Clustering of Analogies for Inter-Language Similarities

Software project

Justine Diliberto, Cindy Pereira, Anna Nikiforovskaja

Université de Lorraine, IDMC

12.10.2021

Work done

- Find approximately 10-15 papers on language similarities, including both linguistics and deep learning approach.
- Run the program by Safa et al and check that we receive the same things from their paper
- Think of possible visualisations and implement a few of them

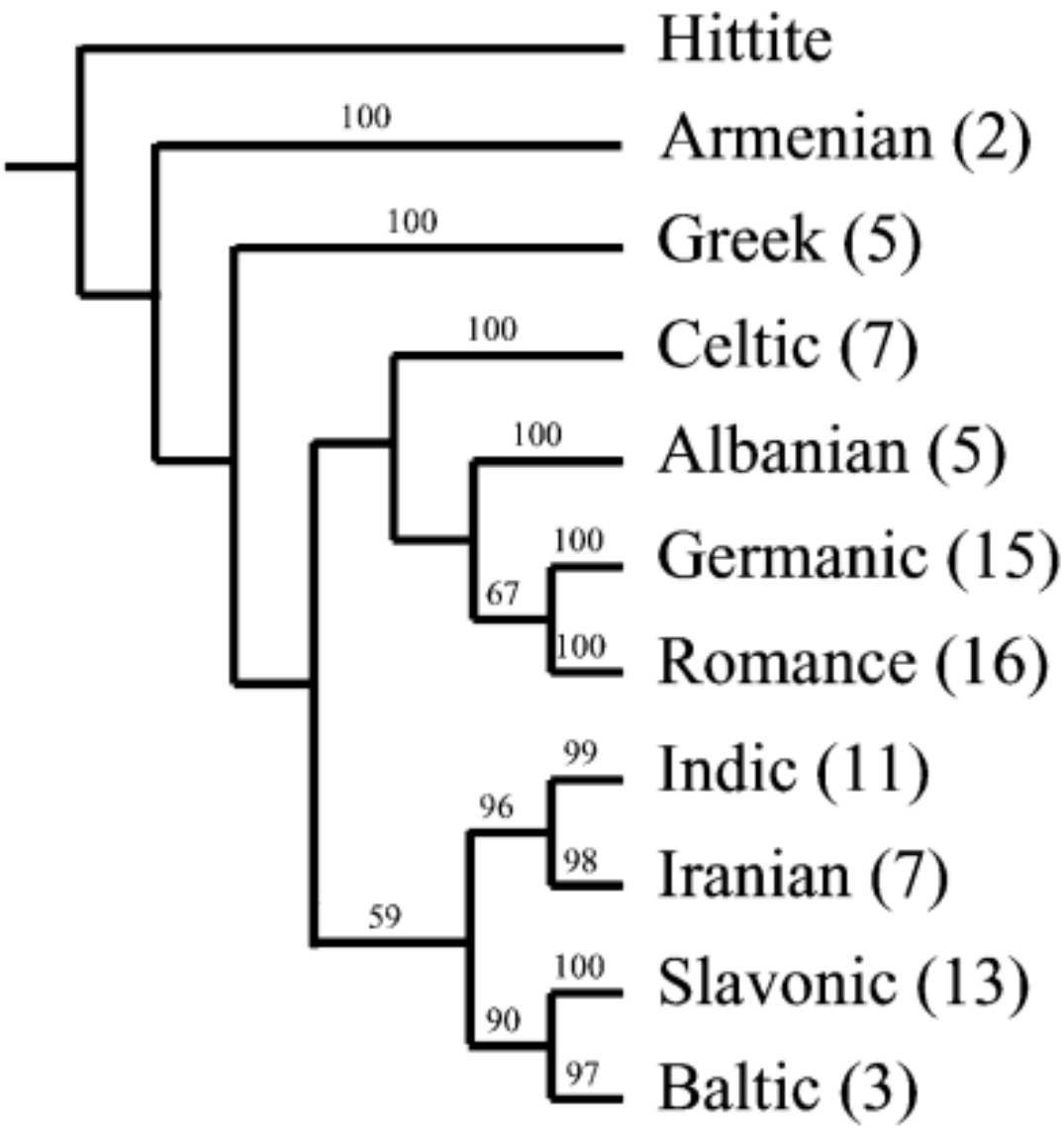
Linguistic approach

- Studied approach[1]: lexicostatistical methods
- Other approaches: phonetics, genetics, archaeology

- Cladistic analysis of languages: Indo-European classification based on lexicostatistical data
  - hierarchy of Indo-European languages
  - analysis of lexical data (basic vocabulary)
  - using the maximum parsimony approach
  - evaluate robustness through primary homologies and to different character-coding strategies

[1] Rexová et al., *Cladistic analysis of languages: Indo-European classification based on lexicostatistical data.*

Linguistic approach



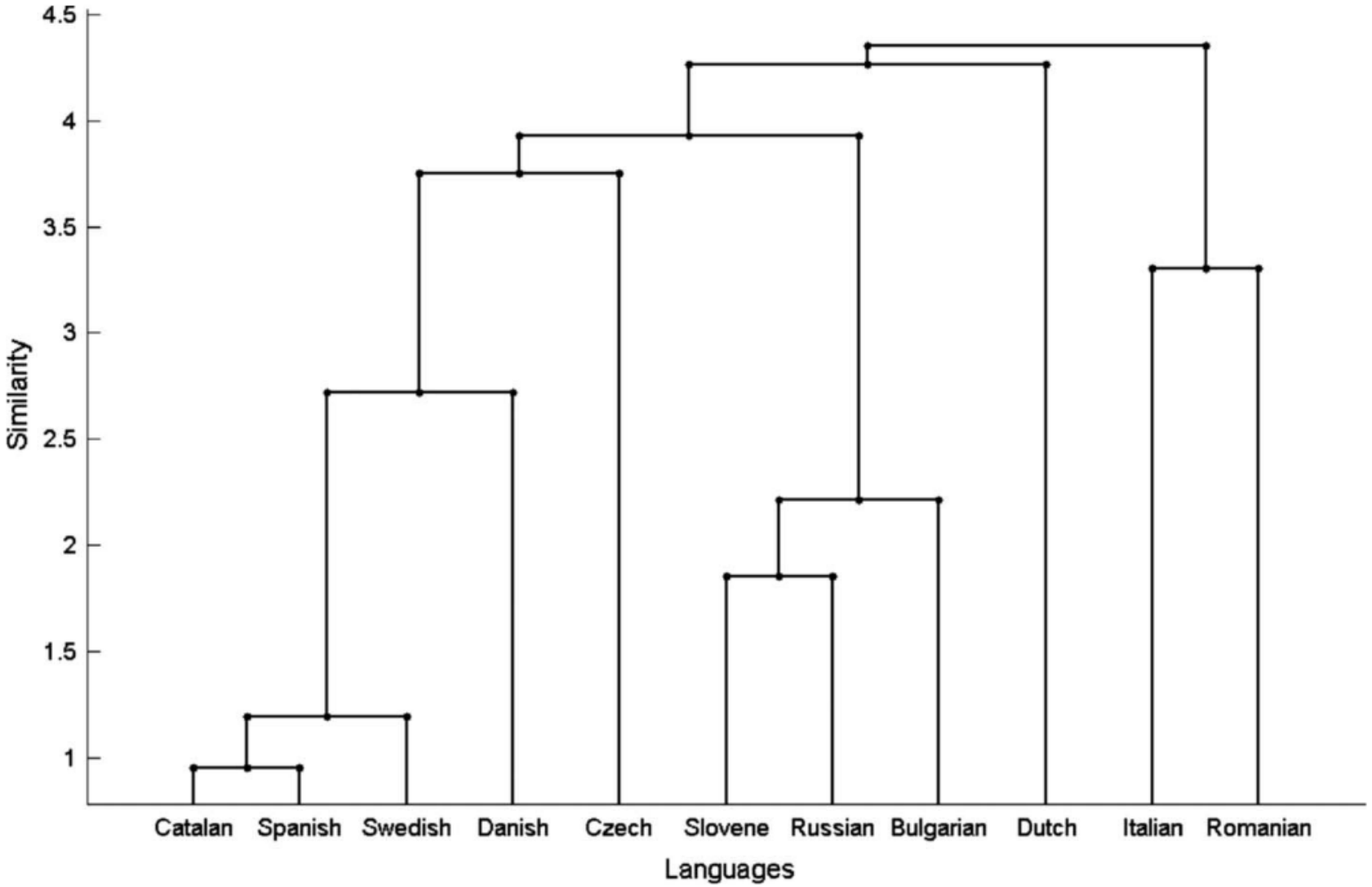
Computational approach

Studied approach[2]:

- 11 languages
- Classification algorithms using:
  - Dependencies
  - n-grams
  - Quantitative typological indices
- Results match with genealogical similarities of languages

[2] Abramov et al., *Automatic Language Classification by means of Syntactic Dependency Networks.*

Computational approach



Problems

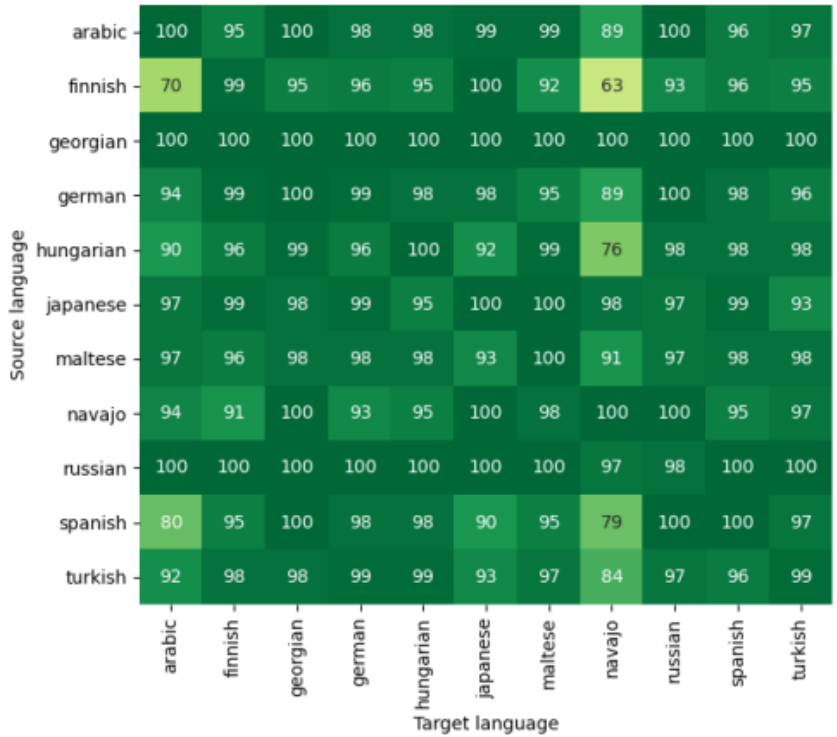
Computers die

- 5h to run half of the evaluation of the classifiers
- Process killed by lack of memory
- Need CSV file to implement visualisation

Compared results on full transfer

100	95	100	98	98	99	99	89	100	96	97
70	99	95	96	95	100	92	63	93	97	95
100	100	100	100	100	100	100	100	100	100	100
94	99	100	100	98	98	95	89	100	98	96
90	96	99	96	100	92	99	76	98	98	98
97	99	98	99	95	100	100	98	97	99	93

Figure: Positive analogies

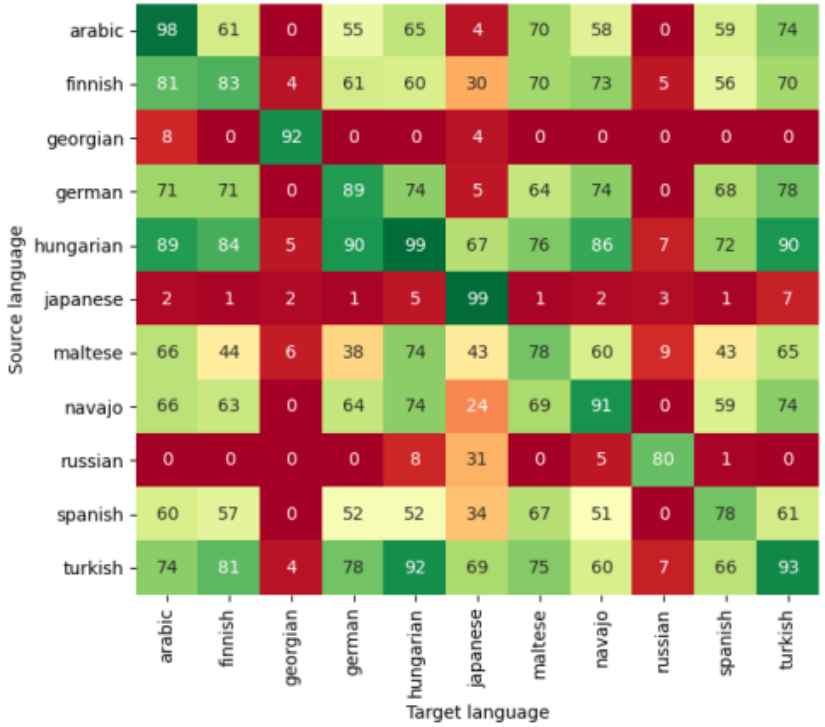


(b) Positive

Compared results on full transfer

98	61	0	55	65	4	70	58	0	59	74
81	83	4	61	60	30	70	73	5	55	70
8	0	92	0	0	4	0	0	0	0	0
71	71	0	89	74	5	64	74	0	67	78
89	84	5	90	99	67	76	86	7	73	90
2	1	2	1	5	99	1	2	3	1	7

Figure: Negative analogies

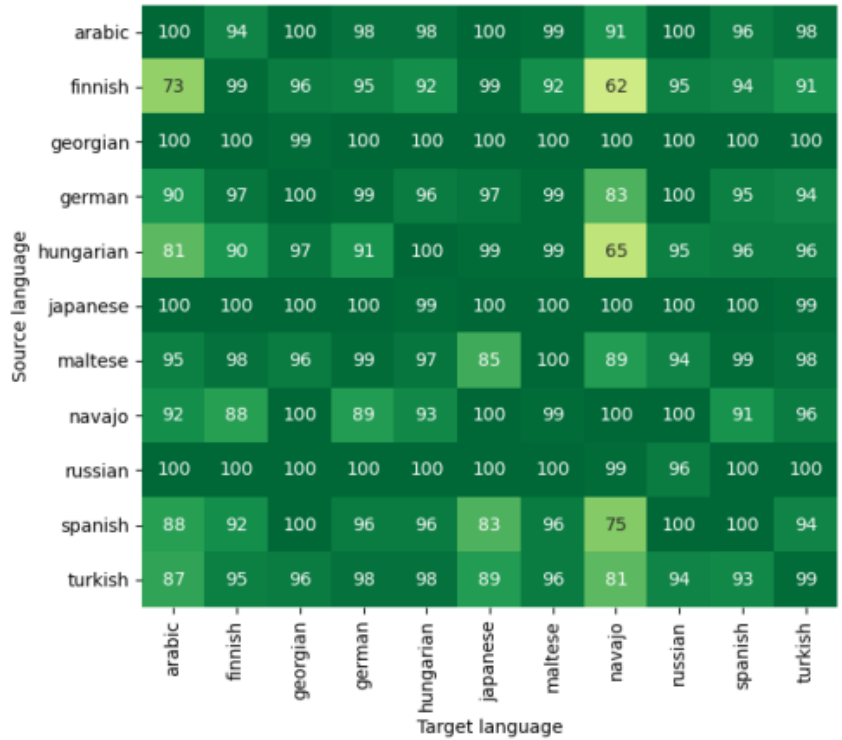


(c) Negative

Compared results on full transfer

100	94	100	98	98	100	99	91	100	96	98
74	99	96	95	92	99	92	62	95	94	91
100	100	99	100	100	100	100	100	100	100	100
90	97	100	99	96	97	99	83	100	95	94
81	90	97	91	100	99	99	65	95	96	96
100	100	100	100	99	100	100	100	100	100	99

Figure: Raw analogies



(a) Base

Methods for visualisation

- Scikit network library
  - Louvain clustering algorithm
  - Node layout using Spring
  - Ward for hierarchy dendrogram
- Transfer only on negative analogies
  - More representative - big differences
- Full and partial transfer

Visualisation

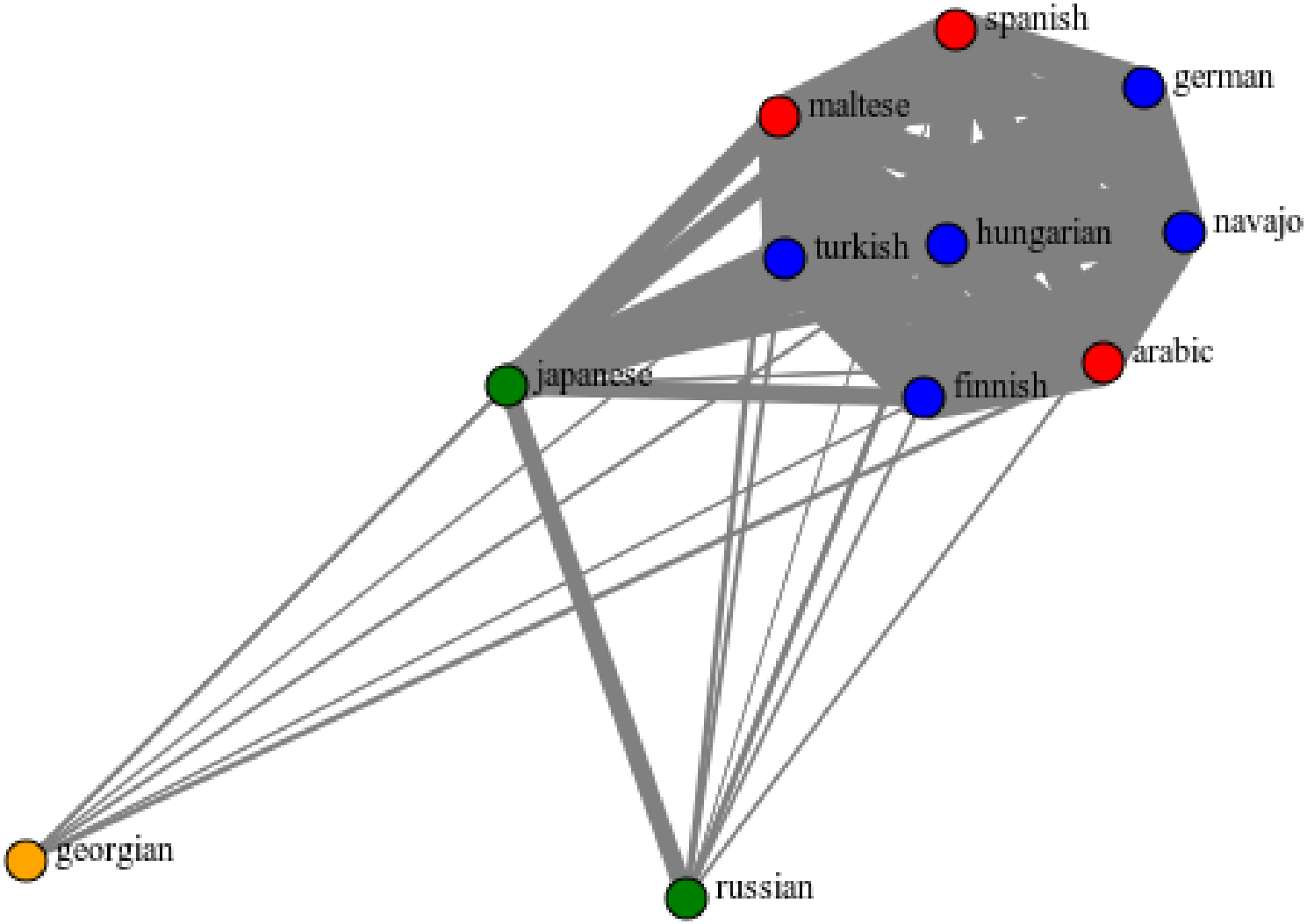


Figure: Full transfer on 1000 analogies

Visualisation

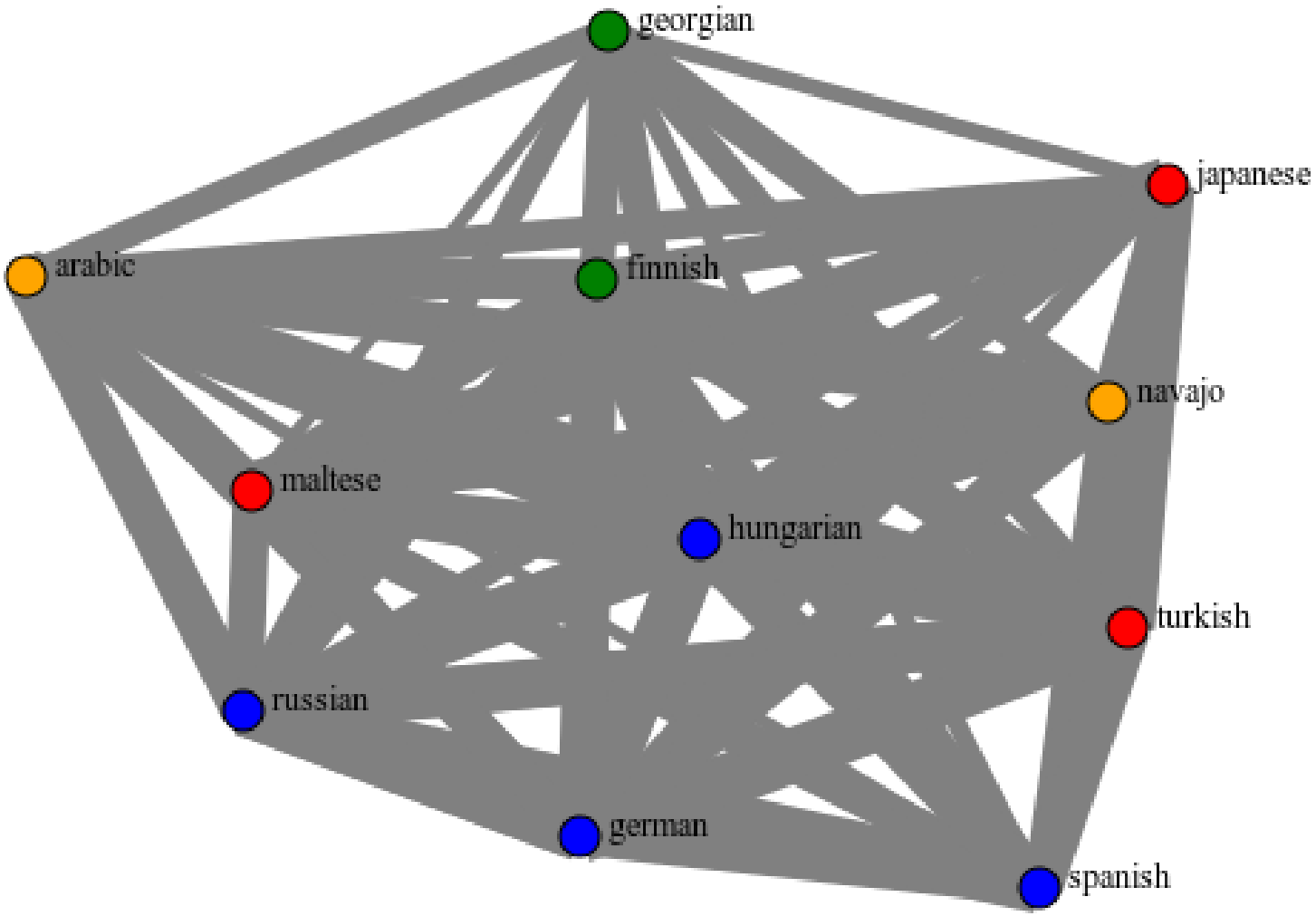
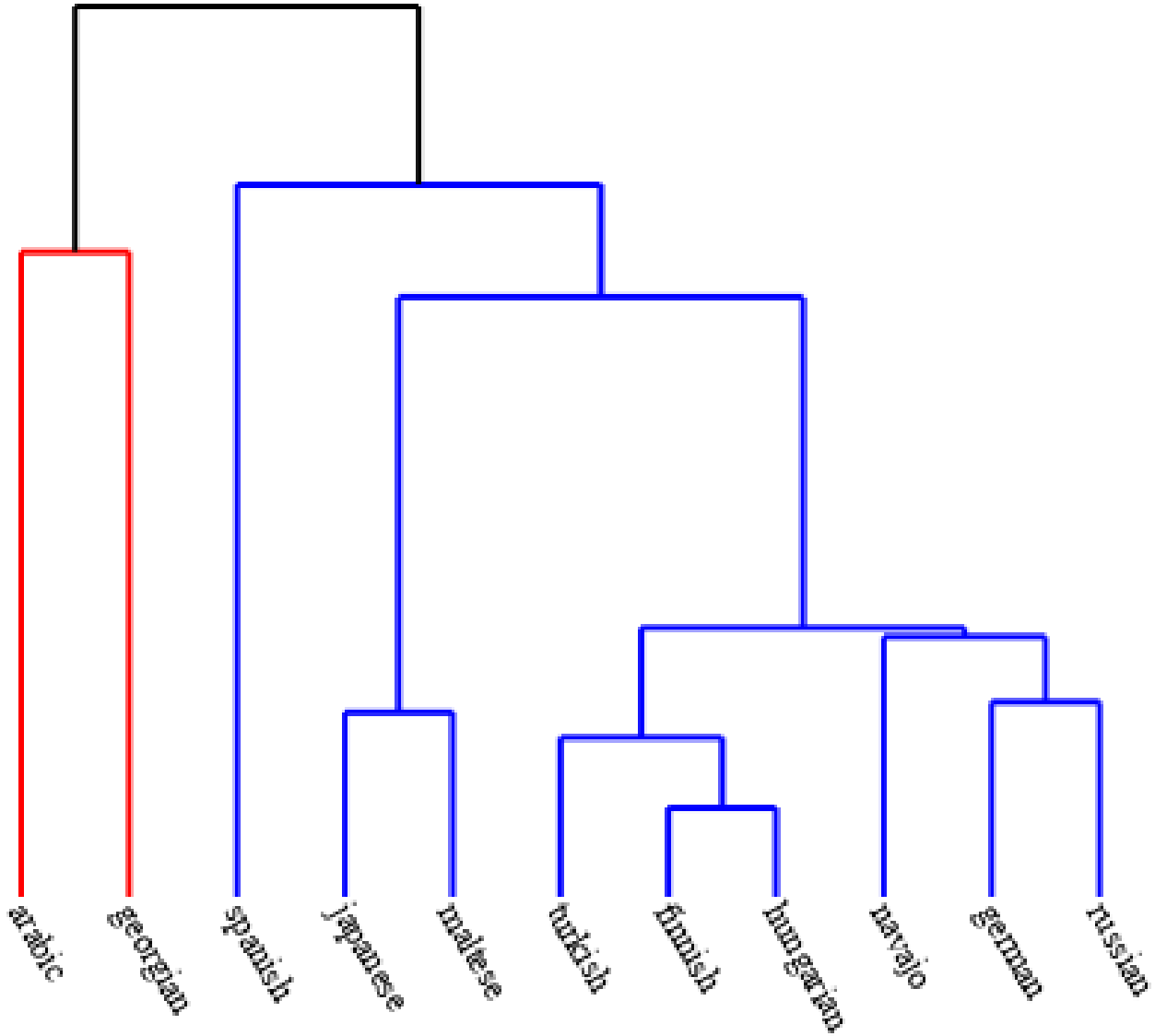


Figure: Partial transfer on 1000 analogies

Visualisation



Plan

- Run and adapt the program on SIGMORPHON 2020
- Compare the results to linguistic researches
- Try other deep learning models and compare
- Start building the library

Thanks

Thank you for your attention.