Model 1 (Python version)

October 22, 2023

1 Model 1: A probabilistic model of segment borrowability

This report includes supplementary materials for:

`Operationalizing borrowability: A case study from phonological segments`

[6]:

```python
from collections import defaultdict, Counter
import pandas as pd
```

[7]:

```python
def load_cldf_dataset(path_to_values, path_to_languages):
    values = pd.read_csv(path_to_values)
    languages = pd.read_csv(path_to_languages)
    return pd.merge(left = values, right = languages, how="left",
                    left_on="Language_ID", right_on="ID")
```

A helper routine for computing typological frequencies.

When several inventories (doculects) are available for a single language, we collapse the inventories in them (i.e. take their union).

[8]:

```python
def get_frequencies_w_inventory_collapsing(dataset):
    glottocode_to_inventory = defaultdict(set)
    for row in dataset.itertuples():
        if not pd.isnull(row.Language_ID):
            glottocode_to_inventory[row.Language_ID].add(row.Value)
    print(f'{len(glottocode_to_inventory)} languages')
    frequencies_absolute = Counter()
    for segments in glottocode_to_inventory.values():
        for segment in segments:
            frequencies_absolute[segment] += 1
    frequencies_relative = {
        segment: count / len(glottocode_to_inventory)
        for segment, count in frequencies_absolute.items()
    }
    return frequencies_absolute, frequencies_relative
```

First, load the data from the CLDF format (Forkel et al. 2018) and combine the tables into single data frames including PHOIBLE (Moran and McCloy 2019) and SegBo (Grossman et al. 2020).

[9]:

```
segbo = load_cldf_dataset('../data/segbo/cldf/values.csv',
                          '../data/segbo/cldf/languages.csv')
phoible = load_cldf_dataset('../data/phoible/cldf/values.csv',
                            '../data/phoible/cldf/languages.csv')
```

Number of different inventories in PHOIBLE:

[10]:

```
n_phoible_inventories = len(phoible.Language_ID.unique())
n_phoible_inventories
```

[10]:

```
2177
```

Number of different inventories in SEGBO:

[11]:

```
len(segbo.Language_ID.unique())
```

[11]:

```
498
```

Some languages in SEGBO are missing from PHOIBLE:

[12]:

```
len(set(segbo.Language_ID) - set(phoible.Language_ID))
```

[12]:

```
199
```

We need to exclude them:

[13]:

```
phoible_langs = set(phoible.Language_ID)
segbo = segbo.loc[ segbo.Language_ID.map(lambda gltc: gltc in phoible_langs) ]
```

[14]:

```
len(segbo.Language_ID.unique())
```

[14]:

```
299
```

[15]:

```
len(set(segbo.Language_ID) - set(phoible.Language_ID))
```

[15]:

```
0
```

We compute borrowability factors for a segment s (b_s) following the approach by Eisen (2019). We assume that the marginal probability of borrowing of s ($P_s(\text{borrowing})$) is equal to probability of contact between a language with this segment and a language lacking this segment ($P_s(\text{contact})$) multiplied by the conditional probability of borrowing of this segment in a contact situation ($P_s(\text{borrowing}|\text{contact})$):

$$P_s(\text{borrowing}) = P_s(\text{contact})P_s(\text{borrowing}|\text{contact})$$

We approximate $P_s(\text{borrowing})$ with the empirical relative frequency of borrowing (q_s) provided by SEGBO and PHOIBLE and assume, following Eisen (2009), that $P_s(\text{contact})$ can be estimated as a product of the relative typological frequency of a segment (f_s) and its comlement ($1 - f_s$):

$$P_s(\text{contact}) \propto f_s(1 - f_s)$$

Here and in Model 2, we define b_s to be $P_s(\text{borrowing}|\text{contact})$, which gives

$$\begin{aligned} q_s &= f_s(1 - f_s)b_s \\ b_s &= \frac{q_s}{f_s(1 - f_s)} \end{aligned}$$

f_s is equal to the number of occurrences of s in PHOIBLE divided by the number of distinct languages in PHOIBLE.

q_s is equal to the number of occurrences of s as a borrowed segment in SEGBO again divided by the number of distinct languages in PHOIBLE: languages without borrowed segments were not included in SEGBO, which therefore cannot be used as a source of negative data.

[16]:

```python
# We cannot use vanilla relative frequencies: for SEGBO, we need the number of distinct languages from PHOIBLE;
# for PHOIBLE see below.

(
    phoible_frequencies_absolute,
    _
) = get_frequencies_w_inventory_collapsing(phoible)

(
    segbo_frequencies_absolute,
    _
) = get_frequencies_w_inventory_collapsing(segbo)
```

```
2177 languages
299 languages
```

[17]:

```python
segbo_frequencies_relative = {
    segment: count_segbo / n_phoible_inventories
    for segment, count_segbo in segbo_frequencies_absolute.items()
}
```

Vanilla relative frequencies from PHOIBLE produce valid results in most cases, but problems arise with some rare segments. E.g., when a rare segment was borrowed from language A to language B, it may happen that language A then quickly loses it. As a result, this segment may have a higher frequency in SEGBO than in PHOIBLE, which makes the derivation ill-defined.

In order to avoid this issue we create two versions of absolute PHOIBLE frequencies – one where the values are greater than or equal than in SEGBO and one where they are strictly greater (through Laplace smoothing) – and then use these absolute frequencies to compute relative typological frequencies.

```
[18]: phoible_greater_or_equal = {}
      phoible_strictly_greater = {}
      for segment, count_segbo in segbo_frequencies_absolute.items():
          if count_segbo >= phoible_frequencies_absolute[segment]:
              print(segment, count_segbo, phoible_frequencies_absolute[segment])
              phoible_greater_or_equal[segment] = count_segbo
              phoible_strictly_greater[segment] = count_segbo + 1
          else:
              phoible_greater_or_equal[segment] = phoible_frequencies_absolute[
                  segment]
              phoible_strictly_greater[segment] = phoible_frequencies_absolute[
                  segment] + 1
```

```
  1 0
 ai 1 0
   1 1
ts  1 1
p  1 1
d  1 1
  1 0
uə̯ 1 1
 ə̯ 1 0
n  1 1
l  1 1
ndz 1 0
ð  1 1
```

```
[19]: phoible_freqs_relative = {
          segment: count / n_phoible_inventories
          for segment, count in phoible_greater_or_equal.items()
      }
      phoible_freqs_relative_laplace = {
          segment: count / n_phoible_inventories
          for segment, count in phoible_strictly_greater.items()
      }
      for segment, f_s in sorted(phoible_freqs_relative.items(),
                                 key=lambda el: el[1], reverse=True)[:10]:
          print(f'{segment}: {f_s}, {phoible_freqs_relative_laplace[segment]}')

      # Smoothing has no effect on frequent segments.
```

```
m: 0.9701423977951309, 0.9706017455213597
k: 0.9205328433624254, 0.9209921910886542
j: 0.915480018373909, 0.9159393661001378
u: 0.9150206706476803, 0.915480018373909
a: 0.9108865411116215, 0.9113458888378503
p: 0.870463941203491, 0.8709232889297198
w: 0.864951768488746, 0.8654111162149747
```

```
n: 0.8474965548920533, 0.847955902618282
t: 0.7606798346348186, 0.7611391823610473
l: 0.7266881028938906, 0.7271474506201194
```

[20]:

```
# Now we can compute borrowability scores using Eisen's formula with the␣
 ↪normalisation constant

def borrowability_score(q_s, f_s):
    return q_s / f_s / (1 - f_s) # / 6 Ingoring the normalisation constant for␣
 ↪simplicity

borrowability_scores = {}
borrowability_scores_laplace = {}
for segment in segbo_frequencies_relative:
    borrowability_scores[segment] = {
        'Segment': segment,
        'Borrowability': borrowability_score (
            segbo_frequencies_relative[segment],
            phoible_freqs_relative[segment]
        ),
        'PHOIBLE_frequency_absolute': phoible_greater_or_equal[segment],
        'PHOIBLE_frequency_relative': phoible_freqs_relative[segment],
        'SEGBO_frequency_absolute': segbo_frequencies_absolute[segment],
        'SEGBO_frequency_relative': segbo_frequencies_relative[segment]
    }
    borrowability_scores_laplace[segment] = {
        'Segment': segment,
        'Borrowability': borrowability_score (
            segbo_frequencies_relative[segment],
            phoible_freqs_relative_laplace[segment]
        ),
        'PHOIBLE_frequency_absolute': phoible_strictly_greater[segment],
        'PHOIBLE_frequency_relative': phoible_freqs_relative_laplace[segment],
        'SEGBO_frequency_absolute': segbo_frequencies_absolute[segment],
        'SEGBO_frequency_relative': segbo_frequencies_relative[segment]
    }
```

[21]:

```
borrowability_df = pd.DataFrame.from_dict(borrowability_scores).T.
 ↪sort_values(by='Borrowability', ascending=False)
```

[22]:

```
# Frequently borrowed segments
borrowability_df.loc[ borrowability_df.SEGBO_frequency_absolute >= 10 ]
```

[22]:

```
    Segment Borrowability PHOIBLE_frequency_absolute  \
f         f      0.187879                        968
p         p      0.126288                       1895
                0.097833                       1255
```

```
b         b      0.081371                        1385
z         z      0.081137                         682
                0.078383                        331
d      d       0.07082                        640
v         v      0.070115                         617
d         d      0.069825                        1097
x         x      0.065986                         411
h         h      0.064024                        1258
                 0.06386                        782
l         l      0.055507                        1582
r         r      0.053289                        1099
t      t      0.052772                        916
o         o      0.051489                        1446
                0.049857                        615
s         s      0.044023                        1531
                0.043184                        327
ts       ts      0.040479                         519
e         e      0.040159                        1482
ŋ         ŋ      0.024363                        1424
                0.019333                        846

    PHOIBLE_frequency_relative SEGBO_frequency_absolute  \
f                     0.444649                      101
p                     0.870464                       31
                     0.576481                      52
b                     0.636197                       41
z                     0.313275                       38
                     0.152044                      22
d                   0.293983                     32
v                     0.283418                       31
d                     0.503904                       38
x                     0.188792                       22
h                     0.577859                       34
                      0.35921                      32
l                     0.726688                       24
r                     0.504823                       29
t                   0.420763                     28
o                     0.664217                       25
                     0.282499                      22
s                     0.703261                       20
                     0.150207                      12
ts                    0.238401                       16
e                     0.680753                       19
ŋ                     0.654111                       12
                     0.388608                      10

    SEGBO_frequency_relative
```

```
f                    0.046394
p                     0.01424
                     0.023886
b                    0.018833
z                    0.017455
                     0.010106
d                   0.014699
v                     0.01424
d                    0.017455
x                    0.010106
h                    0.015618
                     0.014699
l                    0.011024
r                    0.013321
t                   0.012862
o                    0.011484
                     0.010106
s                    0.009187
                     0.005512
ts                    0.00735
e                    0.008728
ŋ                    0.005512
                     0.004593
```

[23]:
```
# Rare segments
borrowability_df.loc[ borrowability_df.SEGBO_frequency_absolute <= 2 ][:10]
```

[23]:
```
       Segment Borrowability PHOIBLE_frequency_absolute  \
ts       ts        1.00046                                1
l         l       1.00046                               1
p          p        1.00046                               1
uə̯         uə̯       1.00046                               1
 ai         ai        1.00046                               1
ndz  ndz       1.00046                              1
                  1.00046                               1
n         n       1.00046                               1
                   1.00046                               1
d     d       1.00046                              1

       PHOIBLE_frequency_relative SEGBO_frequency_absolute  \
ts                      0.000459                       1
l                      0.000459                       1
p                       0.000459                       1
uə̯                      0.000459                       1
 ai                      0.000459                       1
ndz                    0.000459                       1
                       0.000459                       1
```

```
n                       0.000459                        1
                        0.000459                        1
d                     0.000459                        1

       SEGBO_frequency_relative
ts                    0.000459
l                     0.000459
p                     0.000459
uȩ                    0.000459
 ai                    0.000459
ndz                 0.000459
                      0.000459
n                     0.000459
                      0.000459
d                   0.000459
```

[24]:

```
borrowability_df.to_csv('model_1_borrowability.csv', index=False)
```

[25]:

```
borrowability_laplace_df = pd.DataFrame.from_dict(borrowability_scores_laplace).
 ↪T.sort_values(
    by='Borrowability', ascending=False)
```

[26]:

```
# Frequently borrowed segments
borrowability_laplace_df.loc[ borrowability_laplace_df.SEGBO_frequency_absolute␣
 ↪>= 10 ]
```

[26]:

```
    Segment Borrowability PHOIBLE_frequency_absolute  \
f         f       0.18784                        969
p         p       0.12667                       1896
                0.097862                       1256
b         b      0.081415                       1386
z         z      0.081072                        683
                0.078189                        332
d      d      0.070755                        641
v         v      0.070046                        618
d         d      0.069826                       1098
x         x      0.065863                        412
h         h      0.064043                       1259
                0.063824                        783
l         l      0.055565                       1583
r         r       0.05329                       1100
t      t      0.052757                        917
o         o      0.051524                       1447
                0.049808                        616
s         s      0.044063                       1532
                0.043075                        328
ts       ts      0.040425                        520
```

```
e         e      0.040189                       1483
ŋ         ŋ      0.024378                       1425
                0.019325                        847

    PHOIBLE_frequency_relative SEGBO_frequency_absolute  \
f                     0.445108                      101
p                     0.870923                       31
                     0.576941                       52
b                     0.636656                       41
z                     0.313734                       38
                     0.152503                       22
d                   0.294442                       32
v                     0.283877                       31
d                     0.504364                       38
x                     0.189251                       22
h                     0.578319                       34
                     0.359669                       32
l                     0.727147                       24
r                     0.505282                       29
t                   0.421222                       28
o                     0.664676                       25
                     0.282958                       22
s                     0.703721                       20
                     0.150666                       12
ts                    0.238861                       16
e                     0.681213                       19
ŋ                     0.654571                       12
                     0.389068                       10

    SEGBO_frequency_relative
f                   0.046394
p                    0.01424
                   0.023886
b                   0.018833
z                   0.017455
                   0.010106
d                 0.014699
v                    0.01424
d                   0.017455
x                   0.010106
h                   0.015618
                   0.014699
l                   0.011024
r                   0.013321
t                 0.012862
o                   0.011484
                   0.010106
```

```
s                     0.009187
                     0.005512
ts                     0.00735
e                     0.008728
ŋ                     0.005512
                     0.004593
```

[27]:

```
# Rare segments
borrowability_laplace_df.loc[ borrowability_laplace_df.SEGBO_frequency_absolute <= 2 ][:10]
```

[27]:

```
       Segment Borrowability PHOIBLE_frequency_absolute  \
p         p        0.50046                                2
                   0.50046                                2
uə̯         uə̯        0.50046                                2
n         n       0.50046                                2
l         l       0.50046                                2
ndz  ndz       0.50046                                2
ð           ð        0.50046                                2
                  0.50046                                2
 ai         ai        0.50046                                2
                  0.50046                                2

       PHOIBLE_frequency_relative SEGBO_frequency_absolute  \
p                        0.000919                                1
                         0.000919                                1
uə̯                        0.000919                                1
n                        0.000919                                1
l                        0.000919                                1
ndz                    0.000919                                1
ð                         0.000919                                1
                        0.000919                                1
 ai                       0.000919                                1
                        0.000919                                1

       SEGBO_frequency_relative
p                      0.000459
                       0.000459
uə̯                      0.000459
n                      0.000459
l                      0.000459
ndz                  0.000459
ð                       0.000459
                      0.000459
 ai                     0.000459
                      0.000459
```

[28]:

```
borrowability_laplace_df.to_csv('model_1_borrowability_laplace.csv', index=False)
```

1.1 References

Eisen, Elad. 2019. “The Typology of Phonological Segment Borrowing.” Masters thesis, Jerusalem, Israel: Hebrew University of Jerusalem.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. “Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics.” Scientific Data 5: 180205.

Grossman, Eitan, Elad Eisen, Dmitry Nikolaev, and Steven Moran. 2020. “SegBo: A Database of Borrowed Sounds in the World’s Language.” In Proceedings of the 12th Language Resources and Evaluation Conference, 5316–22.

Moran, Steven, and Daniel McCloy, eds. 2019. PHOIBLE 2.0. Jena: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.2562766.