DistaLs: A Comprehensive Collection of Language Distance Measures

Rob van der Goot[1], Esther Ploeger[2], Verena Blaschke[3], Tanja Samardžić[4]
[1]IT University of Copenhagen, robv@itu.dk
[2]Aalborg University, espl@cs.aau.dk
[3]LMU Munich & Munich Center for Machine Learning, blaschke@cis.lmu.de
[4]University of Zurich, tanja.samardzic@uzh.ch

Abstract

Languages vary along a wide variety of dimensions. In Natural Language Processing (NLP), it is useful to know how "distant" languages are from each other, so that we can inform NLP models about these differences or predict good transfer languages. Furthermore, it can inform us about how diverse language samples are. However, there are many different perspectives on how distances across languages could be measured, and previous work has predominantly focused on either intuition or a single type of distance, like genealogical or typological distance. Therefore, we propose DistaLs, a toolkit that is designed to provide users with easy access to a wide variety of language distance measures. We also propose a filtered subset, which contains less redundant and more reliable features. DistaLs is designed to be accessible for a variety of use cases, and offers a Python, CLI, and web interface. It is easily updateable, and available as a pip package. Finally, we provide a case-study in which we use DistaLs to measure correlations of distance measures with performance on four different morphosyntactic tasks.[1]

1 Introduction

Since language resources are limited for the majority of the world's languages (Joshi et al., 2020), multi-lingual Natural Language Processing (NLP) models are highly relevant. Knowing how diverse a set of languages is can be important for many crucial steps in the NLP pipeline, e.g. language selection to ensure broad coverage, predicting which source language to use for a target language, or predicting performance. However, cross-lingual performance of NLP models can be influenced by many factors. While quantitative language-level measures of distance have shown to be good predictors for performance (e.g. Lin et al., 2019; Lauscher et al., 2020), it has also been shown that factors such as data size or pre-training exposure (token overlap) can be stronger predictors than structural features (de Vries et al., 2022) or that different factors matter for different NLP tasks (Blaschke et al., 2025). However, some cross-lingual results are harder to explain, e.g. when transfer works well between intuitively distant languages such as Indonesian and Irish (Lynn et al., 2014) or when a particular language type turns out to be suitable for transfer to various target languages (Pelloni et al., 2022).

In practice, intuition, relying on properties such as language family, script, or geolocation, is often used to select a source language to transfer from. Another line of work uses more objective typological distances, commonly from lang2vec (Littell et al., 2017), to gauge language similarities, but this option too has several known issues (Toossi et al., 2024; Khan et al., 2025).

In this context, we collect and compare language distance measures more systematically. To this end, we create DistaLs, a toolkit that provides users with properties of languages, and a wide variety of language distances measuring different dimensions of diversity. We foresee at least three main use cases for DistaLs: 1) selecting languages to transfer from; 2) quickly estimating which measure of language distance has a correlation to performance; 3) measuring diversity in language selection.

DistaLs is based on the ISO 639-3 language classification standard. For each language in ISO 639-3, we aggregate language information from a variety of data sources, and use existing distance metrics, or design new ones where necessary. We complement these with text-based features. DistaLs is updateable from the original sources with a single command. We also provide an exploratory data analysis of our data sources, correlations across different distance measures, and a case-study on Universal Dependencies (UD) parsing, where we

[1] Code and data: https://bitbucket.org/robvanderg/distals

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 307–318
November 4-9, 2025 ©2025 Association for Computational Linguistics

check which distance measure correlate with performance for four morphosyntactic tasks.

2 Distance Measures

The distances collected in DistaLs can be roughly grouped into four categories: meta-data, typological distances, wordlist-based, and text-based. For each of these sources, we only collect information for valid ISO 639-3 language codes. If only the language name is available (the case for “nlp_fate”) we check if Glottolog or ISO 639-3 includes the exact match of the name. Some of the macro-languages have a variant that is the majority variant which can be assumed to be meant by the user, in these cases we convert the language label with a manually curated lookup table (Appendix A), other macro-languages are not supported due to their internal diversity. The user is informed when an automatic conversion is done, and macro-label conversion can be disabled with `–disable_macro_conversion`.

We focus on features that are available for at least 1,000 languages. Our metrics aim to cover a variety of dimensions, but within each dimension, they might lack specificity. We have for example a single feature representing phoneme inventory distance; if one is interested in more fine-grained information for a specific usecase (i.e. does a language include stressed consonants), we refer to the original data sources we included, which are designed for this exact purpose.

All metrics are converted to reflect distance (as opposed to similarity) and are normalized to be between 0.0 and 1.0. The distances are also bi-directional, so the distance of language 1 to language 2 is the same as from language 2 to language 1. The information on which the distance measures are based is also indexed and quickly retrievable so that the measures can easily be implemented differently, or the features can be used for other usecases. We provide an overview of the features in Table 1. We also provide an example of how they could be aggregated, as taking a naive average over all features will lead to undesirable weighting of certain categories, and include conflicting/redundant sources. Our selection is based on coverage, redundancy (Section 4), and quality. However, it should be noted that this is not the only possible way to aggregate the different features, and different situations will require different selections.

Category	Feature	Source	Coverage
Metadata	wiki_size	Wikipedia	7,856
	nlp_state	State and fate	2,269
	speakers	LinguaMeta	5,539
	AES	Glottolog	7,725
	loc	Glottolog	7,629
Typology	lang2vec	URIEL	3,910
	lang2vec_knn	URIEL	3,910
	PHOIBLE	PHOIBLE	2,078
	grambank_all	Grambank	2,326
	grambank_.*	Grambank	2,326
	glot_tree	Glottolog	7,856
	scripts	LinguaMeta, GlotScript	6,431
Wordlists	**ASJP**	ASJP	6,117
	concepts	Conceptualizer	1,274
Text-driven	whitespace	LTI LangID	2,525
	punctuation	LTI LangID	2,525
	char_distr.	LTI LangID	2,525
	textcat	LTI LangID	2,525

Table 1: All features provided by DistaLs. Bold: recommended to use for average for category. grambank_.* refers to the sub-categories within Grambank. Coverage of wiki_size is large, becuase if there is no Wikipedia for a language, we assign the size 0 (happens for 7,570 languages).

Below, we describe each feature. Appendix B contains further implementation details.

2.1 Metadata

Wikipedia We use the number of articles per language as a statistic of a language, which can be considered a proxy to online presence of languages. The distance metric is the proportional difference in size: $1 - min(size1, size2)/max(size1, size2)$.

State and Fate The amount of resources available for a language can be a strong predictor for performance for universally trained models. There are many different catalogues of resour-ces, and also many less standardized resources. We here use the categorization provided by Joshi et al. (2020); they divide languages into one of 6 groups based on the availability of raw text data, and annotated NLP datasets. The groups are ranked, so we use the distance in rank as our metric: $(max - min)/5$.

LinguaMeta LinguaMeta (Ritchie et al., 2024) is an effort to calculate metadata of languages into a unified format. It is combining a variety of existing resources, including manual corrections/additions where possible. We extract the **number of speakers** and **scripts** from LinguaMeta. The number of speakers only counts L1 speakers, and uses a variety of sources; CLDR,[2] Wikipedia, and Google-

[2] cldr.unicode.org

internal information. The scripts are mostly from internal Google data which is used for keyboard selection in their products. We complement the scripts information with data from GlotScript (Kargaran et al., 2024), if there is more than 1 script from both sources, we only use the intersection. For the scripts, we use $1 - \%overlap$ as metric, and for the speakers we use the same formula as for the Wikipedia size.

Glottolog (Hammarström et al., 2024) is a database containing information and references for different languages, dialects, and families. We here use Agglomerated Endangerment Status (AES). The endangerment status has 6 ranked classes, we use the same formula as for "state and fate".

2.2 Typology

lang2vec We use the average values over all data sources for the syntax, phonology, and inventory categories from the URIEL database through the lang2vec toolkit (Littell et al., 2017), which is in turn based on WALS (Dryer and Haspelmath, 2013), SSWL (Collins and Kayne, 2011), Ethnologue (Campbell and Grondona, 2008), and PHOIBLE (Moran and McCloy, 2019). These values are concatenated and used as feature vector for a language. We additionally use lang2vec's imputation method for missing features based on a k-nearest-neighbours selection of similar languages. The distance metric is cosine distance, where we remove features from both languages if a feature is missing for one of the languages. Reproducibility of the lang2vec distances is non-trivial (Toossi et al., 2024; Khan et al., 2025), so we calculate the cosine distance based on their representation vectors ourselves.

PHOIBLE (Moran and McCloy, 2019) is a cross-linguistic phonological database. It contains phoneme inventories based on International Phonetic Alphabet (International Phonetic Association, 2005) collected from a wide variety of sources. We use the set of the defined *GlyphId*s for each language as a representation, and use the % of overlap between these sets as a distance metric.

Grambank (Skirgård et al., 2023) is a database containing morphosyntactic information about languages. It contains 195 features with a higher language coverage compared to lang2vec. Similar to Ploeger et al. (2024), we first binarize the data, and then we take the euclidean distance ignoring empty features. Languages with fewer than 25% of the features covered are removed. We divide this distance by the square root of the total number of features to make it range between 0–1.

Glottolog Besides the AES (described above), we also extract family trees from Glottolog. We calculate a distance based on the position in the tree structure. If two languages are in different language families, the distance is maximal (1.0), if the languages are in the same tree, we calculate the number of overlapping edges divided by the depth of the deepest language of the two.

2.3 Wordlist-based metrics

ASJP Automated Similarity Judgment Program (ASJP) is a database containing standardized word lists of concepts in many languages (Wichmann et al., 2022). Their word list is based on the Swadesh lists (Swadesh, 1955). Both lists are created to cover concepts that are expected to exist in cultures and languages all over the world. For each concept, ASJP collected a phonetic description of the concept in each language. We follow their original implementation (Bakker et al., 2009) and use average normalized Levenshtein distance over the phonetic sequences of the concepts.

Conceptualizer Liu et al. (2023) use 51 concepts from the bible combined with 32 concepts defined in Swadesh lists to compare representations of different concepts in different languages. They model the concepts as a bipartite graph, in which a concept (represented as a set of English strings) links to all correlated translations. We use the language distance metric as proposed by Liu et al. (2023); the cosine distance over the representations of each concept for each language, where a concept representation is the number of steps a concept needs to get to the English concept.

2.4 Text-driven distances

We use a combination of the GlotLID (Kargaran et al., 2023) and the LTI LangID corpus (Brown, 2014) as our source data because of their large language coverage. These datasets are combinations of a variety of sources, but the majority of data comes from Wikipedia and a variety of bible translation sets. We take 1,000 lines per script (equal amount of each source) of each available data source to represent a language. We apply NFC normalization before collecting our features.

Character categories There are two categories of characters that are commonly used across different scripts: **whitespace characters**, and **punctuation characters**. The amount of usage of these categories can be used to distinguish languages with whitespace-based writing systems, lengths of words within these, and the amount of punctuation information. For both categories, we use the definition from the huggingface library (specifically, the `_is_whitespace` and `_is_punctuation` function) for classification. We then convert the percentages to a distance score through the following formula: $1 - min(prob1, prob2)/max(prob1, prob2)$

Character distribution distance We first extract the character distributions from each language. Following the original LTI LangID Corpus we use UTF-8 encoding for defining characters, and for each character estimate their frequency as a probability. We then use the Jensen-Shannon distance over the union of the character sets of both languages.

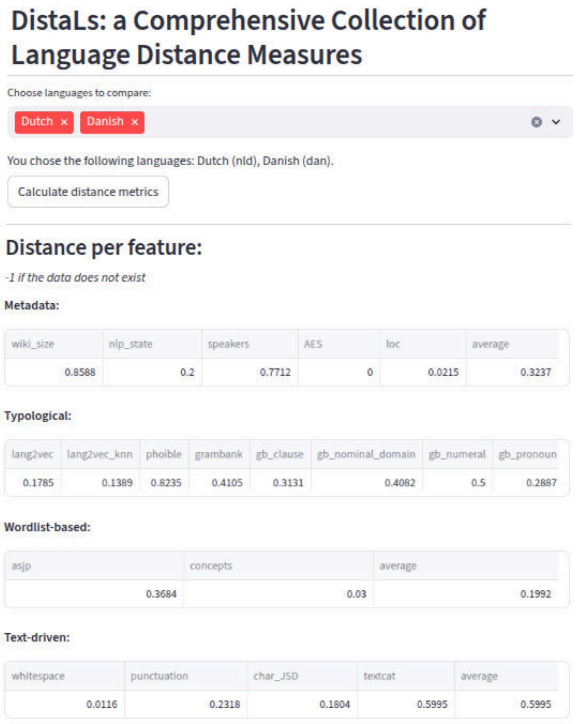
Textcat distance Cavnar et al. (1994) proposed to use n-gram frequency-lists for language classification. Specifically, they extract the 300 most common 1–5 character n-grams, and sort them by frequency to represent a language. For an input, they then create a similar frequency list, and calculate a distance to the representation of each language in the training data to obtain a similarity ranking. We use the same setup to calculate distances across texts of different languages, but use the 400 most common n-grams, following van Noord (1997).

3 Interface

DistaLs provides language data in the following three ways:

- Pre-calculated distances: we provide all metrics for all language pairs in csv files.
- Database with language information: this will automatically be downloaded by the package if it is not found. The code uses this database to calculate the distances.
- Scrape the data: download all data sources with a single bash script, and then use this to populate a database (as described above). This allows for easy updating of all data sources.

DistaLs provides three different interfaces (besides the pre-calculated distances), which are described in the next sections.



DistaLs: a Comprehensive Collection of Language Distance Measures

Choose languages to compare:

You chose the following languages: Dutch (nld), Danish (dan).

Distance per feature:

-1 if the data does not exist

Metadata:

wiki_size	nlp_state	speakers	AES	loc	average
0.8588	0.2	0.7712	0	0.0215	0.3237

Typological:

lang2vec	lang2vec_knn	phoible	grambank	gb_clause	gb_nominal_domain	gb_numeral	gb_pronoun
0.1785	0.1389	0.8235	0.4105	0.3131	0.4082	0.5	0.2887

Wordlist-based:

asjp	concepts	average
0.3684	0.03	0.1992

Text-driven:

whitespace	punctuation	char_JSD	textcat	average
0.0116	0.2318	0.1804	0.5995	0.5995

Figure 1: Screenshot of the web interface.

3.1 Web interface

We provide an online interface to DistaLs on `https://distals.streamlit.app/`. The user is presented with a text field, in which language names can be added (search results will pop up for easy selection after typing). After selecting a number of languages, the users clicks on the button or presses enter, and after a short wait the distances will be shown per category (see Figure 1).

3.2 Command-line Interface (CLI)

DistaLs is available as a pip package. After installing the package, the main functionality is accessed through the parameter `-langs`. The user specifies a list of languages, which can be ISO639-3 codes, ISO639-2 codes, or language names (which will be converted according to the procedure described in Section 2). DistaLs will first print all the information it has available for each language. If there are two languages defined, it will then print all the distances for each category, including their average. When information for a feature is not available for both languages, it will print a –1 value. If more than two languages are included, it will print language×language matrices. An example of usage and its output can be seen in Figure 2.

The CLI command also provides an interface

```
$ distals --langs fry dan
loading from: ./distals-db.pickle.gz
7856 languages loaded
========================================
Information for fry
wiki_size: 57,027
nlp_state: 1. The Scraping-Bys
speakers: 740,000
AES: 5. not endangered
loc: (5.86091, 53.143)
lang2vec: [1.0, 1.0, 0.0, ...]
lang2vec_knn: [1.0, 1.0, 0.0, ...]
phoible: ['0061', '0061+0069', '0061+0075', ...]
grambank: {'GB020': 1, 'GB021': 1, 'GB022': 1, ...}
glot_tree: ["'Western Frisian [west2354][fry]-1-'", "'
     Westlauwers-Terschelling Frisian [west2902]'", "'Modern
      West Frisian [mode1264]'", ...]
scripts: {'latn'}
asjp: [['1', 'ik'], ['2', 'do, yo'], ['3', 'vEi'], ...]
whitespace: 0.160835
punctuation: 0.031726
char_JSD: {' ': 0.1608, 'e': 0.1195, 'n': 0.0754, ...}
textcat: [' ', 'e', 'n', ...]

========================================
Information for dan
wiki_size: 308,911
nlp_state: 3. The Rising Stars
speakers: 5,510,600
AES: 5. not endangered
loc: (9.36284, 54.8655)
lang2vec: [1.0, 0.0, 0.0, ...]
lang2vec_knn: [1.0, 0.0, 0.0, ...]
phoible: ['0061', '0062+0325', '0062+0325+02B0', ...]
grambank: {'GB020': 1, 'GB021': 1, 'GB022': 1, ...}
glot_tree: ["'Danish [dani1285][dan]-1-'", "'South
     Scandinavian [sout3248]'", "'North Germanic [nort3160
     ]'", "'Northwest Germanic [nort3152]'", "'Germanic [
     germ1287]'", "'Classical Indo-European [clas1257]'", "'
     Indo-European [indo1319]'"]
scripts: {'latn'}
asjp: [['1', 'yoy'], ['2', 'du'], ['3', 'vi'], ...]
whitespace: 0.156298
punctuation: 0.028514
char_JSD: {' ': 0.1563, 'e': 0.1249, 'r': 0.0675, ...}
textcat: [' ', 'e', 'r', ...]

========================================
Distances between fry and dan (-1 if feature not available)
METADATA
wiki_size: 0.8154
nlp_state: 0.4000
speakers: 0.8657
AES: 0.0000
loc: 0.0149
average: 0.5203

TYPOLOGY
lang2vec: 0.1598
lang2vec_knn: 0.1204
phoible: 0.8148
grambank: 0.3841
gb_clause: 0.3742
gb_nominal_domain: 0.3482
gb_numeral: 0.5000
gb_pronoun: 0.0000
gb_verbal_domain: 0.4644
glot_tree: 0.5325
scripts: 0.0000
average: 0.5995

WORDLISTS
asjp: 0.3397
concepts: 0.0400
average: 0.1898

TEXTBASED
whitespace: 0.0282
punctuation: 0.1012
char_JSD: 0.1979
textcat: 0.5859
average: 0.3919
```

Figure 2: Example output of DistaLs. It first reports information about the provided language(s), and then reports all features per category.

to the updating of the database. There are three separate arguments (for language labels and names, databases, and text-based features), which can be used separately or jointly. The resulting database is stored in a dictionary which is saved in a compressed pickle file. The toolkit will automatically download a recent database from the repository if –database_path is not specified. The code can also be ran directly from the repository (python3 src/distals/distals.py) without installation.

```
>>> from distals import distals
>>> model = distals.Distals()
>>> model.get_dists('nld', 'cmn')
{'metadata': {'wiki_size': 0.99378,
              'nlp_state': 0.2,
              'speakers': 0.98131,
              'AES': 0.0,
              'loc': 0.39121,
              'average': 0.39377},
 'typology': {'lang2vec': 0.31654,
              'lang2vec_knn': 0.33795,
              'phoible': 0.82278,
              'grambank': 0.58478,
              'gb_clause': 0.55470,
              'gb_nominal_domain': 0.59761,
              'gb_numeral': 0.0,
              'gb_pronoun': 0.64550,
              'gb_verbal_domain': 0.60302,
              'glot_tree': 1.0,
              'scripts': 0.66667,
              'average': 0.80252},
 'wordlists': {'asjp': 0.49636,
               'concepts': 0.08,
               'average': 0.28818}
 'textbased': {'whitespace': 0.21244,
               'punctuation': 0.67855,
               'char_JSD': 0.54401,
               'textcat': 0.87235,
               'average': 0.87235}
}
```

Figure 3: Example output of DistaLs when used in Python.

3.3 Python

For easy integration into other projects and codebases, we also provide a python interface. The information stored for each language is directly available from the DistaLs database, which is a python dictionary in a pickle file. One can also import DistaLs to get direct access to the distances, which can be returned as a list or as a dictionary (containing the four main categories as a hierarchy). The language names and code conversion scripts are also available after loading DistaLs. Example usage is shown in Figure 3. Updating the DistaLs database is done by first running a bash script that downloads/updates the data, and then the python package has functionality to update through a single command.

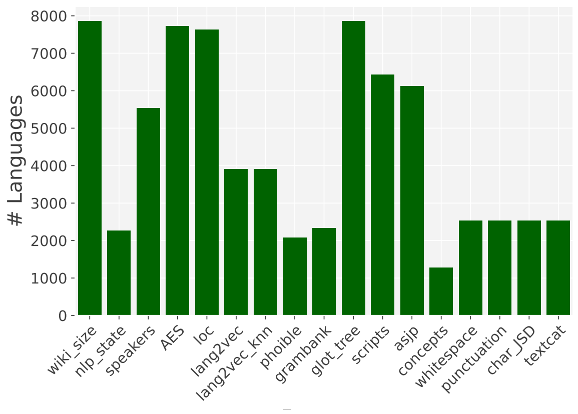


Figure 4: The number of languages (y-axis) supported by each feature (x-axis).

4 Exploratory Data Analysis

To get a clearer picture of the coverage, we count the number of features supported by each language, and plot the number of languages with N supported features (Figure 4). This shows that there are a few features for which almost all languages are covered. These are mainly meta-features and genealogical information. Many features have a coverage around 2,000 languages, which often have a large overlap in languages. DistaLs contains 437 languages that have information for all 17 features.

We normalized all features to have a value between 0 and 1, but different features might still have different distributions within this range. Hence, for each feature, we sort the values of all language-pairs, disregarding the missing values. We then plot the scores to get an overview of how the distributions differ. Results (Figure 5) show that there is indeed a disparity in the distributions of the probabilities, with AES having lower distances, because there are only 6 possible labels, and the text-based feature (textcat) and typology based features (PHOIBLE, Grambank, and Glot_tree) having larger distances for most pairs.

Some of the features have a similar goal, so they can be expected to correlate. For example, the main categories in the typology category all aim to capture typological distances. If features have a large overlap, some of them can be left out for the sake of simplicity and efficiency. We therefore perform a correlation study for feature-pairs across all language-pairs. For each pair of features, the languages included can be different (based on data availability). Hence, the results across features are not directly comparable, but should give a rough idea of which features contain similar information.

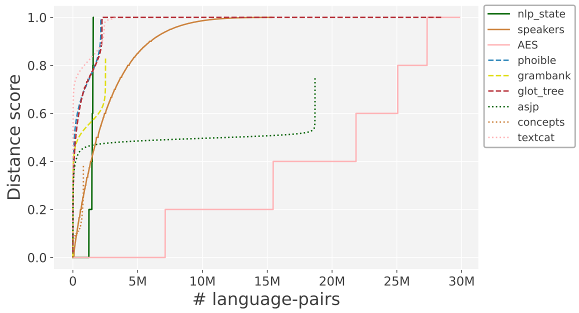


Figure 5: The cumulative probabilities of selected features.

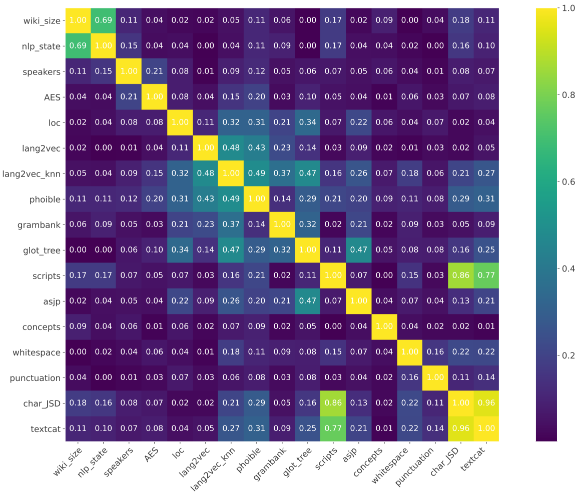


Figure 6: Pearson correlations across all features (except the grambank sub-categories). P-values in Appendix D.

Results (Figure 6) show only a few strong Pearson correlations (i.e. > 0.5), which are all within the main categories we defined in Section 2, except script which has a strong correlation to the text-based distances. Within the typology category, there are many moderate correlations (0.3–0.5), and across categories there are mostly correlations close to 0, where mainly the text-based distances (char JSD and textcat) have some weak correlations (~0.2) across the other features.

5 Case Study

Inspired by the studies of de Vries et al. (2022), Samardžić et al. (2022) and Blaschke et al. (2025), who analyze the effect of certain language distances on downstream NLP model performance, we execute a similar case study on cross-lingual transfer with our extended feature set. We train a multi-task model on the first 10k words of each UD v2.15 treebank (Nivre et al., 2020) that has a training split resulting in 64 source and 126 target languages.

	UPOS	Dep	UFeats	Lemma
nlp_state	−0.42 *	−0.37 *	−0.08 *	−0.12 *
speakers	−0.24 *	−0.20 *	−0.02	−0.05 *
AES	−0.32 *	−0.23 *	−0.03	−0.05 *
loc	−0.29 *	−0.28 *	0.02	0.02
phoible	−0.19 *	−0.17 *	0.02	−0.02
grambank	−0.42 *	−0.45 *	−0.20 *	−0.11 *
glot_tree	−0.27 *	−0.27 *	−0.03	−0.07 *
asjp	−0.26 *	−0.28 *	−0.14 *	−0.11 *
concepts	−0.32 *	−0.31 *	0.01	−0.10 *
textcat	−0.21 *	−0.23 *	−0.06 *	−0.06 *

Table 2: Pearson's r between language distance and accuracy or labelled attachment score (Dep). *p<0.05.

We use all treebanks for part-of-speech tagging (UPOS) and dependency parsing (Dep), and add lemmatization (Lemma) and morphological tagging (UFeats) when available. For languages with multiple treebanks, we average the results. We use the MaChAmp toolkit (van der Goot et al., 2021) v0.4.2 with default hyperparameters. We train with XLM-R large (Conneau et al., 2020) and Glot500 (Imani et al., 2023) as an encoder, and report the average results (trends were highly similar across models).

We evaluate the correlations on the subset of most informative features from Section 2 (bold in Table 2). The correlations (Table 2) are weak for the morphological tasks (UFeats/Lemma), but much stronger for the syntactic tasks (UPOS/Dep), whose performance can be better predicted with our distance features. Interestingly, the most strongly correlated distances are scattered across our distance categories.

6 Comparison to Other Toolkits

We compare existing toolkits for estimating language diversity in Table 3. DistaLs covers the most categories of distances, but some other toolkits have more functionality within a specific category. For example, Delta[3] and LangDive (Samardzic et al., 2024) focus on the diversity of datasets with respect to linguistic and syntactic information. Within the domain of syntax, they will provide a much more granular perspective on distance, but at the cost of a lower language coverage. QwanQwa[4] instead focuses on aligning metadata across different language-code systems, and typdiv (Ploeger et al., 2024) provides metrics for assesing typological diversity. LangRank (Lin et al., 2019) directly focuses on predicting transfer performance, providing a variety of metrics.

7 Conclusion

We propose DistaLs, a toolkit that aggregates language information from a variety of sources, and provides distance measures based on this. DistaLs contains a variety of easy to use interfaces, a webpage, csv files, python, and a command-line interface. It includes a wide variety of measures covering a variety of dimensions of "distance", all with a high language coverage. We showed its usefulness by reporting correlations of the features with four morphosyntactic tasks. Based on this, we conclude that syntactic tasks have higher correlations than morphological tasks (i.e., performance transfer is easier to predict), and that no features are close to a perfect correlation, however, features with a moderate correlations are quite diverse.

Acknowledgements

We thank Barbara Plank, Joakim Nivre, Jörg Tiedemann, Steven Moran, and the members of NLP-north for their feedback. We also thank Ralf Brown for helping with cleaning LTI-LangID.

This work is supported by the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

[3] https://gitlab.lisn.upsaclay.fr/esteve/delta

[4] https://github.com/WPoelman/qwanqwa

Toolkit	Focus	Lang. cat.	Coverage	Typology	Metadata	Wordlists	Textbased	Updatable	Interface
Lang2vec	typology	ISO 639-3	4,005	✓	✓	✗	✗	✗	python, CLI
Delta	syntactic diversity	—	156	✓	✗	✗	✗	✓	python, c
LangDive	dataset diversity	ISO 639-3	—	✗	✗	✗	✗	✓	python
QwanQwa	metadata	many	7,511	✓	✓	✗	✗	✗	python
typdiv	diversity	glottocodes	—	✓	✗	✗	✗	✓	python, CLI
LinguaMeta	metadata	BCP-47,ISO 639-3	7,512	✓	✓	✗	✗	✗	tsv file
LangRank	multiple	ISO 639-2	—	✓	✓	✗	✓	✗	python
DistaLs	multiple	ISO 639-3	1,271–7,855	✓	✓	✓	✓	✓	python, CLI, web

Table 3: Comparison of existing toolkits for measuring language diversity. The main categories refer to the ones described in Section 2. A '—' in coverage means that these toolkits are supposed to be used with datasets to estimate the distances. Updateable refers to automatically updateable (i.e. through a single command).

Limitations

We limited the language coverage to the ISO 639-3 standard, as it is one of the most widely used set of language labels. However, this standard is known to have biases (Morey et al., 2013). At the same time, we ignore in-language variation, and we make the (flawed) assumption that the textual data we use serves as a proxy for a representation of the language as a whole. Also, the data sources we include are carefully chosen to have a wide coverage, but there is definitely more information for high-resource languages. The case-study on the UD data also has a biased sub-selection of languages, which have a higher coverage in Western languages.

Furthermore, each feature can be seen as an abstraction to actual diversity as it occurs within a language. This is a necessary step to take when providing smaller numbers of distance metrics, but it is obscuring a lot of potentially interesting information. If more detailed information on a specific dimension of language is required, we refer to the original data sources.

While we were careful in selecting the information sources, with data on this scale there are undoubtedly errors in the data. We have done an automatic and manual correction of some of the LTI-LangID data in cooperation with Ralf Brown, and have done inspection, cleaning and merging of the other sources where possible, but there are of course many data points that are hard to verify. The text-based features are also biased in domain. Most of the data comes from bible translations.

References

Dik Bakker, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W Holman. 2009. Adding typology to lexicostatistics: A combined approach to language classification.

Verena Blaschke, Masha Fedzechkina, and Maartje ter Hoeve. 2025. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8653–8684, Vienna, Austria. Association for Computational Linguistics.

Cecil H. Brown, Eric W. Holman, Søren Wichmann, and Viveka Velupillai. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *Language Typology and Universals*, 61(4):285–308.

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

Lyle Campbell and Verónica Grondona. 2008. Ethnologue: Languages of the world. *Language*, 84(3):636–641.

William B Cavnar, John M Trenkle, and 1 others. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, page 14, Las Vegas, NV.

Chris Collins and Richard Kayne. 2011. Syntactic structures of the world's languages. New York University, New York.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.0.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

International Phonetic Association. 2005. International phonetic alphabet.

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

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.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. GlotScript: A resource and tool for low resource writing system identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7774–7784, Torino, Italia. ELRA and ICCL.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

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

Yihong Liu, Haotian Ye, Leonie Weissweiler, Philipp Wicke, Renhao Pei, Robert Zangenfeind, and Hinrich Schütze. 2023. A crosslingual investigation of conceptualization in 1335 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12969–13000, Toronto, Canada. Association for Computational Linguistics.

Teresa Lynn, Jennifer Foster, Mark Dras, and Lamia Tounsi. 2014. Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *Proceedings of the First Celtic Language Technology Workshop*, pages 41–49, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

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

Stephen Morey, Mark W. Post, and Victor A. Friedman. 2013. The language codes of iso 639: A premature, ultimately unobtainable, and possibly damaging standardization.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Olga Pelloni, Anastassia Shaitarova, and Tanja Samardzic. 2022. Subword evenness (SuE) as a predictor of cross-lingual transfer to low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7428–7445, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, and Johannes Bjerva. 2024. A principled framework for evaluating on typologically diverse languages. *Preprint*, arXiv:2407.05022.

Sandy Ritchie, Daan van Esch, Uche Okonkwo, Shikhar Vashishth, and Emily Drummond. 2024. LinguaMeta: Unified metadata for thousands of languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10530–10538, Torino, Italia. ELRA and ICCL.

Tanja Samardzic, Ximena Gutierrez, Christian Bentz, Steven Moran, and Olga Pelloni. 2024. A measure for transparent comparison of linguistic diversity in multilingual NLP data sets. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3367–3382, Mexico City, Mexico. Association for Computational Linguistics.

Tanja Samardžić, Ximena Gutierrez-Vasques, Rob van der Goot, Max Müller-Eberstein, Olga Pelloni, and Barbara Plank. 2022. On language spaces, scales and cross-lingual transfer of UD parsers. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 266–281, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, and 86 others. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16).

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

Hasti Toossi, Guo Huai, Jinyu Liu, Eric Khiu, A. Seza Doğruöz, and En-Shiun Lee. 2024. A reproducibility study on quantifying language similarity: The impact of missing values in the URIEL knowledge base. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 233–241, Mexico City, Mexico. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multitask learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Gertjan van Noord. 1997. Textcat.

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

Appendix

A Macro Language Conversion

Table 4 reports the list of manually added conversions of macro labels.

Macro	Language
est	ekk
zho	cmn
grn	gug
toki	tok
nep	npi
lav	lvs
ara	arb
ori	ory
msa	zlm
kom	kpv

Table 4: Conversion of macro labels to language labels.

B Implementation Details for Distance Measures

Below, we describe implementation details of version 0.1.1 of DistaLs.

B.1 Metadata

Wikipedia We collect the page counts from the webpage `https://en.wikipedia.org/wiki/List_of_Wikipedias`. Where possible, we use the `language` attribute from the column with sample text in each wiki's language. For the relatively few wikis where this information is not available, we derive the language from the wikicode. We manually verified that this fallback option results in the correct ISO639-3 codes as of September 12, 2025.

B.2 Typology

lang2vec We use the `syntax_average+phonology_average+inventory_average` category for original features, and the `syntax_knn+phonology_knn+inventory_knn` category for the KNN completed features.

We only compare lang2vec vectors for language pairs where at least 25% of the features have values for both languages. Users can change this threshold.

PHOIBLE Some languages have multiple phoneme inventories on PHOIBLE (coming from different sources and/or describing different dialects/sociolects).[5] If this applies to one or both languages, we calculate the distances between each inventory of the first language and each inventory of between each pair of the second language, and return the minimum distance. We do not take into account the allophone information.

Grambank We only compare Grambank feature vectors for language pairs where at least 25% of the features have values for both languages. Users can change this threshold.

Glottolog Where pseudo-families[6] are used in Glottolog for bookkeeping purposes, we remove the pseudo-families. For instance, the family tree path for German Sign Language is originally *German Sign Language – DSGic – L1 Sign Language – Sign Language*, but we remove the last two nodes.

B.3 Wordlist-based metrics

ASJP Some languages have multiple word form entries for one concept (because a wordlist contains multiple entries for a concept, and/or because ASJP contains multiple wordlists with the same language code). In such cases, we use the word form with the lowest Levenshtein distance to the other word form it is compared to.

We ignore vowel and consonant modifiers (cf. Brown et al., 2008) when calculating Levenshtein distances. In addition to normalizing the Levenshtein distance by the length of the longer word, we normalize it by the average distance between entries with *different* meanings in the two wordlists being compared. We only compare ASJP wordlists for language pairs where at least 25% of the concepts have word form entries for both languages. Users can change this threshold.

C Licenses

Licenses for each data source are listed in Table 5. DistaLs is released under the CC BY-SA 4.0, as it is required by some of the included data sources.

D P-values of Correlations Across Features

Figure 7 has the same shape as Figure 6 in the paper, but contains the p-values instead of the actual

[5] `https://phoible.org/faq#why-do-some-languages-have-multiple-entries-in-phoible`

[6] `https://glottolog.org/meta/glossary#sec-pseudofamilies`

Category	Feature	Source	License
Metadata	wiki_size	Wikipedia	CC BY-SA
	nlp_state	state and fate	—
	speakers	LinguaMeta	CC BY-SA 4.0
	scripts	LinguaMeta, GlotScript	MIT License
	AES	Glottolog	CC BY 4.0
	loc	Glottolog	CC BY 4.0
Typology	lang2vec	URIEL	CC BY-SA 4.0
	lang2vec_knn	URIEL	CC BY-SA 4.0
	PHOIBLE	PHOIBLE	GPL 3
	grambank_all	Grambank	CC BY 4.0
	grambank_.*	Grambank	CC BY 4.0
	glot_tree	Glottolog	CC BY 4.0
Wordlists	ASJP	ASJP	CC BY 4.0
	concepts	Conceptualizer	—
Text-driven	whitespace	LTI LangID	CC
	punctuation	LTI LangID	CC
	char_distr.	LTI LangID	CC
	textcat	LTI LangID	CC

Table 5: Licenses for all data sources included in DistaLs.

correlation. The values are generally very low, this is because of the large data size.

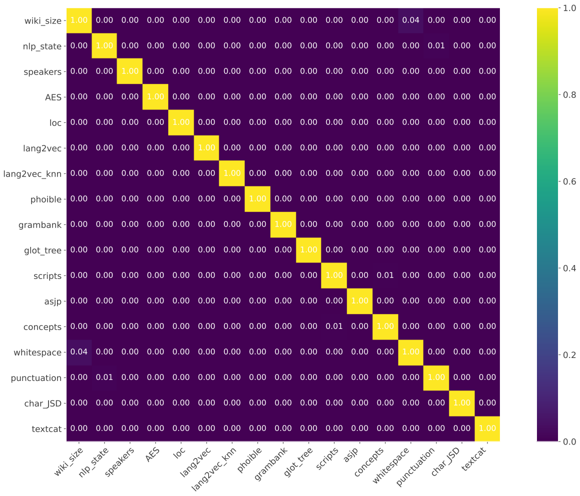


Figure 7: P-values of correlations across features.