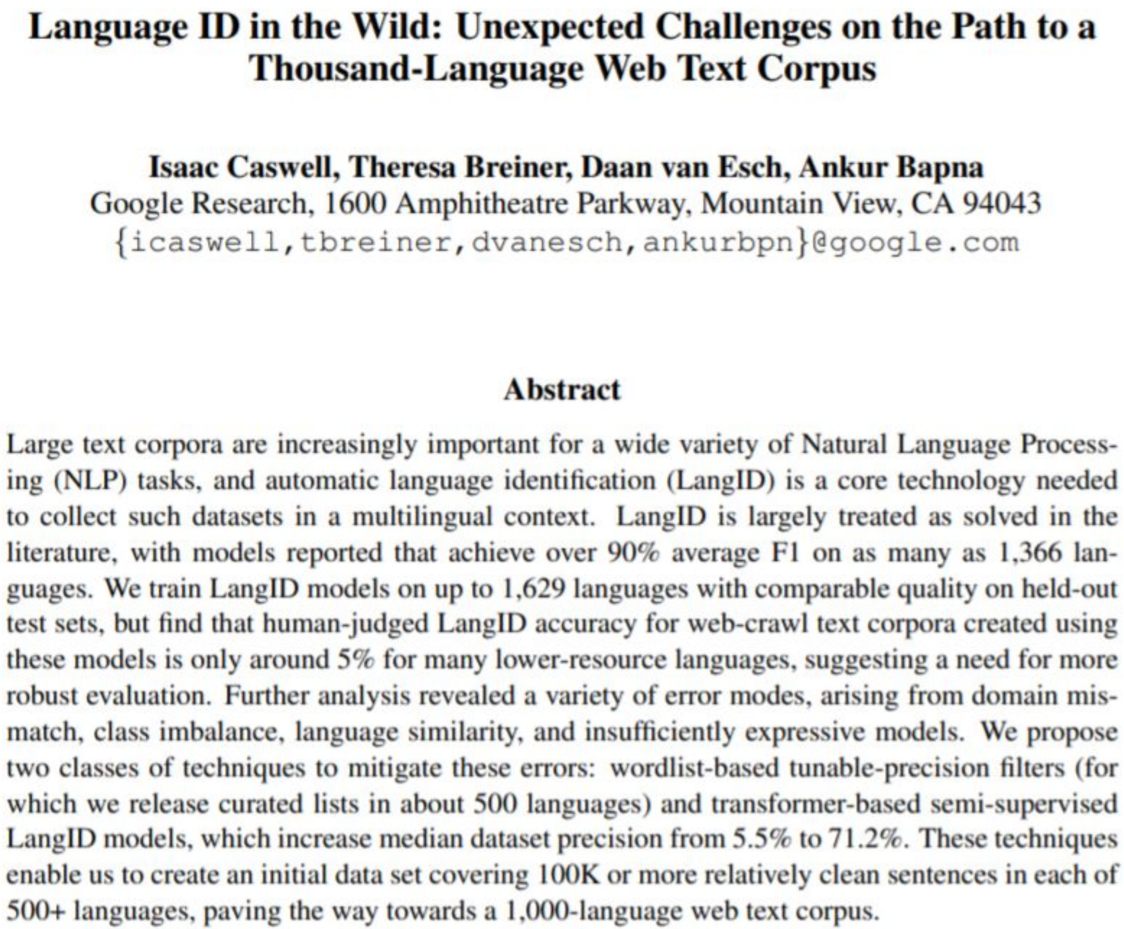
[Submitted on 22 Mar 2021 (v1), last revised 21 Feb 2022 (this version, v4)]

Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets

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, Mofetoluwa Adeyemi

With the success of large-scale pre-training and multilingual modeling in Natural Language Processing (NLP), recent years have seen a proliferation of large, web-mined text datasets covering hundreds of languages. We manually audit the quality of 205 language-specific corpora released with five major public datasets (CCAligned, ParaCrawl, WikiMatrix, OSCAR, mC4). Lower-resource corpora have systematic issues: At least 15 corpora have no usable text, and a significant fraction contains less than 50% sentences of acceptable quality. In addition, many are mislabeled or use nonstandard/ambiguous language codes. We demonstrate that these issues are easy to detect even for non-proficient speakers, and supplement the human audit with automatic analyses. Finally, we recommend techniques to evaluate and improve multilingual corpora and discuss potential risks that come with low-quality data releases.

Comments: Accepted at TACL; pre-MIT Press publication version

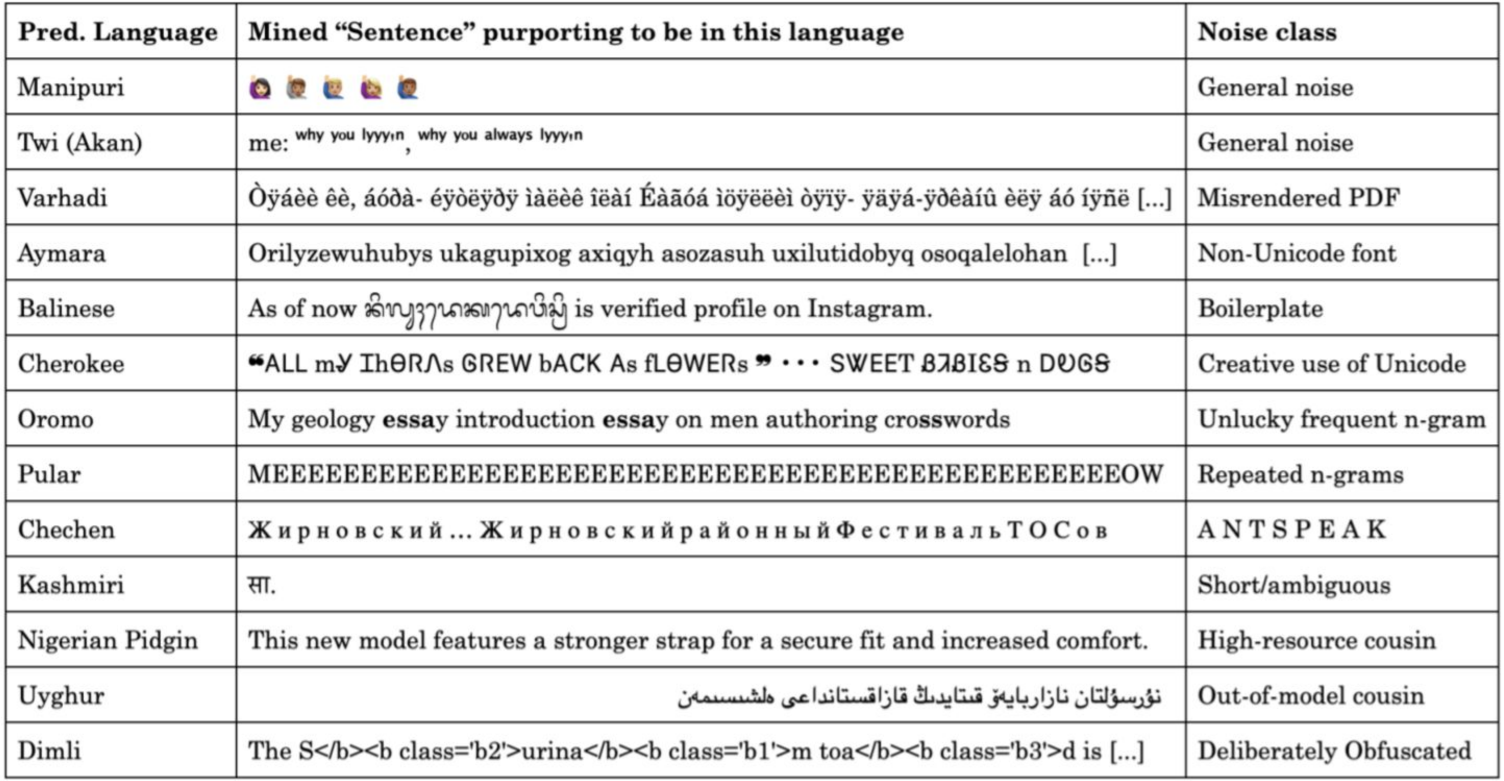


Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus

Isaac Caswell, Theresa Breiner, Daan van Esch, Ankur Bapna
Google Research, 1600 Amphitheatre Parkway, Mountain View, CA 94043
{icaswell,tbreiner,dvanesch,ankurbpn}@google.com

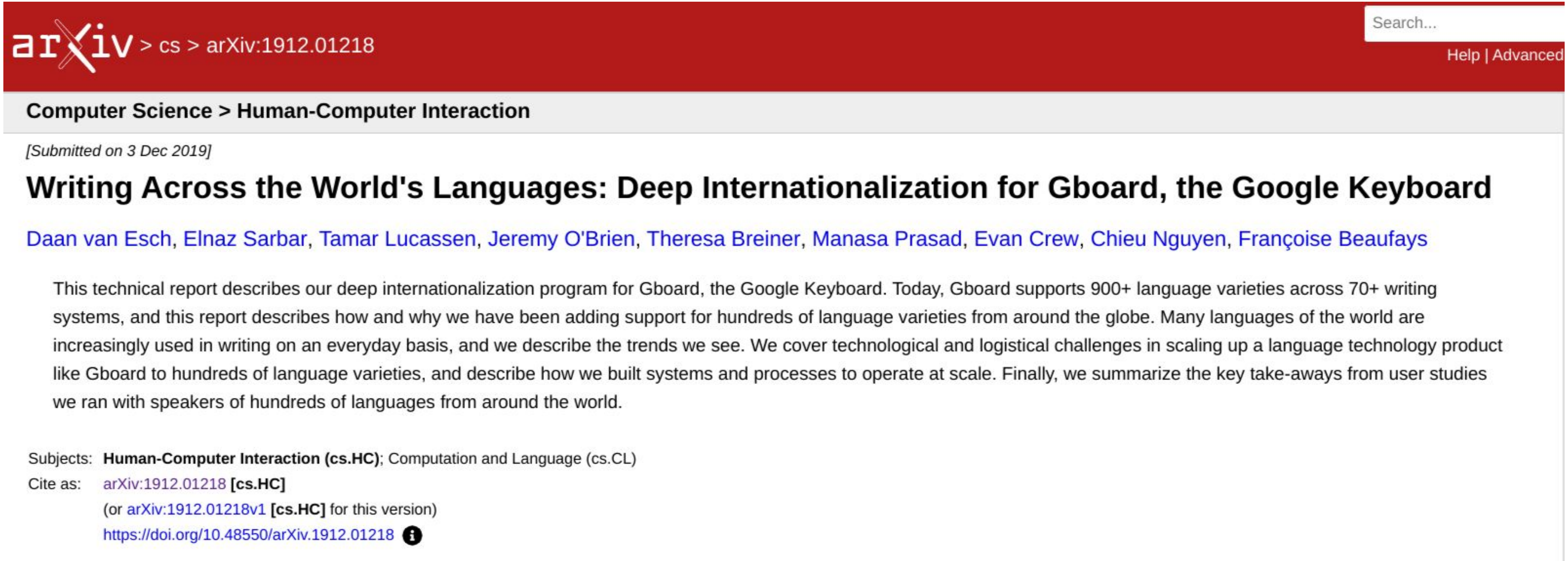
Abstract

Large text corpora are increasingly important for a wide variety of Natural Language Processing (NLP) tasks, and automatic language identification (LangID) is a core technology needed to collect such datasets in a multilingual context. LangID is largely treated as solved in the literature, with models reported that achieve over 90% average F1 on as many as 1,366 languages. We train LangID models on up to 1,629 languages with comparable quality on held-out test sets, but find that human-judged LangID accuracy for web-crawl text corpora created using these models is only around 5% for many lower-resource languages, suggesting a need for more robust evaluation. Further analysis revealed a variety of error modes, arising from domain mismatch, class imbalance, language similarity, and insufficiently expressive models. We propose two classes of techniques to mitigate these errors: wordlist-based tunable-precision filters (for which we release curated lists in about 500 languages) and transformer-based semi-supervised LangID models, which increase median dataset precision from 5.5% to 71.2%. These techniques enable us to create an initial data set covering 100K or more relatively clean sentences in each of 500+ languages, paving the way towards a 1,000-language web text corpus.



Pred. Language	Mined “Sentence” purporting to be in this language	Noise class
Manipuri	🙋🏻‍♀️ 🙋🏽‍♂️ 🙋🏼‍♂️ 🙋🏼‍♀️ 🙋🏽‍♂️	General noise
Twi (Akan)	me: why you lyyyın, why you always lyyyın	General noise
Varhadi	Òÿáèè êè, áóðà- éÿòëÿðÿ ìàëèê îëàí Éàãóá ìöÿëëèì òÿïÿ- ÿäÿá-ÿðêàíû èëÿ áó íÿñë [...]	Misrendered PDF
Aymara	Orilyzewuhubys ukagupixog axiqyh asozasuh uxilutidobyq osoqalelohan [...]	Non-Unicode font
Balinese	As of now [illegible] is verified profile on Instagram.	Boilerplate
Cherokee	❝ALL mᎩ ᏆhᎾᏒᐱs ᏀREW bACK As fᏞᎾWERs ❞ · · · SᎳEET ᏰᏘᏰIᏋᏕ n DᎤᏩᏕ	Creative use of Unicode
Oromo	My geology **essa**y introduction **essa**y on men authoring cro**ss**words	Unlucky frequent n-gram
Pular	MEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEOW	Repeated n-grams
Chechen	Ж и р н о в с к и й ... Ж и р н о в с к и й р а й о н н ы й Ф е с т и в а л ь Т О С о в	A N T S P E A K
Kashmiri	सा.	Short/ambiguous
Nigerian Pidgin	This new model features a stronger strap for a secure fit and increased comfort.	High-resource cousin
Uyghur	نۇرسۇلتان نازاربايەۋ قىتايدىڭ قازاقستانداعى ٴلشىسىمەن	Out-of-model cousin
Dimli	The S</b><b class='b2'>urina</b><b class='b1'>m toa</b><b class='b3'>d is [...]	Deliberately Obfuscated

Table 2: Examples of several representative classes of noise in our initial web-crawl corpora.



Computer Science > Human-Computer Interaction

[Submitted on 3 Dec 2019]

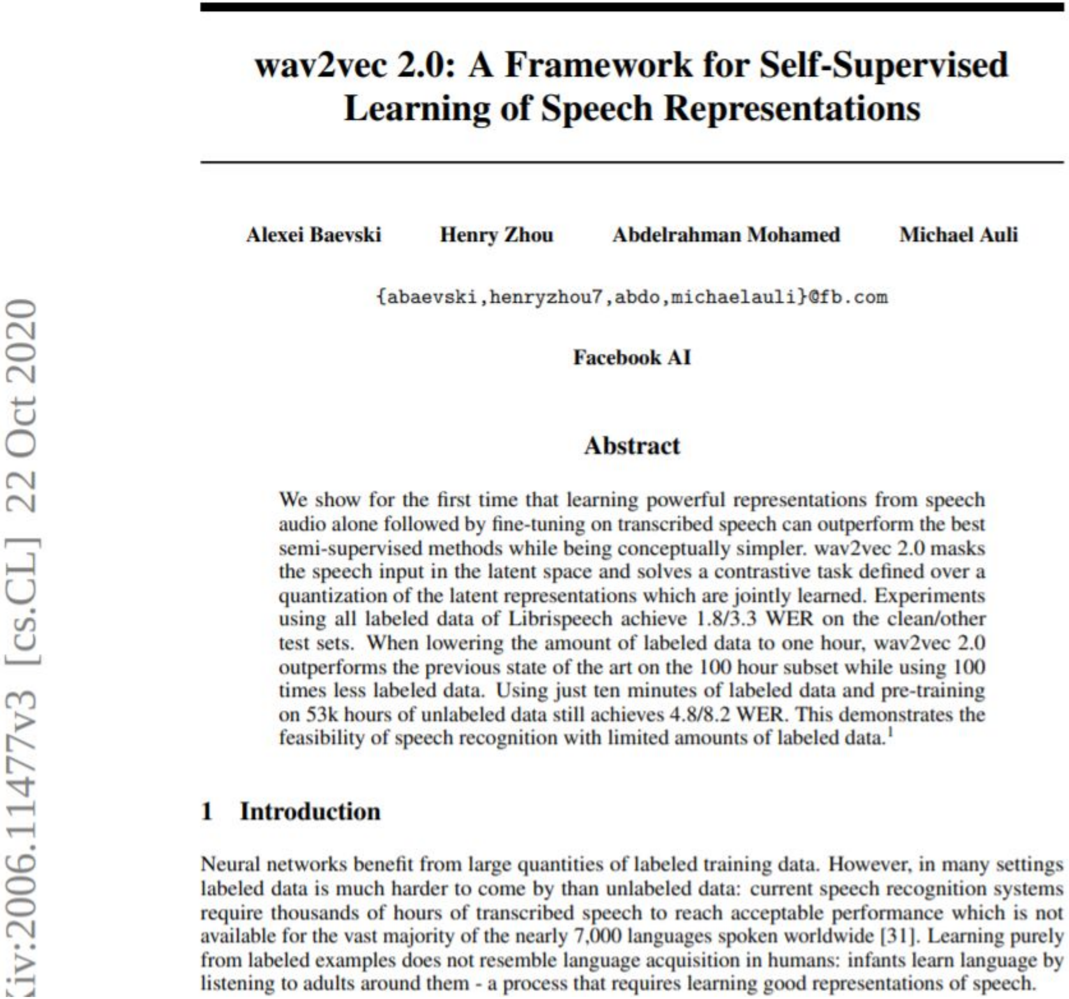
Writing Across the World's Languages: Deep Internationalization for Gboard, the Google Keyboard

Daan van Esch, Elnaz Sarbar, Tamar Lucassen, Jeremy O'Brien, Theresa Breiner, Manasa Prasad, Evan Crew, Chieu Nguyen, Françoise Beaufays

This technical report describes our deep internationalization program for Gboard, the Google Keyboard. Today, Gboard supports 900+ language varieties across 70+ writing systems, and this report describes how and why we have been adding support for hundreds of language varieties from around the globe. Many languages of the world are increasingly used in writing on an everyday basis, and we describe the trends we see. We cover technological and logistical challenges in scaling up a language technology product like Gboard to hundreds of language varieties, and describe how we built systems and processes to operate at scale. Finally, we summarize the key take-aways from user studies we ran with speakers of hundreds of languages from around the world.

Subjects: **Human-Computer Interaction (cs.HC)**; Computation and Language (cs.CL)
Cite as: arXiv:1912.01218 **[cs.HC]**
(or arXiv:1912.01218v1 **[cs.HC]** for this version)
https://doi.org/10.48550/arXiv.1912.01218

Bringing ASR to more languages



wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

Alexei Baevski **Henry Zhou** **Abdelrahman Mohamed** **Michael Auli**

{abaevski,henryzhou7,abdo,michaelauli}@fb.com

Facebook AI

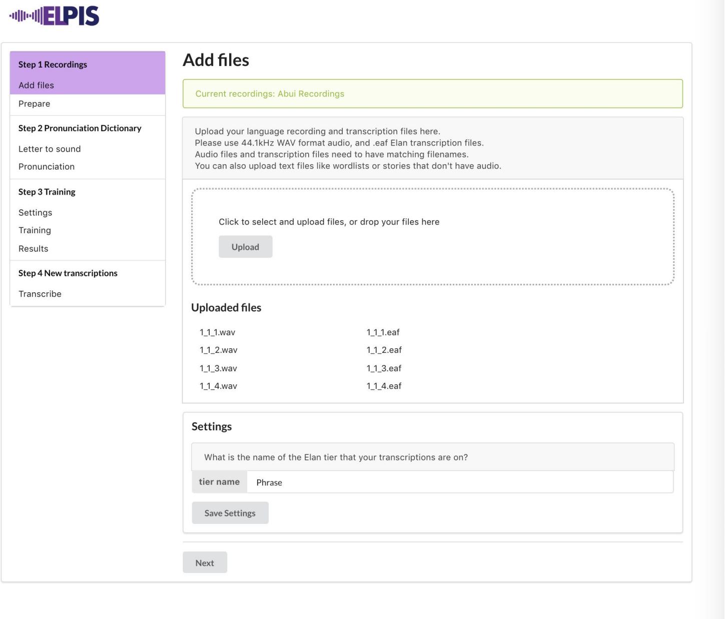
Abstract

We show for the first time that learning powerful representations from speech audio alone followed by fine-tuning on transcribed speech can outperform the best semi-supervised methods while being conceptually simpler. wav2vec 2.0 masks the speech input in the latent space and solves a contrastive task defined over a quantization of the latent representations which are jointly learned. Experiments using all labeled data of Librispeech achieve 1.8/3.3 WER on the clean/other test sets. When lowering the amount of labeled data to one hour, wav2vec 2.0 outperforms the previous state of the art on the 100 hour subset while using 100 times less labeled data. Using just ten minutes of labeled data and pre-training on 53k hours of unlabeled data still achieves 4.8/8.2 WER. This demonstrates the feasibility of speech recognition with limited amounts of labeled data.[1]

1 Introduction

Neural networks benefit from large quantities of labeled training data. However, in many settings labeled data is much harder to come by than unlabeled data: current speech recognition systems require thousands of hours of transcribed speech to reach acceptable performance which is not available for the vast majority of the nearly 7,000 languages spoken worldwide [31]. Learning purely from labeled examples does not resemble language acquisition in humans: infants learn language by listening to adults around them - a process that requires learning good representations of speech.

Elpis: Speech Tech in a User-Friendly GUI



Add files

Current recordings: Abui Recordings

Upload your language recording and transcription files here.
Please use 44.1kHz WAV format audio, and .eaf Elan transcription files.
Audio files and transcription files need to have matching filenames.
You can also upload text files like wordlists or stories that don't have audio.

Click to select and upload files, or drop your files here

Uploaded files

1_1_1.wav	1_1_1.eaf
1_1_2.wav	1_1_2.eaf
1_1_3.wav	1_1_3.eaf
1_1_4.wav	1_1_4.eaf

Settings

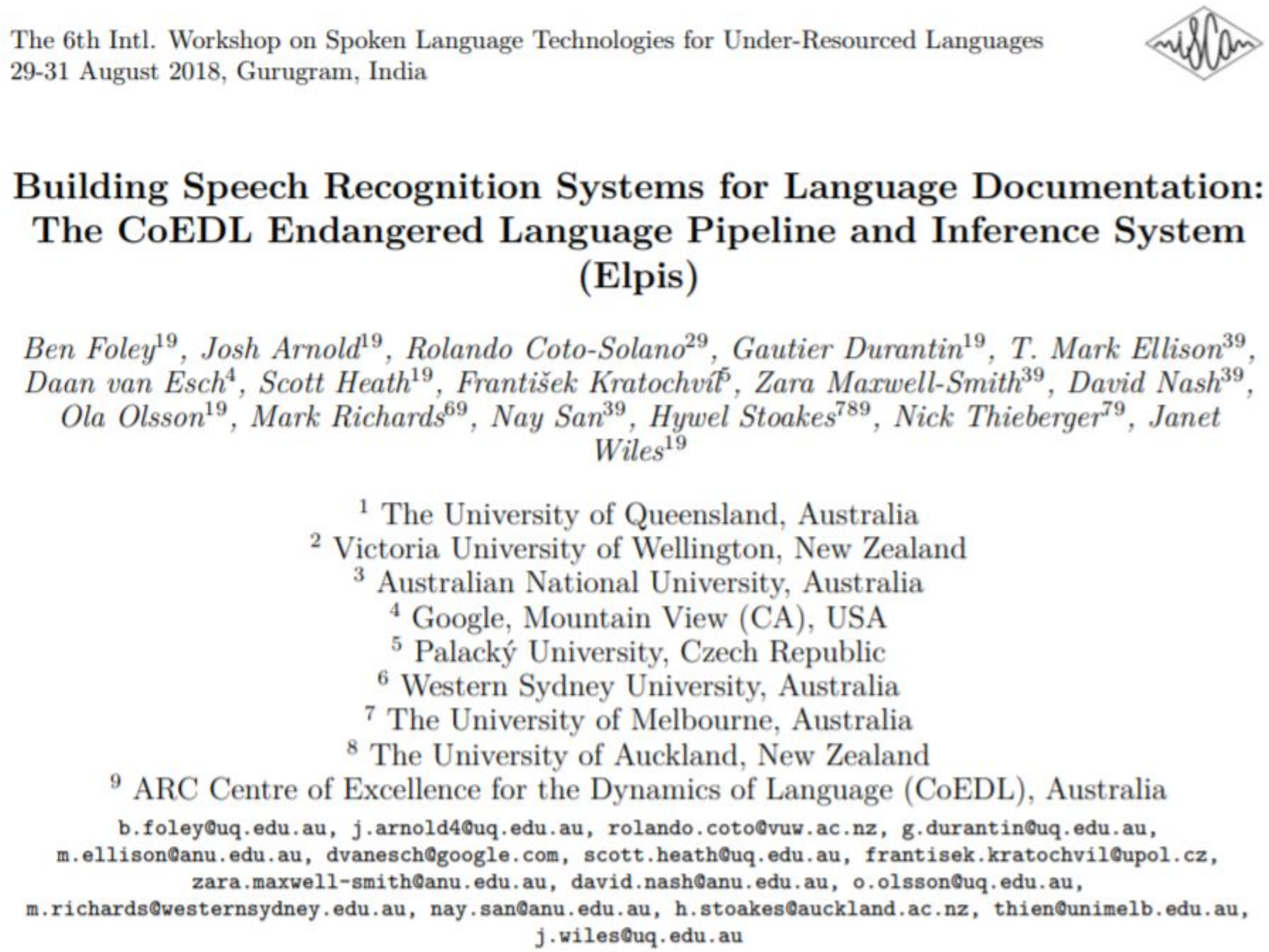
What is the name of the Elan tier that your transcriptions are on?

tier name: Phrase

An **easy-to-use** open-source speech toolkit for **fieldwork linguists & communities**
Promising results across many languages, even with relatively little data (~2 hours). Runs **on your laptop!**

Built by the Australian **Centre of Excellence for the Dynamics of Language**, with help from Google
Learn more at github.com/CoEDL/elpis

(Source: Elpis user interface and elpis.readthedocs.io)



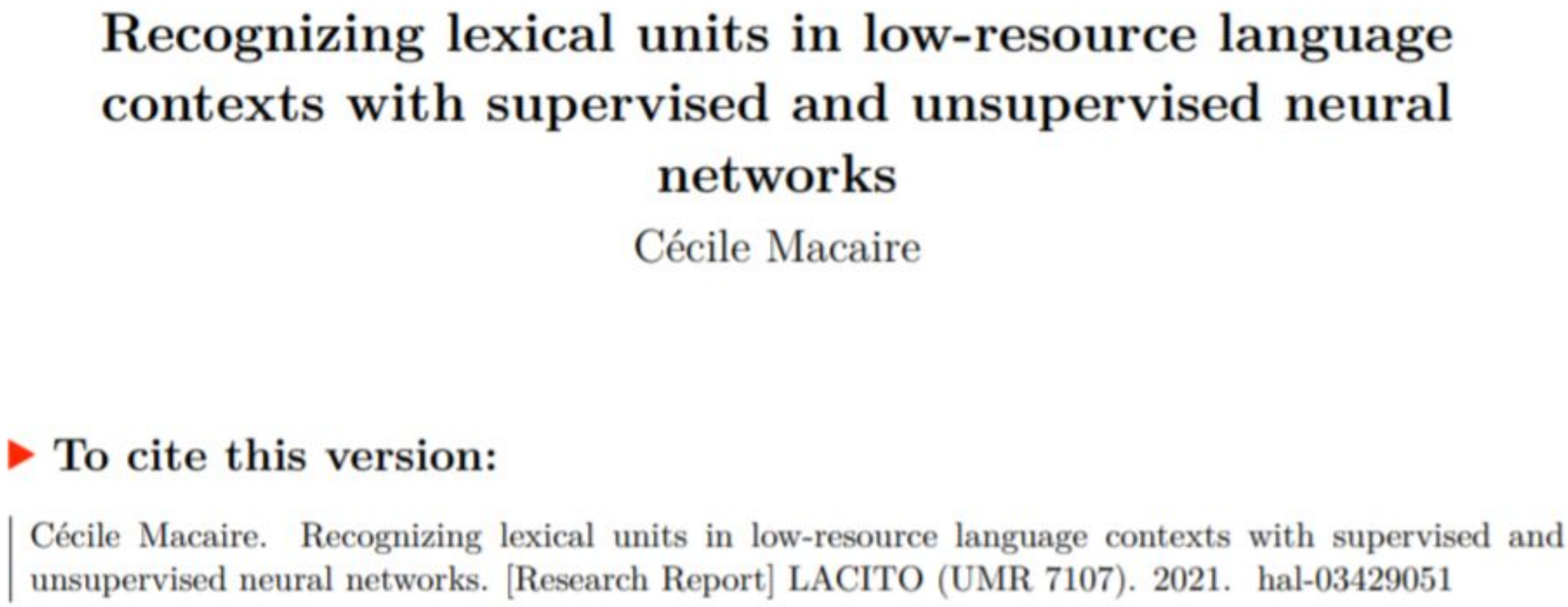
The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (Elpis)

Ben Foley[19], *Josh Arnold*[19], *Rolando Coto-Solano*[29], *Gautier Durantin*[19], *T. Mark Ellison*[39], *Daan van Esch*[4], *Scott Heath*[19], *František Kratochvíl*[5], *Zara Maxwell-Smith*[39], *David Nash*[39], *Ola Olsson*[19], *Mark Richards*[69], *Nay San*[39], *Hywel Stoakes*[789], *Nick Thieberger*[79], *Janet Wiles*[19]

[1] The University of Queensland, Australia
[2] Victoria University of Wellington, New Zealand
[3] Australian National University, Australia
[4] Google, Mountain View (CA), USA
[5] Palacký University, Czech Republic
[6] Western Sydney University, Australia
[7] The University of Melbourne, Australia
[8] The University of Auckland, New Zealand
[9] ARC Centre of Excellence for the Dynamics of Language (CoEDL), Australia

b.foley@uq.edu.au, j.arnold4@uq.edu.au, rolando.coto@vuw.ac.nz, g.durantin@uq.edu.au, m.ellison@anu.edu.au, dvanesch@google.com, scott.heath@uq.edu.au, frantisek.kratochvil@upol.cz, zara.maxwell-smith@anu.edu.au, david.nash@anu.edu.au, o.olsson@uq.edu.au, m.richards@westernsydney.edu.au, nay.san@anu.edu.au, h.stoakes@auckland.ac.nz, thien@unimelb.edu.au, j.wiles@uq.edu.au

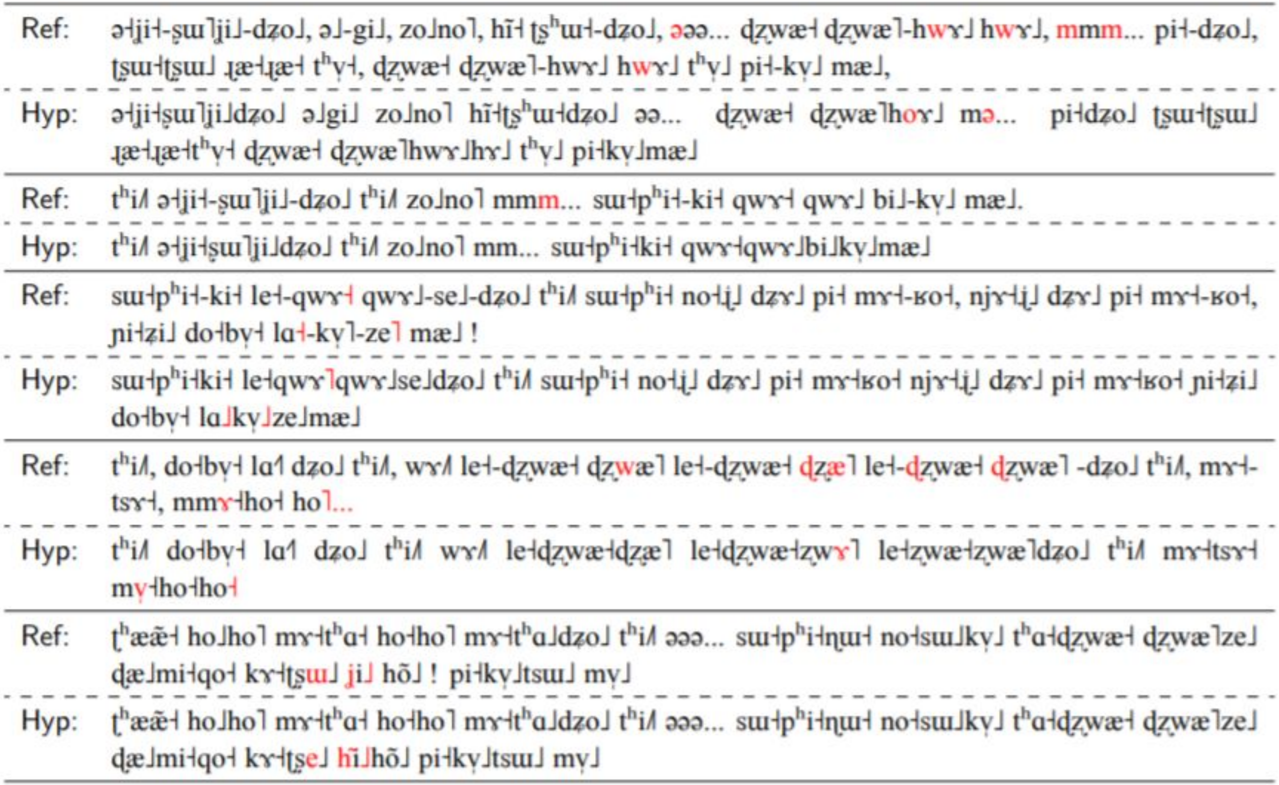


Recognizing lexical units in low-resource language contexts with supervised and unsupervised neural networks

Cécile Macaire

To cite this version:

> Cécile Macaire. Recognizing lexical units in low-resource language contexts with supervised and unsupervised neural networks. [Research Report] LACITO (UMR 7107). 2021. hal-03429051



Ref:	ə˧ʝi˧-ʂɯ˥ʝi˩-dʑo˩, ə˩-gi˩, zo˩no˥, hĩ˧ ʈʂʰɯ˧-dʑo˩, əəə... ɖʐwæ˧ ɖʐwæ˥-hwɤ˩ hwɤ˩, mmm... pi˧-dʑo˩, ʈʂɯ˧ʈʂɯ˩ ɻæ˧ɻæ˧ tʰv̩˧, ɖʐwæ˧ ɖʐwæ˥-hwɤ˩ hwɤ˩ tʰv̩˩ pi˧-kv̩˩ mæ˩,
Hyp:	ə˧ʝi˧ʂɯ˥ʝi˩dʑo˩ ə˩gi˩ zo˩no˥ hĩ˧ʈʂʰɯ˧dʑo˩ əə... ɖʐwæ˧ ɖʐwæ˥hoɤ˩ mə... pi˧dʑo˩ ʈʂɯ˧ʈʂɯ˩ ɻæ˧ɻæ˧tʰv̩˧ ɖʐwæ˧ ɖʐwæ˥hwɤ˩hɤ˩ tʰv̩˩ pi˧kv̩˩mæ˩
Ref:	tʰi˩˥ ə˧ʝi˧-ʂɯ˥ʝi˩-dʑo˩ tʰi˩˥ zo˩no˥ mmm... sɯ˧pʰi˧-ki˧ qwɤ˧ qwɤ˩ bi˩-kv̩˩ mæ˩.
Hyp:	tʰi˩˥ ə˧ʝi˧ʂɯ˥ʝi˩dʑo˩ tʰi˩˥ zo˩no˥ mm... sɯ˧pʰi˧ki˧ qwɤ˧qwɤ˩bi˩kv̩˩mæ˩
Ref:	sɯ˧pʰi˧-ki˧ le˧-qwɤ˧ qwɤ˩-se˩-dʑo˩ tʰi˩˥ sɯ˧pʰi˧ no˧ɭ̩˩ dʑɤ˩ pi˧ mɤ˧-ʁo˧, ɲɟɤ˧ɭ̩˩ dʑɤ˩ pi˧ mɤ˧-ʁo˧, ɲi˧ʑi˩ do˧bv̩˧ lɑ˧-kv̩˥-ze˥ mæ˩ !
Hyp:	sɯ˧pʰi˧ki˧ le˧qwɤ˥qwɤ˩se˩dʑo˩ tʰi˩˥ sɯ˧pʰi˧ no˧ɭ̩˩ dʑɤ˩ pi˧ mɤ˧ʁo˧ ɲɟɤ˧ɭ̩˩ dʑɤ˩ pi˧ mɤ˧ʁo˧ ɲi˧ʑi˩ do˧bv̩˧ lɑ˩kv̩˩ze˩mæ˩
Ref:	tʰi˩˥, do˧bv̩˧ lɑ˧˥ dʑo˩ tʰi˩˥, wɤ˩˥ le˧-ɖʐwæ˧ ɖʐwæ˥ le˧-ɖʐwæ˧ ɖʐæ˥ le˧-ɖʐwæ˧ ɖʐwæ˥ -dʑo˩ tʰi˩˥, mɤ˧-tsɤ˧, mmɤ˧ho˧ ho˥...
Hyp:	tʰi˩˥ do˧bv̩˧ lɑ˧˥ dʑo˩ tʰi˩˥ wɤ˩˥ le˧ɖʐwæ˧ɖʐæ˥ le˧ɖʐwæ˧ʐwɤ˥ le˧ʐwæ˧ʐwæ˥dʑo˩ tʰi˩˥ mɤ˧tsɤ˧ mv̩˧ho˧ho˧
Ref:	ʈʰææ̃˧ ho˩ho˥ mɤ˧tʰɑ˧ ho˧ho˥ mɤ˧tʰɑ˩dʑo˩ tʰi˩˥ əəə... sɯ˧pʰi˧ɳɯ˧ no˧sɯ˩kv̩˩ tʰɑ˧ɖʐwæ˧ ɖʐwæ˥ze˩ ɖæ˩mi˧qo˧ kɤ˧ʈʂɯ˩ ʝi˩ hõ˩ ! pi˧kv̩˩tsɯ˩ mv̩˩
Hyp:	ʈʰææ̃˧ ho˩ho˥ mɤ˧tʰɑ˧ ho˧ho˥ mɤ˧tʰɑ˩dʑo˩ tʰi˩˥ əəə... sɯ˧pʰi˧ɳɯ˧ no˧sɯ˩kv̩˩ tʰɑ˧ɖʐwæ˧ ɖʐwæ˥ze˩ ɖæ˩mi˧qo˧ kɤ˧ʈʂe˩ hĩ˩hõ˩ pi˧kv̩˩tsɯ˩ mv̩˩

Table 3.11: Samples of the predicted transcriptions by the xlsr-na-180 model of the "Appeal to the gods to settle a quarrel" speech file. In red, the deletions, insertions and substitutions.

(Source for this visual and the map on the preceding slide: Macaire 2021)

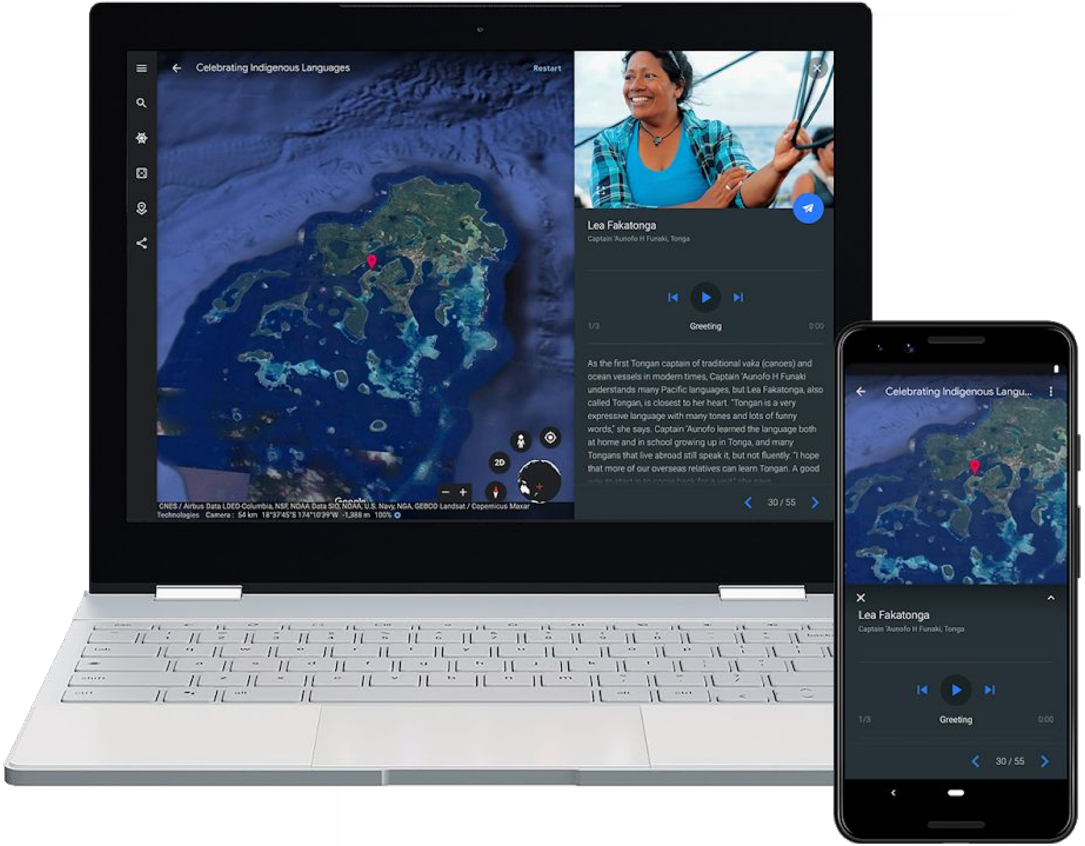
Bringing Language Technologies to More Languages

- On top of algorithms, need:
  - Clear language roadmap
  - Solid data gathering pipeline
  - Scalable data quality controls
  - Robust, easy-to-use trainer
  - Input from native speakers
  - Automatic dashboards to monitor progress & quality

- ASR
  - Unlabelled audio data
  - Transcribed audio data
  - Text in many languages

- NMT
  - Text in many languages
  - Ideally parallel, but monolingual also works

Many thanks!
Any questions?


Celebrating Indigenous Languages
Restart
2D
CNES / Airbus Data LDEO-Columbia, NSF, NOAA Data SIO, NOAA, U.S. Navy, NGA, GEBCO Landsat / Copernicus Maxar Technologies Camera: 54 km 18°37'45"S 174°10'39"W -1,388 m 100%
Celebrating Indigenous Langu...

Lea Fakatonga

Captain 'Aunofo H Funaki, Tonga

1/3 Greeting 0:00

As the first Tongan captain of traditional vaka (canoes) and ocean vessels in modern times, Captain 'Aunofo H Funaki understands many Pacific languages, but Lea Fakatonga, also called Tongan, is closest to her heart. "Tongan is a very expressive language with many tones and lots of funny words," she says. Captain 'Aunofo learned the language both at home and in school growing up in Tonga, and many Tongans that live abroad still speak it, but not fluently. "I hope that more of our overseas relatives can learn Tongan. A good way to start is to come back for a visit," she says.

30 / 55

Check this out in Google Earth at
https://goo.gle/indigenouslanguages