Glotto Dstat

Theo

May 13 2020

1/ Quartets Dstat

For each quartet, languages were ordered to have the topology (P1,P2)P3)P4) (cf fig below). Then 4 D were computed. One for each dataset, lexicon, morphosyntax and phono. And final D was computed using all available data from all three dataset. The significance of each D was determined by a simple binomial test.

2/ Results

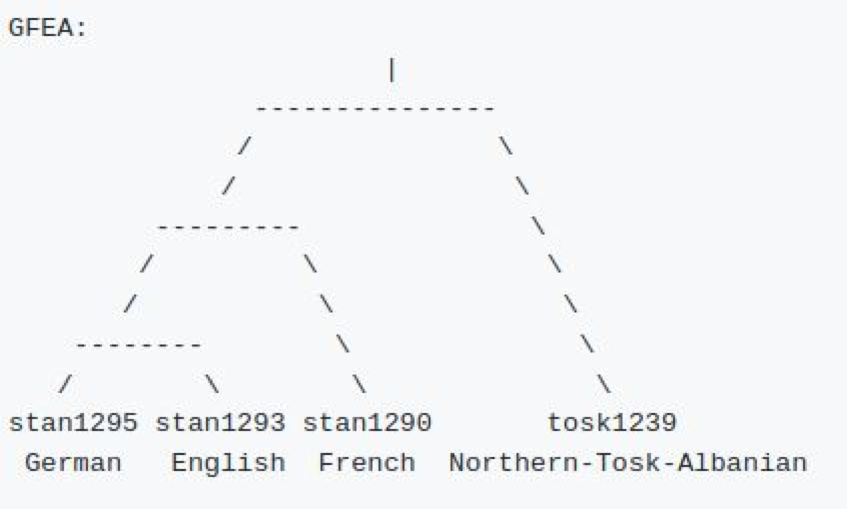
1) GFEA

For the lexicon dataset, we have a positive D=-0.54 significantly different from 0 (Pvalue=0.0169). This mean that French and English share more lexicon features that expected from the tree topologies (fig 1).

For the morphosyntax dataset, we have a negative D=-0.3 but this result is not significantly different from 0 (Pvalue=0.263).

For the phono dataset, we have a negative D=-0.75 but this result is not significantly different from 0 (Pvalue=0.07).

For all dataset, we have a null D=0. This suggests that overall there is no excess of feature shared between F and G than F and E.



```
GFEA:
                       |
               ---------------
              /               \
             /                 \
         ---------              \
        /         \              \
       /           \              \
    --------        \              \
   /        \        \              \
stan1295 stan1293 stan1290       tosk1239
 German   English  French  Northern-Tosk-Albanian
```

data	sum	abba	baba	D	Pvalue
lexi	579	17	5	0.545454545454545	0.0169005393981934
morpho	116	7	13	-0.3	0.263175964355469
phono	88	1	7	-0.75	0.0703125
all	783	25	25	0	1

Figure 1: D statistic for the quartet GFEA. A, the (P1,P2)P3)P4) topology used to compute the ABBA/BABA test. B, the summary statistics related to the 4 Dstat computed.

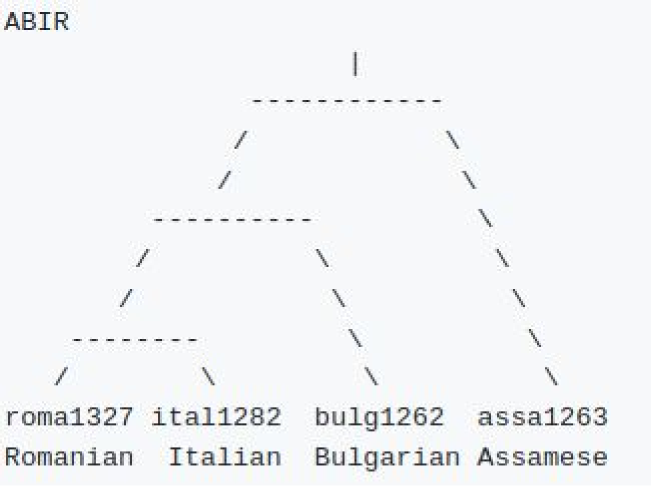
2) ABIR

For the lexicon dataset, we have a negative D=-0.384, this result is not significantly different from 0 (Pvalue=0.266) (fig 2).

For the morphosyntax dataset, we have a positive D=1, this result is not significantly different from 0 (Pvalue=0.125).

For the phono dataset, we have a positive D=0.33, this result is not significantly different from 0 (Pvalue=1).

For all dataset, we have a null D=0. Overall there is no excess of feature shared between B and R than B and I.



```
ABIR
                     |
               ------------
              /            \
             /              \
         ----------          \
        /          \          \
       /            \          \
    --------         \          \
   /        \         \          \
roma1327 ital1282  bulg1262  assa1263
Romanian  Italian  Bulgarian Assamese
```

data	sum	abba	baba	D	Pvalue
lexi	560	4	9	-0.384615384615385	0.266845703125
morpho	40	4	0	1	0.125
phono	68	2	1	0.333333333333333	1
all	668	10	10	0	1

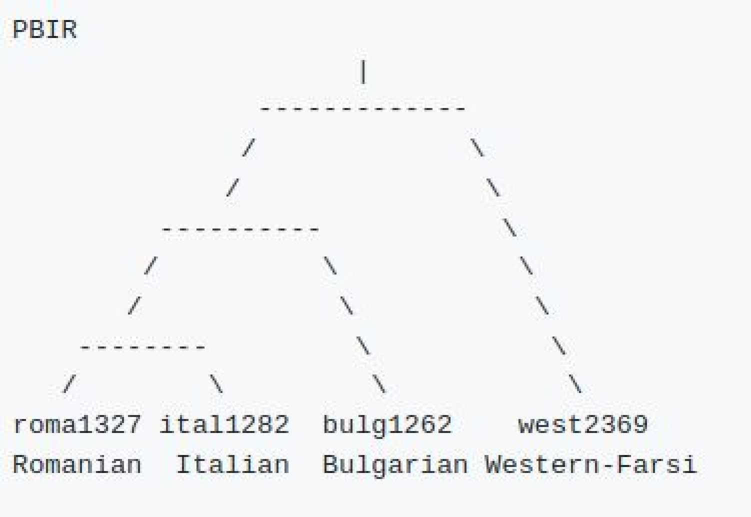
Figure 2: D statistic for the quartet ABIR. A, the (P1,P2)P3)P4) topology used to compute the ABBA/BABA test. B, the summary statistics related to the 4 Dstat computed.

3) PBIR

For the lexicon dataset, we have a positive D=0.058 significantly different from 0 (Pvalue=1) (fig 3).

For the morphosyntax dataset, we have a positive D=0.33 but this result is not significantly different from 0 (Pvalue=0.5). For the phono dataset, we have a positive D=1 but this result is not significantly different from 0 (Pvalue=1).

For the dataset, we have a positive D=0.18, this result is not significantly different from 0 (Pvalue=0.44).



```
PBIR
                     |
               -------------
              /             \
             /               \
         ----------           \
        /          \           \
       /            \           \
    --------         \           \
   /        \         \           \
roma1327 ital1282  bulg1262    west2369
Romanian  Italian  Bulgarian Western-Farsi
```

data	sum	abba	baba	D	Pvalue
lexi	576	9	8	0.0588235294117647	1
morpho	71	6	3	0.333333333333333	0.5078125
phono	68	1	0	1	1
all	715	16	11	0.185185185185185	0.442068338394165

Figure 3: D statistic for the quartet PBIR. A, the (P1,P2)P3)P4) topology used to compute the ABBA/BABA test. B, the summary statistics related to the 4 Dstat computed.

3/ Github and data

All results are available in:
https://github.com/theotricou/watch_your_mouth/tree/master/4_languages.
You can also find, for each quartet, the list of features that exhibit a ABBA or BABA signature.

4/ Remarks

It should be noted that the number of features presenting the ABBA and BABA signature is really low compare to the number of feature available. This could explain why only one test was significant.
Once again more data could help use detecting more accurately pattern of exchange between languages.