Language Technology for All

Industry Initiatives to Serve Low Resource Languages

Blaise Hylak

About Myself

Blaise Hylak is a localization industry professional with six years of experience. A graduate of Villanova University, he rose through the ranks from intern to Program Manager. He holds a master's degree in Technical Communication and Localization from the University of Strasbourg. He speaks at language industry events nationwide, manages teams for local/national clients as a Program Manager at Come Alive Communications, Inc., and participates as a consultant/researcher to local, national, and international organizations requiring guidance on their current tech stacks, DEI, and how to enhance processes. He has been a member of ATA's Language Technology Division Leadership Council for the past three years.

Paradigm Shift in the LOC Industry

"Equal access representation that is [...] pushed by governmental regulations [...] [make clients] more interested in ensuring [...] [that] access to all kinds of [...] baseline services or information is guaranteed" (Beccaletto, 2023).

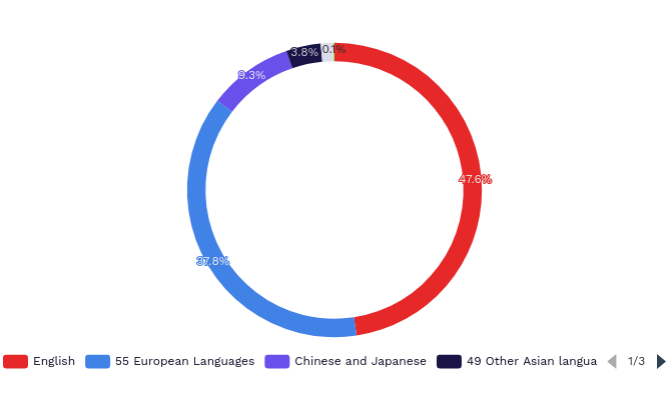
Simona Beccaletto, Head of TAUS' Human Language Project Operations

Legislation Mandating Language Access

- **United States of America (USA)**
  - Americans with Disabilities Act (ADA)
  - Affordable Care Act (ACA)
  - Civil Rights Act of 1964, Title VI
  - Executive Order 13166 (2000)
- **European Union (EU)**
  - European Charter for Regional or Minority Languages
  - European Accessibility Act (EAA)

Tech Bias Toward Low Resource Languages

- In a December 2023 exclusive interview for my thesis, Don DePalma of CSA Research provided an excellent graph produced by CSA Research that illustrates all training data by language group in the Common Crawl as of May 2023
- 85.4% of the data is for European Languages
- English alone astonishingly accounts for nearly half of all training data



Research Overview

Research Question

Abstract

Methodology

Principle Findings

Meta	Google	Microsoft	TAUS	Translators without Borders (TWB)
Meta is developing high quality MT and S2ST tools.	Google is developing high quality datasets and developing a quality S2ST tool.	Microsoft is developing high quality data.	TAUS is developing high quality data.	TWB is developing high quality data.
No Language Left Behind (NLLB), Seamless	**MADLAD-400, Translatotron 3**	**Project ELLORA**	**Human Language Project (HLP)**	**Language Data Initiative**

Meta

- **No Language Left Behind (NLLB)**
  - Uses Human-Translated Datasets and Tools to Create Large Bitext Datasets
  - Achieved a 44% improvement in BLEU scores, advancing the goal of a universal translation system
  - Open-sourced tools and models for wider community use

Google

- **MADLAD-400**
  - Spans 419 languages, aiming to provide comprehensive data for MT and NLP research.
  - Includes 3 trillion clean tokens and 100 billion words, with a focus on LRLs
  - Models trained on MADLAD-400 have shown competitive performance with larger models

Microsoft

- **Project ELLORA**
  - **Prioritizes Indian Languages:** This initiative focuses on Indian languages with limited digital presence, such as Gondi, Mundari, and Idu Mishmi
  - **Partnerships in Data Collection:** Microsoft collaborates with local communities to gather and preserve language data, ensuring cultural sensitivity and accuracy
  - **Tailored Digital Resources:** They develop digital dictionaries, translation services, and educational tools specifically for these languages, fostering digital inclusion and language preservation

TAUS

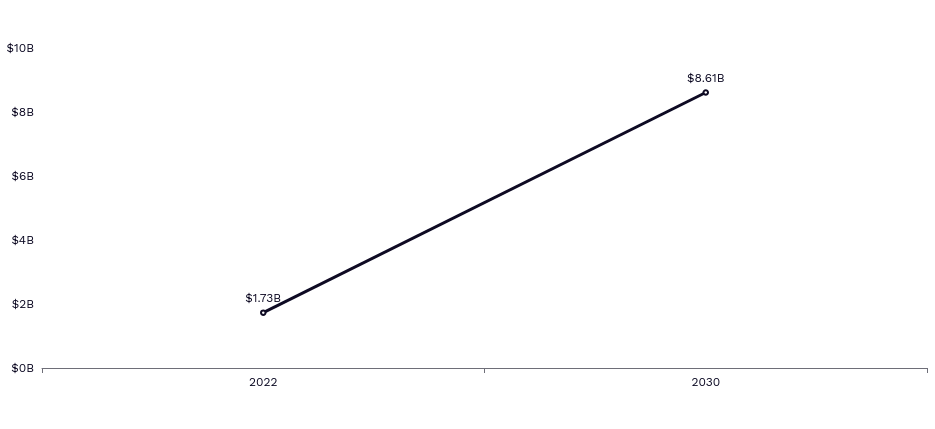
- **Human Language Project (HLP)**
  - Focuses on creating data for machine translation (MT) and speech-to-speech translation (S2ST)
  - Involves crowd-sourced data collection from diverse global communities
  - Has expanded to cover over 30 languages across 20 countries

Translators without Borders (CLEAR Global)

- **Language Data Initiative**
  - 58 datasets covering almost 60 countries
    - Curated, cleaned, and reformatted the data to be more accessible for humanitarian and developmental purposes
  - Learn which languages are spoken where and more about the state of language data globally with the Global Language Data Review
  - Pre-formatted and translated questions for language data collection

Language Data for AI (LD4AI)

- Since there are no reliable estimates available for the LD4AI market, LD4AI is treated as an emerging sub-sector of the overall AI training data market.
- According to a market report by Grand View Research (2023), the AI training data market, including LD4AI, is "expected to reach $8.61 billion by 2030 and expand at a CAGR of 22.1% from 2023 to 2030."



DATAFORCE BY TRANSPERFECT

TransPerfect, the world's largest LSP, redirect their efforts towards AI data solutions.

Data Collection	Data Annotation	Data Relevance	Localize Chatbots
"Gather high-quality data for your unique model training/evaluation needs configuration options" (TransPerfect, 2024).	"DataForce accelerates your data labeling processes with our range of human annotation services at scale" (TransPerfect, 2024).	"DataForce supports and improves your relevance models, accuracy, and recall to ensure the content you showcase to your customers is relevant, culturally acceptable, and geographically precise" (TransPerfect, 2024).	"Create chatbots that sound human and are culturally appropriate" (TransPerfect, 2024).
Data Moderation	**Transcription**	**User Studies**	**Gen AI Training**
"A multicultural and multilingual solution for your moderation needs" (TransPerfect, 2024).	"Scale speech and audio recognition capabilities with DataForce" (TransPerfect, 2024).	"DataForce utilizes its global footprint in over 46 countries to build your personalized experience and collect the data you need through piloted, situational, and custom user studies" (TransPerfect, 2024).	"Whether you are developing new foundational models, such as LLMs, or customizing an existing model for a new use case, DataForce tailors solutions that address the unique data challenges your organization faces" (TransPerfect, 2024).

MT Methods for LRLs

Transfer Learning

Both few-shot and zero-shot MT can be considered forms of transfer learning. In few-shot, the model benefits from additional fine-tuning, while in zero-shot, the transfer occurs without direct data.

Multilingual Training

Multilingual training serves as the foundation for zero-shot MT. It enables the model to learn shared representations across languages, which is crucial for zero-shot capabilities.

Back-Translation

Back-translation is often used to generate synthetic data for both few-shot and zero-shot scenarios. This method can improve performance in both contexts, particularly when combined with few-shot fine-tuning.

Data Augmentation

Data augmentation is a general strategy that can be applied to enhance few-shot MT by increasing the diversity of training examples. It can also benefit zero-shot MT indirectly by improving the robustness of the model.

Unsupervised MT

Unsupervised MT aims to perform translation without parallel corpora, similar to zero-shot MT. However, it typically requires large monolingual datasets, which zero-shot MT does not.

Semi-Supervised Learning

Semi-supervised learning often complements few-shot learning by using small amounts of labeled data alongside large amounts of unlabeled data. Zero-shot MT can benefit from semi-supervised techniques as well but often in a more indirect manner.

Active Learning

Active learning is primarily relevant to few-shot scenarios, where the goal is to optimize the use of limited data by selecting the most informative examples for annotation. Zero-shot MT doesn't directly benefit from active learning because it relies on transfer learning rather than data selection.

Meta-Learning

Meta-learning is similar to zero-shot learning in that it prepares the model to adapt quickly to new tasks (languages) with minimal data. Few-shot MT can be seen as a practical application of meta-learning principles.

Language-Adaptive Fine-Tuning

Language-adaptive fine-tuning is more applicable to few-shot scenarios, where a model is further adjusted using limited data from the target language. Zero-shot MT benefits from this only if the fine-tuning indirectly improves the model's cross-lingual capabilities.

Crowdsourcing and Community Involvement

Few-shot MT can benefit significantly from crowdsourced data, where small but highly relevant datasets are collected. Zero-shot MT typically doesn't rely on such data but can benefit from the improved model architectures and representations developed through few-shot methods.

LLMs for LRLs?

1. A study found that LLMs, including ChatGPT, struggle significantly with LRLs compared to high-resource languages
2. New approaches like Linguistically-Diverse Prompting (LDP) have been developed to help LLMs better handle LRLs by leveraging their strength in high-resource languages, particularly English
3. Despite advancements, LLMs continue to underperform in low-resource settings, often failing to surpass traditional machine learning models in tasks like machine translation and named-entity recognition
4. While there have been technical improvements, the performance gap between LRLs and high-resource languages for LLMs remains significant, highlighting the need for continued research and development

LRLs Can Jailbreak GPT-4...

- Researchers in a study jailbroke GPT-4 by translating unsafe English input into a low resource language, and consequently inputting that output into GPT-4. In essence, requesting a "back translation" (English to a low resource language and then back to English)
- "GPT-4 engages with the unsafe translated inputs and provides actionable items that can get the users towards their harmful goals 79% of the time"

Conclusion

Generating/gathering high quality data is essential for language technology initiatives. In regards to low resource languages, this is a particular challenge that must be addressed and budgeted.

However, the matter of supporting low resource languages represents an interesting industry crossroads. Since governments worldwide are increasingly mandating language access with legislation, that also implies FUNDING. The financial incentives alone are a compelling reason to continue supporting initiatives for low resource languages besides mere compliance of the law.

References

- Shiyas, A. (2023, November 13). *Microsoft Research project helps languages survive — and thrive.* Microsoft. https://news.microsoft.com/source/asia/features/microsoft-research-project-helps-languages-survive-and-thrive/
- Stasimioti, M. (2023, March 10). *Google sheds light on 1000-languages universal speech model.* Slator. https://slator.com/google-sheds-light-on-1000-languages-universal-speech-model/
- Albarino, S. (2023, May 23). *Meta challenges Whisper with massively multilingual speech launch.* Slator. https://slator.com/meta-challenges-whisper-with-massively-multilingual-speech-launch/
- TransPerfect. (2024). *DataForce by TransPerfect.* https://www.dataforce.ai
- Grand View Research. (2023). *AI Training Dataset Market Size, Share and Trends Analysis Report, By Type (Text, Image/Video, Audio), By Vertical (IT, Automotive, Government, Healthcare, BFSI), By Regions, And Segment Forecasts, 2023 - 2030.*
- Goyal, N., Le, Q., & Vinyals, O. (2021, June 23). Few-shot learning for machine translation. Google AI Blog. https://ai.googleblog.com/2021/06/few-shot-learning-for-machine-translation.html
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5, 339-351. https://www.aclweb.org/anthology/Q17-1024/
- Zoph, B., & Le, Q. V. (2016). Neural architecture search with reinforcement learning. arXiv. https://arxiv.org/abs/1611.01578
- Aharoni, R., Johnson, M., & Firat, O. (2019). Massively multilingual neural machine translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, pp. 3874-3884). https://www.aclweb.org/anthology/N19-1388/
- Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 86-96). https://www.aclweb.org/anthology/P16-1009/
- Fadaee, M., Bisazza, A., & Monz, C. (2017). Data augmentation for low-resource neural machine translation. arXiv. https://arxiv.org/abs/1705.00440
- Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. A. (2018). Unsupervised machine translation using monolingual corpora only. In International Conference on Learning Representations. https://arxiv.org/abs/1711.00043
- Settles, B. (2009). Active learning literature survey. University of Wisconsin, Madison. https://minds.wisconsin.edu/handle/1793/60660
- Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1126-1135). https://proceedings.mlr.press/v70/finn17a.html
- Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., & Gurevych, I. (2020). AdapterFusion: Non-destructive task composition for transfer learning. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1103-1116). https://www.aclweb.org/anthology/2020.acl-main.98/
- Zaidan, O. F., Callison-Burch, C., & Poesio, M. (2011). Crowdsourcing annotation for machine translation: A tale of two approaches. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 104-113). https://www.aclweb.org/anthology/D11-1010/
- CSA Research. "Percentage of Training Data by Language Group." Graph. In Don DePalma, "Wrestling with the Ethics of GenAI," tcworld magazine, December 2023. https://www.tcworld.info/e-magazine/intelligent-information/wrestling-with-the-ethics-of-genai-1285/.
- Yong, Zheng-Xin, Cristina Menghini, and Stephen H. Bach. "Low-ResourceLanguages Jailbreak GPT-4." Submitted October 3, 2023. PDF. Accessed December 14, 2023. https://arxiv.org/pdf/2310.02446.pdf.
- CLEAR Global. "Language Maps and Data." 2023. https://clearglobal.org/language-maps-and-data/.
- Beccaletto, Simona. Virtual interview by Blaise Hylak. December 20, 2023.
- DePalma, Donald. Virtual interview by Blaise Hylak. December 19, 2023.
- "TAUS Human Language Project." Accessed January 14, 2024. https://hlp.taus.net.
- Kudugunta, Sneha, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A.Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. "MADLAD-400: A Multilingual And Document-Level Large Audited Dataset." Preprint. Accessed December 2023. https://arxiv.org/pdf/2309.04662.pdf.
- 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, et al. "No Language Left Behind: Scaling Human-Centered Machine Translation." Accessed December 14, 2023. https://arxiv.org/pdf/2207.04672.pdf.
- Alam, Firoj, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. "Linguistically-Diverse Prompting (LDP) and Advancements in Low-Resource Languages." In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts, June 2024. https://aclanthology.org/2024.eacl-tutorials.5.
- Robinson, Nathaniel R., Perez Ogayo, David R. Mortensen, and Graham Neubig. "ChatGPT MT: Competitive for High- (but not Low-) Resource Languages." September 14, 2023. PDF. Accessed December 14, 2023. https://arxiv.org/pdf/2309.07423.pdf.
- Court, Sara, and Micha Elsner. "Shortcomings of LLMs for Low-Resource Translation: Retrieval and Understanding are Both the Problem." arXiv, June 2024. https://arxiv.org/abs/2406.15625.
- Cahyawijaya, Samuel, Holy Lovenia, and Pascale Fung. "LLMs Are Few-Shot In-Context Low-Resource Language Learners." arXiv, May 2024. https://arxiv.org/abs/2403.16512.

Wanna continue the conversation? Connect with me on LinkedIn!

Blaise Hylak
Program Manager at Come Alive Communications, Inc.