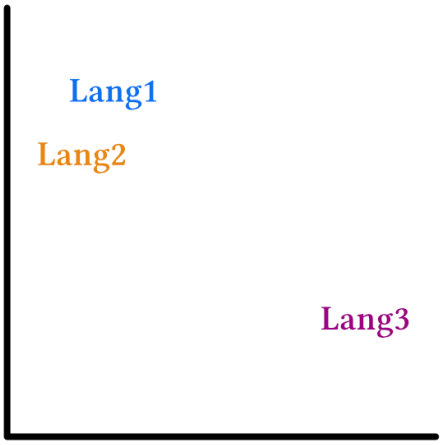
DistaLs: A Comprehensive Collection of Language Distance Measures

Rob van der Goot, Esther Ploeger, Verena Blaschke, and Tanja Samardžić

Typology



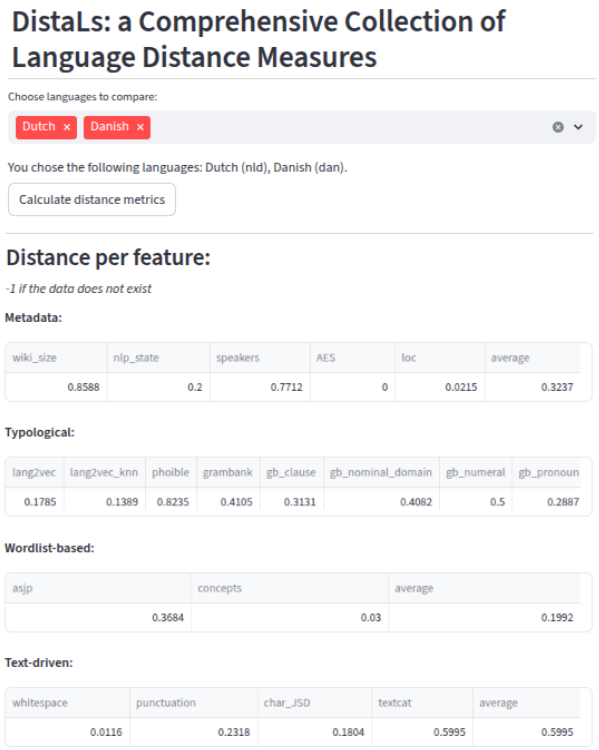
Lexical

Languages can differ among many dimensions!

Why do we need language distances?

- It is super interesting
- Pick good transfer languages in cross-lingual NLP
- Measure correlation between performance and language distance
- Ensure a diverse language sample in experiments

Online interface



DistaLs: a Comprehensive Collection of Language Distance Measures

Choose languages to compare:

Dutch × Danish ×

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

Calculate distance metrics

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

Python

```
>>> 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}
}
```

CLI

```
$ 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]!', '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', ...]
```

```
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
```

How to use: `https://distals.streamlit.app/` `pip3 install distals` `https://bitbucket.org/robvanderg/distals`