The Linguisitc Diversity of the Aya Dataset

Hannes Essfors - hannes.essfors@uivie.ac.at

Universität Wien, Vienna, Austria

Abstract. This report investigates the Aya dataset from a perspective of linguistic diversity, exploring how it aligns with the goal of reducing linguistic inequality. An introduction to the original paper on the Aya dataset by Singh et al. is given after which the phylogenetic, scriptic, endangerment and featural diversity is explored, considering both the number of languages and datapoints in the dataset. The report concludes that the datasets is substantially lacking in endangerment diversity, while being somewhat more diverse phylogenetically and scriptic, but still being very dominated by Indo-european languages and the latin-script. It is concluded that the Aya dataset is a valuable resource for reducing the dominance of English in the digital space and especially NLP, but for reducing linguistic inequality, further measures should be taken.

Keywords: Linguistic Diversity · Multilingual NLP · Aya-dataset.

1 Introduction

The rapid development of NLP (natural language processing) and LLM (large language models) in recent years has certainly not gone anyone unnoticed, and the technology has found use in the most diverse of settings, from helping medical professionals with diagnostics to sport teachers planning their physical education class. While it seems that the development of language technology will impact much of the way that we work, communicate and structure our daily lives, it is not clear what the impact on language itself will be. Generally, the digital space is dominated by a few languages, mainly English. As of January 2024, 52.1 % of all webpages were in English language [3] and similarly, nearly 75 % of all scientific articles published are in English, while in the natural and social sciences, the same number goes up to 95 % [5]. It should thus not be a surprise that most of the language models developed have been designed with English as the primary language of usage in, resulting in bias favoring English language. [7] At the same time, however, only approximately 17 % of the population of the world are proficient in English. [5]

While it is clear that the dominance of English influences the content generated by large language models and discriminate non-English speakers, it is unclear how the digital dominance of English impacts linguistic diversity, the promotion of which is outspokenly recommended by UNESCO. [4] While it could be that the English dominance primarily drives non L1 English speakers to acquire English as an L2, it could also very well be a factor driving people to abandon their native language to gain access to enhanced opportunities presented by new technology. Considering that out of the currently 7000 languages in the world [2], at least 1500 are projected to be lost by 2100 [6], the second scenario might be the more likely, and perhaps a risk that should not be taken as not to exacerbate language loss.

With this as a linguistic background, it is interesting to consider one of the more recent projects aiming to address the situation, namely the Aya-initiative [1], and more specifically, the *Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning.* [11]. In this report, I will first give an overview of the dataset and the paper written by Singh et al., with a special focus on the aims of the project and the data collected. Then, I will look further into the diversity of the dataset and discuss its implication for multilingual NLP and linguistic diversity.

2 The Aya Dataset

In this section, the original paper by Singh et al [11] will be described in further detail, outlining the motivation behind the project and describing its data.

2.1 Motivation

In line with the argumentation in the introduction, the authors of Aya recognize that NLP and large language model has been an English endeavor, putting the

English language in focus. They argue that the recent advances in NLP has been driven by breakthroughs in instruction fine-tuning of large language models, with almost all datasets being in English language. This creates a bias, impairing performance on languages not represented in the data, which according to the authors risks creating a snowball effect, where the poor get poorer and the rich get richer.

The line of reasoning goes as follows:

1. English based models enjoy the best data resources and thus perform the best.
2. English models therefore get more support and become more accessible, cheaper and capable.
3. Speakers of low-resource languages are thus constrained to lower quality technology without the resources to develop their own.

This prompts the need to reduce linguistic inequality by developing high quality datasets in other languages than English, especially in low- to mid-resource languages [1]. While recognizing that work in line with this effort has been carried out, it has been focused on multilingual pretraining and not instruction fine-tuning, thereby creating a gap in the research which Aya aims to fill.

The authors acknowledge that attempts have been made to translate English instruction fine-tuning datasets, but argue that such datasets suffer from translation bias or an inability to take cultural context into account properly. This propels the work of Aya, aiming to gather difficult-to-obtain data from fluent language speakers. The result is the creation of three key resources: Aya Dataset, Aya Collection and Aya Evaluation suit. While the Aya evaluation provide a platform for evaluating multilingual NLP and the Aya collection a large machine translated dataset for many languages, the crowing achievement is the Aya Dataset, creating the largest human annotated instruction fine-tuning dataset to date. In order to constrain the scope of this research report, the focus will be kept on the Aya Dataset

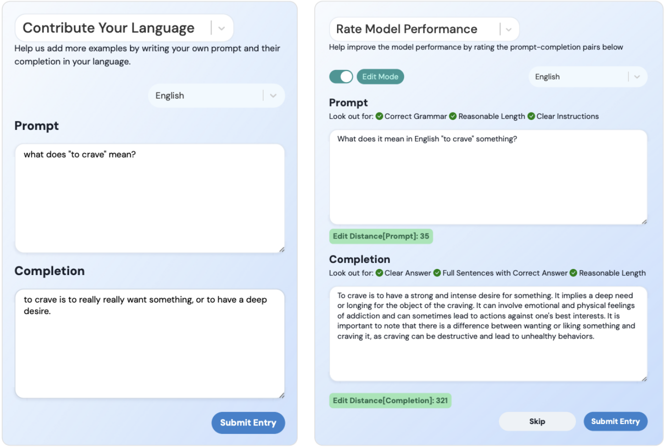
2.2 Data

The original goal of the Aya dataset was to include all languages included in the mT5 model, which covers 101 languages [12]. In the end, not all of these languages received enough contributions, resulting in only 65 languages being included. The contribution was not restricted to the languages of mT5, thus some other languages which received enough data were included. In the end, 65 languages were included.

[1] High-resource (HR), Mid-resource (MR) and Low-resource(LR) is a taxonomy introduced to classify languages according to data availability for NLP tasks. The original clustering was done by Joshi et al. [9], classifying languages into categories from 0 to 5 depending on availability of labeled and unlabeled datasets. These categories are then further categorized into low, mid and high according to e.g. the schema low = category(0,1,2), mid = category(3), high = category(4,5). [11]

The data was collected using an annotation platform, allowing the users to register, sharing their age and self-identified gender and select what language they wanted to provide annotation for with the requirement of being fluent in that language. It was also possible to provide data for a "dialect" however it was decided not to include any dialects in the final dataset. All in all, data was collected from 119 countries on 134 languages.

On the annotation platform, users could contribute through three distinct tasks. The first was to contribute an original annotation, by the means of which a prompt together with its completion was supplied, an example given by the authors can be seen in figure 1. The purpose was to create organic content, allowing the humans to use their creativity and create culturally appropriate content, not biased towards the culture of English speaking countries.


Contribute Your Language
English
Prompt
Completion
Submit Entry
Rate Model Performance
Edit Mode
English
Prompt
Look out for: ✓ Correct Grammar ✓ Reasonable Length ✓ Clear Instructions
Edit Distance[Prompt]: 35
Completion
Look out for: ✓ Clear Answer ✓ Full Sentences with Correct Answer ✓ Reasonable Length
Edit Distance[Completion]: 321
Skip
Submit Entry

Help us add more examples by writing your own prompt and their completion in your language.

what does "to crave" mean?

to crave is to really really want something, or to have a deep desire.

Help improve the model performance by rating the prompt-completion pairs below

What does it mean in English "to crave" something?

To crave is to have a strong and intense desire for something. It implies a deep need or longing for the object of the craving. It can involve emotional and physical feelings of addiction and can sometimes lead to actions against one's best interests. It is important to note that there is a difference between wanting or liking something and craving it, as craving can be destructive and lead to unhealthy behaviors.

(a) Example of an original annotation contribution.

(b) Example of a re-annotation contribution.

Fig. 1. An example of a prompt completion pair going through the process of original annotation and reannotation. Supplied by Sing et al. [11]

Giving annotators the freedom to generate their own content induced some problems pertaining to the lack of knowledge regarding machine learning of the annotators. The project was directed towards laymen willing to contribute to the percevation of their language, thus many might not be familiar with how a prompt completion pair should look, potentially reducing the quality of the dataset. This was addressed by introducing a second task: reannotation. By undertaking a reannotation, an example of which can be found in 1, a user was supplied either, (1) an original annotation, or (2), a machine translated prompt completion pair form established IFT-datasets. The task was then to correct any errors found, and expand the prompt or completion if necessary.

This had the function of serving as a model for annotators, giving examples of the structure expected for prompt completion pairs, while also setting up a feedback loop to improve the quality of the dataset. If an original annotation or translated prompt-completion pair was reannotated and sufficiently improved upon (measured by edit distance between the two instances), then it was included in the dataset. A further measure to ensure high quality of the data apart from the reannotation task was the third possible task of giving annotation feedback. In this process, annotators were given the possibility of rating the quality of other annotators' contribution, allowing for robust evaluation of the data quality.

To be included in the dataset, the content had to be either an original annotation or a reannotation with an edit distance larger than 5. Furthermore, only languages with more than 50 annotations were included, which is why only 65 languages were included even though data existed for 119. Of these 65 languages, 22 are classified as high-resource, 12 as mid-resource and 31 as low-resource. In total, the dataset consists of 204114 prompt completion pairs.

Singh et al. release some information regarding the composition of the annotators. As can be seen in figure 2, the user base is biased towards males (68.1%) in the age range of 18-35 (78.5%)

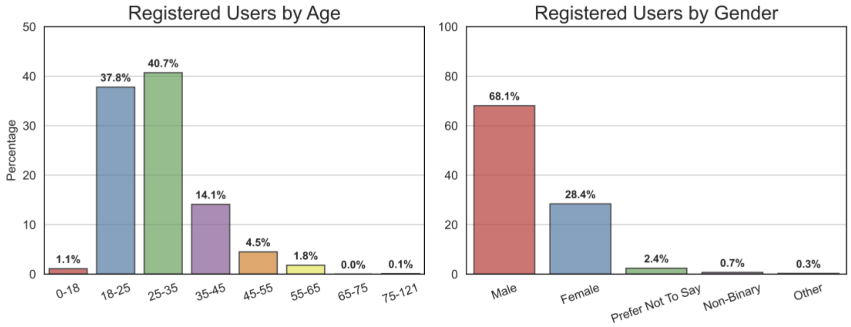


Fig. 2. Distribution of users on the Aya platform by age and gender according to Singh et al. [11]

Considering the geographical distribution, most contributors stem from Asia or Africa (58.8% and 27.4% respectively), as seen in figure 3. The largest source of contributors was India with 347 of 2997 contributors.

3 Assesing the Diversity of the Aya Dataset

In this part, I will carry out a thorough analysis of the linguistic composition of the Aya dataset. We know from the previous section as disclosed in the original paper that the distribution of resource levels is fairly even, with a plurality being low resource. What is not clear however is how diverse it is in regard to e.g. genealogy, script and typology. To assess this, the language data provided in table

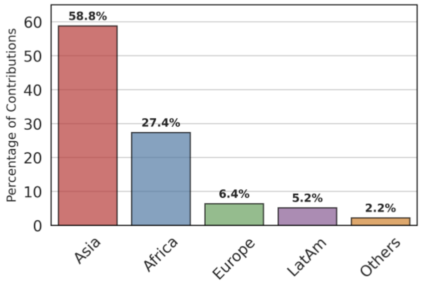


Fig. 3. Geographical distribution of contributions according to Singh et al. [11]

5 in the original paper [11] containing language family, script and resource-level was converted to a .csv file and read into R for analysis. The number of instances for each language in the dataset was aggregated from the dataset, available at its huggingface page (`https://huggingface.co/datasets/CohereForAI/aya_dataset`), Endangerment levels (AES) for the languages were accessed through Glottolog [8] and distances from URIEL+ [10]. For replication purposes, note that the aya dataset in the article uses ISO 639-2 codes while URIEL+ uses ISO 639-3 or glottocodes as language identifiers. Furthermore, the dataset at huggingfaces disaggregates languages onto dialects, which isn't aligned with the reasoning in the article. This analysis aggregates the dialects to stay as closely aligned with Singh et al. as possible. All data and code can be found at `https://github.com/Eszettfors/aya_dataset_linguistic_diversity/tree/main`

3.1 Data Sources

Glottolog [8] is a database collecting bibliographies on languages of the world, containing references for e.g. grammars and dictionaries contributed by expert linguistics. It also features an extensive collection of proven genealogic relationships between as many languages as possible, aiming to classify all "languoids" of the world. In glottolog terminology, a languoid is any family, language or dialect which can be decribed by linguistis, therefore giving it a unique 8 number code consisting of 4 alphanumeric characters and 4 decimal digits. In total, glottolog contains data on 7667 spoken L1 languages and 430 language families and isolates. Furthermore, glottolog aggregates data on language endangerment from multiple sources under the name of *glottoscope*, resulting in the Agglomerated Endangerment Status (AES), classifying languages on a scale from 1 to 6 with 1 denoting no threat, and 6 meaning the language has already gone extinct.

URIEL+ [10] is a knowledge base for language representation with the purpose of enabling large scale language comparisons for multilingual NLP by representing languages as typological vectors. URIEL+ aggregates multiple databases on language typology and allows distances between these vectors to be calculated based on their dissimilarities (cosine or angular distance). URIEL+ allows for the calculation of multiple linguistic distances: phyologenetic, syntactic, phonological, inventory, morphology, and a combined featural distance. This is done by modelling each linguistic feature as a dimension, and decoding its presence in a given language binary. The largest challenge for any attempt at large scale language comparisons is the matter of missing data due to lack of sufficient descriptions. As a result, URIEL+, while featuring vector representations of 7970 languages, miss 87% of possible values. To address this, URIEL+ makes use of imputation algorithms to impute missing values, allowing for distance calculations for 4005 languages.

For the purpose of the analysis, the default settings of URIEL+ is used together with softimpute as imputation algorithm, calculating the featural distances between the languages in the Aya dataset, generating a 65x65 matrices with distance values. These distances are due to the way they are calculated restricted to values between 0 and 1, with 0, meaning the languages have identical feature vectors, and 1, meaning they share no features.

3.2 Distribution of Prompt-Completion Pairs

Considering that the Aya dataset aims to reduce the linguistic inequality, it is positive that a relative majority of the languages included are classified as low-resource, as mentioned in the previous chapter. However, it is still unclear how the distribution of prompt completion pairs in the dataset are distributed over these languages. By considering the number of instances for each language in the dataset, this can be assessed. Viewing the boxplot in figure 4, it seems that the distribution of instances over low-resource languages is fairly similar. The median number of prompt-completion pairs among low-resource languages is 1368 (mean = 3358), while for mid-resource languages it is substantially smaller with 734 instances (mean = 2367), while among the high resource languages it is somewhat larger with 1530 as a median (mean = 3358). For all levels, clear outliers with more than 10000 instances are found, 3 of which are low-resource (Plateau Malagasy, Sinhala and Yoruba), 2 are mid-resource (Tamil and standard Malay) while only Arabic is high-resource.

While the strong representation of low-resource languages is in line with the aims of the Aya-dataset to reduce linguistic inequality, the fact that high-resource languages on average are more well represented counteracts this objective. Especially problematic is the weak representation of mid-resource languages, which not only has the smallest amount of languages, but also the fewest instances in the dataset.

3.3 Phylogenetic diveristy

To assess phylogenetic diversity, the number of language family and the evenness of the distribution of languages across them has to be considered. Theoretically, the maximum richness would be 65, due to 65 languages being present in the dataset, with each belonging to a different language family to maximize evenness.

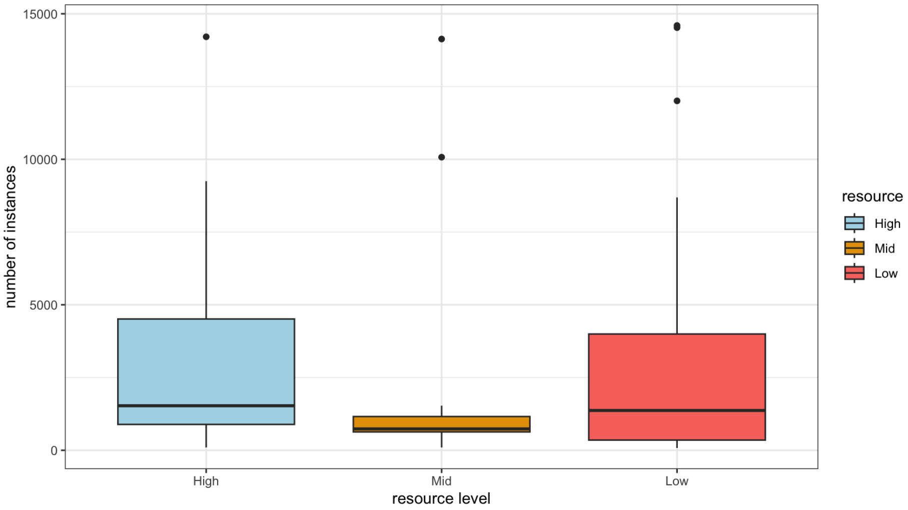


Fig. 4. Boxplots showcasing the distribution of prompt completion pairs in the aya dataset across level of resources according to Joshi et al. [9]

The richness and evenness for the Aya dataset can be observed in figure 5. As can be seen, the dataset is clearly dominated by Indo-European languages, making up almost half the dataset with 30 languages out of 65. In total, there are 13 distinct language families. To assess the evenness of the distribution of language families, each language family is turned into a probability according to

$$p_i = n_i/N \tag{1}$$

where the p_i is probability of randomly picking a language belonging to language family i, n_i is the number of languages belonging to that family and N is the total number of languages. Then, the Shannon diversity index is calculated as

$$H' = -\sum_{i=1}^{k} p_i \ln p_i \tag{2}$$

where k is the total number of language families. It follows that the phylogenetic Shannon entropy of the Aya dataset amounts to 1.856. To arrive at the evenness of the dataset, the Shannon entropy can be normalized to a number between 0 and 1 by dividing with the theoretical maximal entropy of the dataset such that

$$E = \frac{H}{H_{max}} = \frac{-\sum_{i=1}^{k} p_i \ln p_i}{\ln(k)} \tag{3}$$

which gives an evenness of 0.724, signaling a fairly even distribution of language families. Furthermore, for better interpretation, the Shannon entropy can be turned into an effective number of language families by exponentiating it as

$$D = e^{H'} \tag{4}$$

where D is the true diversity, implying what number of equally abundant species would be the equivalent to the given Shannon entropy for this dataset. For the Aya dataset, the true phylogenetic diversity is 6.4, meaning that its entropy equals that of approximately six equally distributed language families.

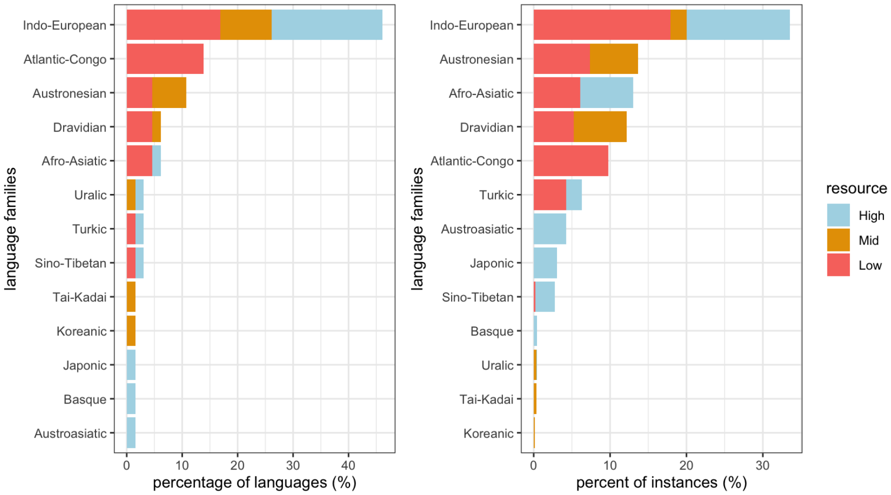


Fig. 5. Barplot showcasing the 13 language families present in the Aya dataset by percentage of the 65 languages and 204 000 instances in the dataset.

If one instead of considering the number of languages in each family considers the distribution of instances i.e. number of prompt-completion pairs as seen in figure 5, then the dominance of Indo-European falls somewhat although clearly still being the most well represented language family in the dataset with more than 30% of all instances. The distribution gets more even with the Austronesian, Afro-Asiatic and Dravidian language families making up larger parts of the dataset. Formally, the richness naturally remains the same with 13 language families, but the Shannon entropy increases to 1.981 due to the increased evenness of 0.773. As such, the true diversity also increases to 7.3, almost equivalent to adding an entire language group.

Considering the distribution of NLP-resources, most of the high-resource languages are Indo-European, while all Atlantic-Congolese language are low-resource, highlighting the geographic divide in NLP-technology of the global south and north. Furthermore, the increased diversity when considering the distribution of instances seems to a large part be driven by mid- and low-resource languages, which see a large increase in the Dravidian, Austronesian and Afro-Asiatic language families, while the high-resource languages drive the increased diversity in the Afro-Asiatic, Austroasiatic and Sino-Tibetan language families.

3.4 Scriptic diversity

Similar to phylogenetic diversity, scriptic diversity can be analyzed by looking at the number of scripts present and evaluate the diversity by richness and evenness of the distribution of languages and instances across the scripts in the datasets. All in all, 19 different scripts are present in the dataset, giving a richness of 19. Considering the evenness, however, it can be seen in figure 6 that the latin script is clearly dominating both the number of languages in the dataset and the number of instances, with almost 60% of all languages and 50% of all instances. Based on the number of languages, the Shannon entropy amounts to 1.788, due to an evenness of 0.608. In total, this gives a true diversity of 5.98 equally distributed scripts.

Considering the number of instances in the dataset, the distribution again gets more even compared to looking at the number of languages, indicating that the underrepresented language speakers contribute more. The Shannon entropy increases to 1.912 and the evenness to 0.650. It follows that the true diversity equals 6.78, again, almost an increase of an entire script, looking at instances compared to languages. Still, the Latin script is clearly dominating the dataset.

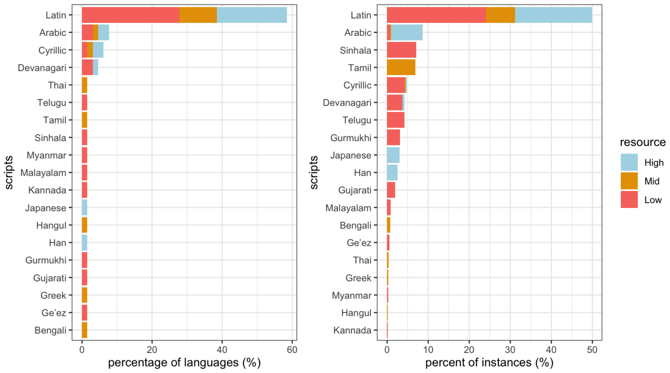


Fig. 6. Barplot showcasing the 19 scripts present in the Aya dataset by percentage of the 65 languages and 204 000 instances in the dataset.

3.5 Endangerment

Since an outspoken aim of the aya-project is to reduce linguistic inequality and thus enhance linguistic diversity, it would be desirable that not only technologically underrepresented languages are included, but that these also are threatened, as this would imply that they are the languages most in need of additional support for their survival.

Looking into the data, it is however apparent that this is not the case, as can be seen in figure 7. Only languages from the three lowest levels of threat are represented, of which 95 % belong to threat level 1, which means they are not endangered. Considering all languages of the world, only 35.13 % of all languages belong to this group [8], meaning they are very overrepresented in the dataset. Only two languages have threat level 2, threatened, namely Basque and Ukrainian, while only a single language is of threat level 3, shifting, namely Irish. Considering the instances, these threatened languages are not very well represented by number of data points in the dataset, which is made clear by the reduction in true diversity when considering instances, which for the threat level amounts to 1.242 while slightly lower for the languages (1.082). Thus, from a diversity point of view, there is more or less only one threat level represented in the dataset: not endangered.

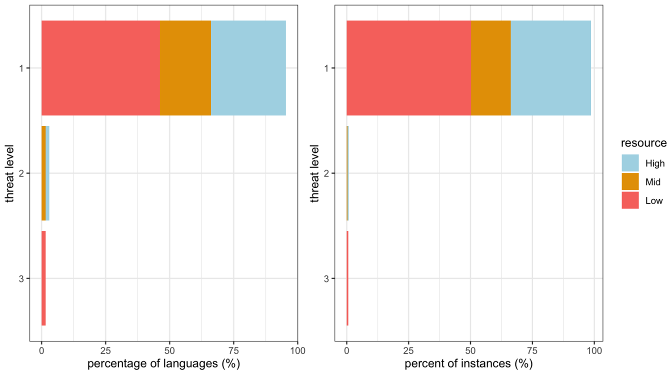


Fig. 7. Barplot showcasing the distribution of threat levels according to AES in the Aya datset

3.6 Feautral diversity

For a dataset to best capture the richness of languages of the world, the languages included should ideally be as different from each other as possible. A dataset containing three very different languages such as Turkish, Chinese and Russian is of course more diverse than a dataset with Serbian, Bosnian and Montenegrin as an extreme example. This aspect of diversity is partly covered by considering the phylogenetic diversity as analyzed in section 4.3, since related languages tend to share features due to a common ancestor. However, languages change over time and influence each other horizontally, thus motivating an analysis of featural diversity, considering which linguistic features are shared between the languages in the aya dataset.

As the distance between the languages are given by comparing their feature vectors from URIEL+, the result is a 65x65 dissimilarity matrix, with each value denoting the distance between these languages. By only considering the values in the upper triangle and ignoring the zero-distances, an idea of the distribution can be given, as seen in figure 8. As can be seen, the distribution of distances is approximately normal. The mean featural diversity is 0.547 (median = 0.545, sd = .079). This suggests that the data set is fairly balanced, tending towards the language being more different than similar, using 0.5 as a threshold (1 is the maximal dissimilarity and 0 is the minimal dissimilarity).

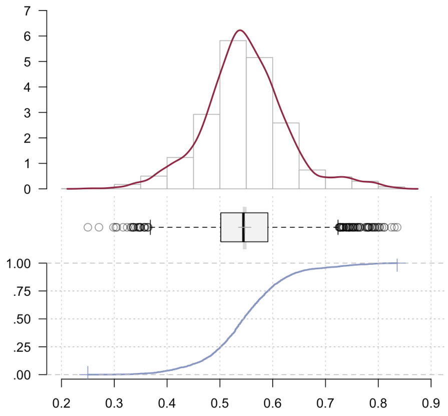


Fig. 8. descriptive plots of the featural distances between languages in the Aya dataset

4 Discussion and Conclusion

I have here investigated the Aya data set to analyze its linguistic diversity. While the Aya dataset aims to reduce linguistic inequality in the world of natural language processing, it falls somewhat short in multiple regards. Altough the relative majority of languages are low-resource languages, high resource languages are more well represented with on average more prompt-completion pairs per language than low-resource language, and mid-resource languages are especially underrepresented. In total, 13 language families are represented in the dataset, but it is heavily skewed towards Indo-European languages, making up almost half of the languages present. When considering the number of prompt completion pairs, the skew is offset somewhat, but Indo-European language still dominate.

A similar situation is observed considering scriptic diversity, with an even larger richness of 19 scripts, but a substantial dominance of the latin script. For both phylogenetic and scriptic diversity, it was interestingly concluded that the diversity increased when considering the distribution of prompt-completion pairs compared to number of languages, suggesting that contributors of non-dominant languages contributed more to the data. With regard to feature diversity, however, the dataset does well in capturing diversity, with a mean featural distance of 0.54 between all language pairs, suggesting that the languages in the dataset are more dissimilar than similar.

As a resource to combat loss of linguistic diversity, the Aya dataset falls substantially short, as it fails to include non-threatened language, only 3 languages of AES-level 2 or above are included. As such, it does not contribute well to provide resources for the languages that might need it the most out of a linguistic diversity perspective. But still, none of this means that the Aya dataset is not a valuable resource. It provides organic data in languages that aren't English, and that is in its own right a significant contribution to the linguistic diversity of the digital world. And while one could wish that the dataset was more diverse, the data was collected on a basis of voluntary contribution, meaning that the diversity is in end effect the result of who decided to contribute. As such, future research needs to try to address why e.g. speakers of mid-resource languages did not decide to contribute more, or why speakers of threatened languages were not reached and allowed to contribute. One could suspect that the decision to not include dialects in the dataset might have influenced this, as many languages which do not have a language-status are threatened, and thus automatically excluded.

In the end, linguistic diversity in the digital space is a fairly unexplored area and the impact of digitization on linguistic diversity has not been evaluated thoroughly. The aya dataset, however, is a significant contributor to highlighting the need for better representation of languages in NLP, and while it does not manage to go the entire way, it certainly is a step in the right direction.

References

1. Aya. `https://cohere.com/research/aya`, accessed: 2025-01-09
2. How many languages are there in the world? `https://www.ethnologue.com/insights/how-many-languages/`, accessed: 2025-01-09
3. Most common languages on the internet. `https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/`, accessed: 2025-01-09
4. Recommendation concerning the promotion and use of multilingualism and universal access to cyberspace. `https://www.unesco.org/en/legal-affairs/recommendation-concerning-promotion-and-use-multilingualism-and-universal-access-cyberspace`, accessed: 2025-01-09
5. Bahji, A., Acion, L., Laslett, A.M., Adinoff, B.: Exclusion of the non-english-speaking world from the scientific literature: Recommendations for change for addiction journals and publishers. Nordic Studies on Alcohol and Drugs **40**(1), 6–13 (2023). https://doi.org/10.1177/14550725221102227
6. Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., Greenhill, S., Hua, X.: Global predictors of language endangerment and

the future of linguistic diversity. Nature ecology evolution **6**, 2 (2022). https://doi.org/10.1038/s41559-021-01604-y

7. Guo, Y., Conia, S., Zhou, Z., Li, M., Potdar, S., Xiao, H.: Do large language models have an english accent? evaluating and improving the naturalness of multilingual llms (2024), `https://arxiv.org/abs/2410.15956`
8. Hammarström, H., Forkel, R., Haspelmath, M., Bank, S.: Glottolog 5.1 (2024), available online at `https://glottolog.org`
9. Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M.: The state and fate of linguistic diversity and inclusion in the nlp world. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. p. 6282–6293. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.560, `https://aclanthology.org/2020.acl-main.560/`
10. Khan, A., Shipton, M., Anugraha, D., Duan, K., Hoang, P.H., Khiu, E., Doğruöz, A.S., Lee, E.S.A.: Uriel+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. arXiv preprint arXiv:2409.18472 (2024)
11. Singh, S., Vargus, F., Dsouza, D., Karlsson, B.F., Mahendiran, A., Ko, W.Y., Shandilya, H., Patel, J., Mataciunas, D., OMahony, L., Zhang, M., Hettiarachchi, R., Wilson, J., Machado, M., Moura, L.S., Krzemiński, D., Fadaei, H., Ergün, I., Okoh, I., Alaagib, A., Mudannayake, O., Alyafeai, Z., Chien, V.M., Ruder, S., Guthikonda, S., Alghamdi, E.A., Gehrmann, S., Muennighoff, N., Bartolo, M., Kreutzer, J., Üstün, A., Fadaee, M., Hooker, S.: Aya dataset: An open-access collection for multilingual instruction tuning (2024), `https://arxiv.org/abs/2402.06619`
12. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. p. 483–498. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.41, `https://aclanthology.org/2021.naacl-main.41/`