Building Data Infrastructure for Low-Resource Languages

Sarah Luger[1], Rafael Mosquera-Gómez[1,2], Alex Miłowski, Thom Vaughan[3], Sara Hincapie-Monsalve[1,2], Pedro Ortiz Suarez[3], Kurt Bollacker[1]

[1]MLCommons, [2]Factored AI, [3]Common Crawl Foundation

Correspondence: sarah@mlcommons.org

Abstract

The MLCommons Datasets Working Group presents a comprehensive initiative to advance the development and accessibility of artificial intelligence (AI) training and testing resources. This paper introduces three key projects aimed at addressing critical gaps in the AI data ecosystem: the Unsupervised People's Speech Dataset, containing over 821,000 hours of speech across 89+ languages; a strategic collaboration with the Common Crawl Foundation to enhance web crawling capabilities for low-resource languages; and a framework for knowledge graph extraction evaluation. By focusing on languages other than English (LOTE) and creating permissively licensed, high-quality datasets, these initiatives aim to democratize AI development and improve model performance across diverse linguistic contexts. This work represents a significant step toward more inclusive and capable AI systems that can serve global communities.

1 Introduction

The MLCommons[1] consortium supports numerous activities in the machine learning innovation space. These activities include building performance benchmarks including those in AI risk and reliability and building benchmarks for evaluating AI systems requires rigorous, standardized test datasets. MLCommons builds open, large-scale, and diverse datasets and a rich ecosystem of techniques and tools for AI data that helps the broader community deliver more accurate and safer AI systems.

Our shared research infrastructure and diverse community aid the scientific research community to derive new insights for new breakthroughs in AI. The Datasets Working Group[2] is one of 17 working groups under the MLCommons umbrella and currently has three main projects it is open-sourcing:

1. Unsupervised People's Speech Dataset,
2. Common Crawl collaboration, and
3. Datasets and Evaluation for Knowledge Graph Extraction.

While these are our current projects, the working group is always open to new topics and projects.

2 Background

Datasets fuel machine learning: a model is only as good as the data upon which it was trained. ImageNet (Deng et al., 2009), created for less than half a million dollars, is arguably the one that gave rise to modern machine learning. Unfortunately, most public datasets today are either small (relative to private commercial datasets), static, licensed for research use only, or some combination of those things. Datasets must be large to train accurate models. To remain relevant, datasets should be constantly improved as gaps in their coverage are identified. Lastly, datasets require a permissive public license to enable new businesses, products, and services globally.

The Datasets Working Group creates and hosts public datasets that are large, actively maintained, and permissively licensed – especially for commercial use. We aim to develop a center of expertise and supporting technologies that dramatically improves the quality and reduces the cost of new public datasets. We believe that a modest investment in public datasets can have an impressive return in terms of machine learning innovation and market growth. The Datasets Working Group's first project was the People's Speech dataset (Galvez et al., 2021), an open speech recognition dataset that is approximately 100 times larger than existing open alternatives.

[1] https://mlcommons.org/

[2] https://mlcommons.org/datasets/

Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025), pages 154–160
May 3, 2025 ©2025 Association for Computational Linguistics

3 Current Projects

3.1 Unsupervised People's Speech Dataset

The MLCommons Unsupervised People's Speech dataset includes over 736,000 hours of speech across 89+ languages with a diverse set of speakers. This open dataset is large enough to train self-supervised speech systems and is available with a permissive license. Building on the success of the impact of Supervised People's Speech dataset's on models like Whisper (Radford et al., 2022), the Unsupervised People's Speech dataset aims to unleash innovation in multilingual speech research and products that are available to users across the globe, with particular benefits for low-resource languages.

3.1.1 Introducing the Unsupervised People's Speech dataset

The MLCommons Dataset Working Group is pleased to announce the release of the Unsupervised People's Speech dataset. Built in collaboration with HuggingFace[3], the Unsupervised People's Speech dataset contains more than 1 million hours of audio that spans dozens of languages. Given the impact the previously released Supervised People's Speech dataset had on models such as Whisper, we expect this new version to drive innovations across different tasks, including self-supervised implementations and improvement in automatic speech recognition pipelines in numerous languages.

3.1.2 The Rationale Behind the Dataset

As the field of speech technology continues to advance, the need for large and diverse audio datasets is increasingly crucial. While several valuable speech datasets already exist—such as LibriSpeech (Panayotov et al., 2015) and Speech Wikimedia (Gómez et al., 2023) (one of MLCommons previous datasets) —they each have limitations in terms of scale and language diversity. LibriSpeech, with its approximately 1,000 hours of English-language audiobook recordings, has been instrumental in advancing English speech recognition but lacks multilingual coverage, while Speech Wikimedia provides some multilingual content but at a significantly smaller scale.

The Unsupervised People's Speech dataset aims to address this need by providing a vast collection of multilingual audio data to support research and development in various areas of speech technology. Supporting broader Natural Language Processing (NLP) research for languages other than English helps bring communication technologies to more people globally – including those speaking low-resource languages.

3.1.3 Ensuring Useful Data

To ensure the dataset is as useful as possible, the MLCommons Datasets Working Group ran different data pipelines to understand the contents across dimensions, such as language distribution and speech detection.

Speech detection The working group created a custom data loader that resampled and converted the audio to a single channel (available in the HuggingFace dataset card), we ran Silero's Voice Activity Detection (Silero-Team, 2024) pipeline, a model that uses a multi-head attention mechanism with STFT as features. The results of using this model delivered a total of 736,036+ hours of speech.

Language identification Language identification is often the first task in a series of NLP tasks, so it is crucial to get it right. The working group used Nvidia's TensorRT-LLM implementation of Whisper Large v3 to run inference on the subset of the dataset for which our speech detection pipeline detected a speech utterance. The results of this pipeline detected a total of 89 languages. While we are certain there are more, since the third most common category the model inferred was a "no speech" tag, it means the model was not able to determine the language. Here are some of the low resource languages detected:

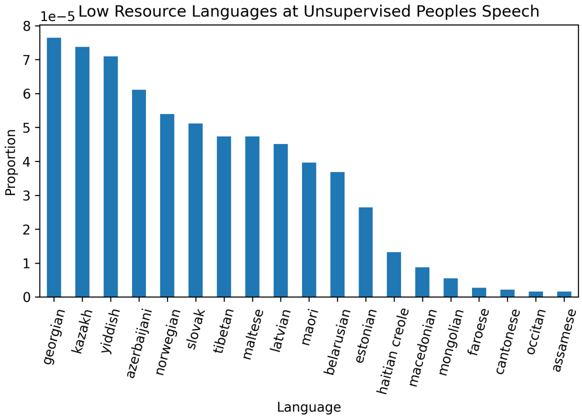


Figure 1: Distribution of Low Resource Languages

[3] https://huggingface.co/

3.1.4 Technical Hurdles

While the focus of the Unsupervised People's dataset is on its potential applications, it is worth noting some of the technical challenges that the working group overcame:

Data Upload and Storage We developed a custom script utilizing Git LFS backend to efficiently upload the 48+TB dataset to S3 and HuggingFace, where it[4] is publicly available right now. This overcame typical speed limitations.

Self-Supervised Learning Potential We have included a training pipeline to train a Wav2Vec model, which could unlock new possibilities in unsupervised speech representation learning.

Deduplication We are in the process of creating embedding representations of the entire dataset using Meta's Encodec. This will allow us to identify and remove duplicate content. This process ensures the uniqueness and quality of the dataset.

3.1.5 Future Work

The Unsupervised People's Speech dataset is built from audio data on Archive.org that is public domain or available with CC-BY or CC-BY-SA licenses. As part of our commitment to updating and maintaining the dataset, we are also releasing the software as open-source to empower the community to build on our contributions.

As the working group completes the deduplication and self-supervised efforts, we anticipate several avenues for the research community to continue to build and develop, especially in the areas of improving low-resource language speech models, enhanced speech recognition across different accents and dialects, and novel applications in speech synthesis.

3.2 Common Crawl Collaboration

The MLCommons Datasets Working Group is collaborating with the Common Crawl Foundation to enhance their web crawling capabilities. The first task the working group has been involved in focuses on improving coverage of low-resource languages. This collaboration centers on developing more effective Language Identification (LID) models that can better detect and classify content in underrepresented languages during the crawling process.

This task involves two different stages: in the first stage, the creation of a held out test set with samples in different languages labeled with the language they are in. In the second stage, a challenge will ask participants to submit language identification models which will be evaluated on the output of the first stage. We expect the best submissions to be incorporated in the stack used by Common Crawl, as long as they comply with certain efficiency and accuracy requirements.

3.2.1 Text Language Identification Task

To advance this goal, a text LID challenge was set up on the Dynabench (Kiela et al., 2021) platform. The challenge dataset includes samples from more than 400 languages, creating one of the most comprehensive language identification benchmarks to date. To ensure quality and provide guidance for annotators, we implemented a multi-model approach for creating soft labels, utilizing three language identification models: LID176 (Joulin et al., 2017; Joulin et al., 2016), LID201 (Burchell et al., 2023), and GLOT-LID (Kargaran et al., 2023).

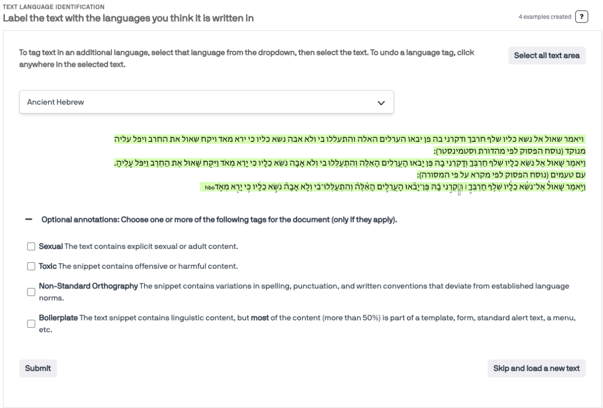
While Common Crawl already annotates their crawls using the CLD2 model[5], this model only covers 80 languages, and they only started distributing this annotation since August 2018[6]. We thus decided to re-annotate with these 3 models so that we can cover earlier crawls, more languages (since the intersection of languages covered by LID201 and LID176 is less than 70) and overcome some of the limitations of n-gram models for language identification pointed by Caswell et al. (2020) and Kreutzer et al. (2022), specially given that GLOT-LID's training data is heavily biased towards bible translations and religious content (Kargaran et al., 2023) which is not representative of the register of languages normally found in web data.

The challenge's approach requires participants to perform fine-grained language identification by highlighting specific text segments where a particular language is present. This granular annotation strategy addresses a key limitation in traditional language identification tasks, which typically operate at the document level. By February 2025, the challenge had significant engagement, with more than 600 annotations submitted by participants. To maintain momentum and expand participation, a series of hackathons are scheduled for February

[4] https://huggingface.co/datasets/MLCommons/unsupervised_peoples_speech

[5] https://github.com/CLD2Owners/cld2

[6] https://commoncrawl.org/blog/august-2018-crawl-archive-now-available



TEXT LANGUAGE IDENTIFICATION

Label the text with the languages you think it is written in

To tag text in an additional language, select that language from the dropdown, then select the text. To undo a language tag, click anywhere in the selected text.

Ancient Hebrew

ויאמר שאול אל נשא כליו שלף חרבך ודקרני בה פן יבאו הערלים האלה והתעללו בי ולא אבה נשא כליו כי ירא מאד ויקח שאול את החרב ויפל עליה
מנוקד (נוסח הפסוק לפי מהדורת וסטמינסטר):
וַיֹּאמֶר שָׁאוּל אֶל נֹשֵׂא כֵלָיו שְׁלֹף חַרְבְּךָ וְדָקְרֵנִי בָהּ פֶּן יָבֹאוּ הָעֲרֵלִים הָאֵלֶּה וְהִתְעַלְּלוּ בִי וְלֹא אָבָה נֹשֵׂא כֵלָיו כִּי יָרֵא מְאֹד וַיִּקַּח שָׁאוּל אֶת הַחֶרֶב וַיִּפֹּל עָלֶיהָ.
עם טעמים (נוסח הפסוק לפי מקרא על פי המסורה):
וַיֹּאמֶר שָׁאוּל אֶל־נֹשֵׂא כֵלָיו שְׁלֹף חַרְבְּךָ ׀ וְדָקְרֵנִי בָהּ פֶּן־יָבֹאוּ הָעֲרֵלִים הָאֵלֶּה וְהִתְעַלְּלוּ־בִי וְלֹא אָבָה נֹשֵׂא כֵלָיו כִּי יָרֵא מְאֹד hbo

— Optional annotations: Choose one or more of the following tags for the document (only if they apply).

☐ **Sexual** The text contains explicit sexual or adult content.

☐ **Toxic** The snippet contains offensive or harmful content.

☐ **Non-Standard Orthography** The snippet contains variations in spelling, punctuation, and written conventions that deviate from established language norms.

☐ **Boilerplate** The text snippet contains linguistic content, but **most** of the content (more than 50%) is part of a template, form, standard alert text, a menu, etc.

Figure 2: Interface of the Language Identification Task

and March 2025. These events will serve multiple purposes: to accelerate the annotation process, to foster community engagement, and potentially discovering novel approaches to language identification in web crawling contexts.

3.2.2 Model Benchmark: Language ID

As mentioned before, the second stage of the collaboration will involve creating a benchmark using Dynabench, through which participants will be able to submit their language identification models.

3.2.3 Future Work

In the case of the Common Crawl collaboration, there are two areas of focus:

1. Identifying and collecting URL's that contain novel data from LOTE. Information on how to submit URL's of low resources can be found here[7].

2. Finding new native language speakers of LOTE for the Dynabench language identification task. This task may be found in the following link[8].

3.3 Datasets and Evaluation for Knowledge Graph Extraction

There is an ongoing advancement in the use of LLMs for various traditional NLP tasks for knowledge extraction. Recent renewed efforts to evaluate LLM models for relation extraction tasks show State Of The Art (SOTA) performance (Wadhwa et al., 2024). The core challenge of these evaluations is to have datasets that mimic the usage in industry. While prior methods for evaluating models focused on human annotation of sentences, recent methods require annotations of larger passages with specific relations and their providence. Of the datasets considered, some are no longer available and others are limited in scope and size; producing annotations is human-intensive with possibly subjective outcomes.

Although these datasets were useful for prior research, they are probably insufficient for the testing and evaluation of LLMs. As such, we need better datasets and evaluation methods. This project intends to collect some existing candidate datasets, reformulate them with the tools ML Commons has developed (e.g., croissant annotations), and to create a framework for evaluating new AI systems for knowledge extraction tasks as a foundational tool for building knowledge graphs.

3.4 Limitations

Datasets are artifacts of the humans that make them, and thus are susceptible to bias. This bias may take the form of the annotations or judgments produced by the individual annotators or contributors, or by the demographic distribution of the crowd-sourced workers themselves.

Further, some of the contributors might not realize that they have contributed to these datasets and do not want their data published. Developing efficient and transparent methods for the removal of voice data by the owners of such data is an ongoing challenge.

3.5 Conclusions

The MLCommons Datasets Working Group is focused on supporting data-centric problem solving including AI system benchmarking, training, evaluation, and research through hosting and socializing datasets for the AI community. The three current projects of the working group are open-sourcing the MLCommons Unsupervised People's Speech dataset, ongoing Common Crawl collaboration, and automating knowledge graph development for LLM use cases.

For knowledge graphs generation, we are continuing to invest in sourcing datasets as well as in a pipeline for navigating datasets, inference, scoring, and judgement for the core relation extraction tasks. This will focus on using and enhancing the tools and approaches that MLCommons has developed for datasets and system benchmarking, and

[7] https://github.com/commoncrawl/web-languages/

[8] https://dynabench.org/tasks/text-language-identification

the project can be found here[9].

There are many ways to get involved in the MLCommons data effort. To join the Datasets working group or any other working group at MLCommons, please visit the Get Involved page[10]. The Datasets Working Group meets weekly to share ideas and implement projects. More information on the MLCommons Unsupervised People's Speech dataset can be found in the dataset official website[11].

Acknowledgments

This research is supported by MLCommons, the Common Crawl Foundation, and HuggingFace. In addition, the engagement and innovation provided by the MLCommons Datasets working group members as well as the in-kind technical support from Factored and MLPerf Automotive helped make these data resources available to the public, under the Creative Commons license[12]. We also would like to thank the ongoing financial support provided by the companies funding MLCommons and other non-profit, open source, technical initiatives. Finally, much of the data discussed in this paper is crowd sourced from individual contributors around the world. Thank you!

References

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *Preprint*, arXiv:2111.09344.

Rafael Mosquera Gómez, Julián Eusse, Juan Ciro, Daniel Galvez, Ryan Hileman, Kurt Bollacker, and David Kanter. 2023. Speech wikimedia: A 77 language multilingual speech dataset. *Preprint*, arXiv:2308.15710.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv e-prints*, arXiv:1612.03651.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

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

[9] https://github.com/mlcommons/datasets-knowledge-graphs

[10] https://mlcommons.org/get-involved/

[11] https://mlcommons.org/datasets/unsupervised-peoples-speech/

[12] https://creativecommons.org/licenses

at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Silero-Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2024. Revisiting relation extraction in the era of large language models. *Preprint*, arXiv:2305.05003.

Table 1: Most Common Languages

Language	Hours	Files
Major Languages (>1000 hours)		
English (en)	334087.22	809278
Vietnamese (vi)	103093.05	506017
Spanish (es)	70838.13	131656
Arabic (ar)	49516.85	111521
Urdu (ur)	11655.23	23871
German (de)	11122.85	36793
Portuguese (pt)	10755.48	27343
French (fr)	7698.57	20914
Hindi (hi)	4756.90	8706
Indonesian (id)	4630.81	9498
Dutch (nl)	3378.06	10500
Italian (it)	3209.77	7077
Kannada (kn)	2637.07	3455
Catalan (ca)	2429.53	4611
Polish (pl)	2390.10	3564
Russian (ru)	2246.14	10590
Nepali (ne)	1684.06	3121
Telugu (te)	1654.55	2883
Turkish (tr)	1360.22	3923
Swedish (sv)	1244.89	3373
Malay (ms)	1212.00	3221
Latin (la)	1195.60	9321
Bengali (bn)	1176.23	2117
Thai (th)	1055.32	2833
Chinese (zh)	1047.80	3576
Galician (gl)	1000.41	2101

A Unsupervised People's Speech Language Distribution

Table 2: Mid Range Resource Languages

Language	Hours	Files
Mid-Range Languages (100-1000 hours)		
Sinhala (si)	981.81	1177
Punjabi (pa)	953.47	2359
Hungarian (hu)	917.72	1104
Sanskrit (sa)	904.15	2549
Romanian (ro)	902.62	5617
Japanese (ja)	871.05	3979
Javanese (jw)	766.20	6295
Basque (eu)	609.62	2395
Tamil (ta)	561.13	2113
Marathi (mr)	513.04	763
Welsh (cy)	505.87	1237
Korean (ko)	498.85	2042
Croatian (hr)	433.07	1002
Greek (el)	398.68	1221
Danish (da)	380.49	417
Persian (fa)	369.20	1155
Swahili (sw)	359.22	1647
Hebrew (he)	358.72	3990
Somali (so)	345.39	438
Tagalog (tl)	299.08	1217
Norwegian Nynorsk (nn)	298.07	1855
Amharic (am)	263.54	302
Bulgarian (bg)	259.34	993
Sindhi (sd)	246.71	708
Serbian (sr)	236.85	396
Ukrainian (uk)	232.10	425
Malayalam (ml)	228.99	623
Pashto (ps)	218.84	373
Afrikaans (af)	158.25	749
Finnish (fi)	153.72	664
Burmese (my)	153.61	346
Khmer (km)	142.35	503
Bosnian (bs)	141.82	262
Gujarati (gu)	135.89	286
Icelandic (is)	116.38	206

Table 3: Lower Resource Languages

Language	Hours	Files
Lower-Resource Languages (<100 hours)		
Yoruba (yo)	97.49	332
Czech (cs)	92.82	425
Shona (sn)	80.83	513
Slovenian (sl)	77.80	354
Albanian (sq)	67.01	171
Tibetan (bo)	57.26	86
Kazakh (kk)	52.26	134
Azerbaijani (az)	49.20	111
Hawaiian (haw)	42.40	354
Norwegian (no)	42.11	98
Maltese (mt)	41.34	86
Armenian (hy)	38.30	446
Yiddish (yi)	33.14	129
Breton (br)	31.29	268
Maori (mi)	29.10	72
Latvian (lv)	28.98	82
Lithuanian (lt)	26.95	145
Slovak (sk)	21.43	93
Georgian (ka)	21.32	139
Macedonian (mk)	10.82	16
Estonian (et)	7.82	48
Mongolian (mn)	7.71	10
Cantonese (yue)	6.21	4
Belarusian (be)	6.04	67
Haitian (ht)	5.41	24
Assamese (as)	2.32	3
Faroese (fo)	2.30	5
Occitan (oc)	0.13	3