Complex Networks Reveal a Glottochronological Classification of Natural Languages

Harith Hamoodat, Younis Al Rozz and Ronaldo Menezes

Abstract The success of humans cannot be attributed to language, but it is certainly true that language and modern humans are inseparable. This work focuses on revealing the structure of 20 Indo-European languages belonging to three sub-families (Romance, Germanic, and Slavic) from a chronological perspective. In order to find the chronological characteristic features of these languages, we use (1) Heaps' law, which describes the growth of vocabulary (distinct words) in a corpora for each language to the total number of words in the same corpora and (2) structural properties of networks created from word co-occurrence in corpora of 20 written languages. Using clustering approaches and entanglement, we show that in spite of differences from years of being used separately and differences in alphabets, one can find language characteristics that lead to cluster of languages resembling the organization according to historical sub-families and chronological relations.

1 Introduction

The development of societies leads to the use of different tones and words creating different dialects for the same language. Over time, those dialects change by adding or removing words until they are considered as a new language. Moreover, the migration of human populations groups contributed to the formation of languages because the geographical separation of populations acts as a catalyst for changes in vocabulary. In fact, this analogy is similar to how different species emerged as a result of geographical separation. This evolution of language formation means that today there are thousand of different languages currently being used [17]. Due to

H. Hamoodat (✉) · Y. Al Rozz · R. Menezes
BioComplex Laboratory, Computer Science, Florida Institute of Technology,
Melbourne, FL, USA
e-mail: hhamdon2013@my.fit.edu

Y. Al Rozz
e-mail: yyounis2013@my.fit.edu

R. Menezes
e-mail: rmenezes@cs.fit.edu

© Springer International Publishing AG 2018

S. Cornelius et al. (eds.), *Complex Networks IX*, Springer Proceedings
in Complexity, https://doi.org/10.1007/978-3-319-73198-8_18

the nature of their formation, many of these languages can be grouped together into a language family. The languages in each family are related through descent to a common ancestral language. Parental languages transfer some of its characteristics to derived languages; thus, we can say that the derived languages within a language family are "genetically" related [23].

There are about 100 language families in the world, e.g., Indo-European, Afro-Asiatic, and Niger-Congo. The Indo-European family has the largest number of speakers among all families known (more than 40% of the human population). It contains about 445 languages many of them widely spoken such as Spanish, English, Russian, and Portuguese [10]. According to Linguists, the Indo-European family can be divided into several sub-families such as Germanic, Italic-Romance, Slavic, and Baltic [7].

The availability of large volumes of data today encourages researchers to study the relation between languages using regularities extracted from corpora of text. In this work, we show that even without lexical distance analysis or word-pair relations, and focusing merely on the structure built from syntax, we can detect useful structure of language families.

2 Related Work

Although a number of studies have been done in the history of languages and how they derived from each other, there is no unanimity on the origin of human languages because of the lack of direct evidence and empirical data [4]. Due to the difficulty to determine the specific date of language separation, the researchers try to study the relationship between languages and convert the result into an estimate for when a pair diverged. However, the calculation of the distance between pair of language is one of the most efficient methods to use it for chronological estimation. Linguistic distance—how different one language or dialect is from another [22]—can be computed by the lexical distance of the language vocabulary [12, 21].

There are several distance measure algorithms that can be applied on text like Hamming distance, Levenshtein distance, and Jaro–Winkler distance [25]. Levenshtein is commonly used, and it is a metric for measuring the difference between two string sequences. The Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one word into the other.

Petroni and Serva [21] created a chronological family tree for Indo-European and the Austronesian group. They used fifty different languages for both cases depending on two Swadesh list dataset, one for Indo-European and one for Austronesian. The authors created matrices of the lexical distances between languages for the two families. Each matrix contains 1225 elements to describe all pairs in a group. Then, they calculated the absolute timescale for those pairs. In order to calculate the distance between each language pair, one takes the average of the distances between the word pairs. They used a modification of the Levenshtein distance and normalized it by the number of characters for longer of the two words, which is reasonable if two words differ by one character this is much more important for short words than it is for long words. They found that the result from the method above is relatively similar to those found by glottochronologists.

The use of a cognate set of words to study the time of language divergence is not new. In fact, Gray et al. [11] studied the time separation between 87 Indo-European languages from a dataset of 2,449 cognate sets coded as discrete binary characters. They applied likelihood models of lexical evolution to solve the problem of accuracy of tree topology and branch length estimation. A Bayesian inference of phylogeny was used to enhance the estimation of tree topology and branch lengths. Also, they used rate-smoothing algorithms to reduce the rate variation across the tree. Last, they tried to examine subsets of languages using split decomposition, and the result showed a strong identity for the tree when comparing a subset result with complete one. They found the results are in agreement with the Anatolian theory for the origin time of Indo-European languages. Furthermore, a number of studies have been done for the classification of languages using text characteristics without looking to the time divergence [2, 15].

3 Methodology

3.1 Data Curation and Model

In this work, we utilize a large amount of textual data called the Leipzig corpora collection [9]. The languages chosen for this work were Romanian (Ron), French (Fra), Catalan (Cat), Italian (Ita), Spanish (Spa), Portuguese (Por), German (Ger), Dutch (Dut), Danish (Dan), Norwegian (Nor), Swedish (Swe), English (Eng), Slovenian (Slv), Bulgarian (Bul), Polish (Pol), Russian (Rus), Ukrainian (Ukr), Croatian (Cro), Czech (Ces), and Slovak (Slk). These languages were chosen because they are good representations of three large sub-families of the Indo-European family, which are Italic, Germanic, and Slavic. The text corpus for each language was constructed from Wikipedia and news pages to ensure vocabulary diversity. We made the size of the corpus for each language consistent; each language corpus is composed of 1 million sentences. After the entire text was converted to lower case, and the punctuation and special characters were removed, we used 100,000 words from each corpus for the work developed in this paper. The second kind of data we used relates to the languages, tree topology, branches length, and divergence period between languages (year the languages separate), which we reconstructed from several works [11, 12, 21] in linguistics. This hierarchy was done for the 20 languages we deal with in this paper and is used as the ground truth (see Fig. 1).

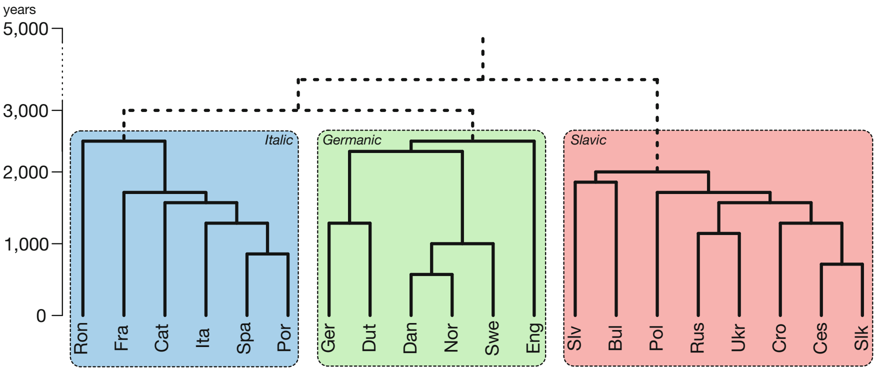


Fig. 1 A dated phylogenetic tree of 20 Indo-European languages with three sub-families, Italic, Germanic, and Slavic. The dates on the y-axis are approximations for when these languages split from a common language

3.2 Feature Extraction

We extracted a set of 19 features for each language; we want to demonstrate that one could use these features (or some of them) to unveil a structure similar to the ground truth. The first two features represent the vocabulary richness of the language as expressed by Heaps' law [13]. The parameters k and β describe the vocabulary growth (distinct words) in texts as a function the total number of words seen [2, 16]. More formally, $V_R(n) = kn^{\beta}$ where V_R is the number of vocabulary words in the text of size n, k and β are parameters determined experimentally from the fitting of Heaps' law.

The other 17 features were obtained from the word co-occurrence network for each language. The network is simple and built having words as nodes and linking words if they appear in the corpus consecutively. The edges' weights represent the frequency in which the two words appear next to each other. The networks follow a power-law distribution and have community structures (we used Louvain modularity [3]); the number of communities is an important feature (com). The features α_d and α_s represent, respectively, the scaling of the degree distribution and the distribution of community sizes. The size of the network is given by the number of nodes n and number of edges m.

There are many other structural characteristics that can be computed from the networks. For this work, we exhaustively added many features without too much concern for an exact number. The purpose is to make sure we are capturing as many uncorrelated metrics as possible. Later we worked reducing the dimensions and identifying the most significant parameters. The degree k of a node is the number of edges connected to it. The higher average degree $\langle k \rangle$ the network has, the more density it is [6]. From Table 1, we can clearly see that the Slavic languages have a

Table 1 Each line in this table represents 19-dimension feature vector for the language shown in the first column

Languages	$\Bbbk$	β	α_d	α_s	n	m	$\langle k \rangle$	C_4	C	$\langle C_d \rangle$	$\langle C_b \rangle$	$\langle C_c \rangle$	D	$trans$	η_∇	ℓ	r	Q	com
Portuguese	6.40	0.702	2.302	1.343	20,641	70,816	6.86	0.044	0.186	0.00033	0.00010	0.305	11	0.0103	11.729	3.331	−0.135	0.392	47
Spanish	7.63	0.694	2.323	1.462	22,258	73,026	6.56	0.059	0.241	0.00030	0.00010	0.315	14	0.0088	12.972	3.217	−0.227	0.351	111
Italian	8.28	0.689	2.291	1.399	22,885	77,693	6.79	0.035	0.170	0.00030	0.00010	0.302	13	0.0113	11.721	3.357	−0.223	0.363	55
Catalan	7.69	0.686	2.324	1.335	20,856	68,005	6.52	0.073	0.277	0.00030	0.00010	0.322	10	0.0084	13.551	3.151	−0.210	0.364	44
French	7.41	0.690	2.289	1.324	20,700	73,241	7.08	0.051	0.257	0.00030	0.00010	0.322	09	0.0109	16.628	3.146	−0.245	0.336	58
Romanian	8.91	0.683	2.307	1.252	22,821	75,361	6.60	0.043	0.175	0.00028	0.00010	0.305	10	0.0106	11.306	3.325	−0.185	0.371	33
Dutch	6.54	0.700	2.175	3.529	20,485	72,745	7.10	0.081	0.320	0.00030	0.00010	0.326	11	0.0157	26.030	3.102	−0.219	0.304	31
German	0.23	1.008	2.214	1.427	24,296	73,841	6.08	0.088	0.260	0.00020	0.00009	0.317	10	0.0120	16.121	3.200	−0.195	0.352	112
Danish	5.70	0.720	2.217	4.804	22,234	71,612	6.44	0.066	0.246	0.00020	0.00010	0.311	10	0.0130	16.535	3.259	−0.183	0.358	34
Norwegian	6.13	0.706	2.231	4.456	20,571	63,997	6.22	0.090	0.298	0.00030	0.00010	0.322	10	0.0108	15.349	3.143	−0.210	0.364	30
Swedish	4.65	0.743	2.186	1.330	24,071	70,887	5.89	0.081	0.278	0.00020	0.00010	0.316	11	0.0086	11.808	3.209	−0.199	0.386	44
English	9.88	0.650	2.368	1.404	17,448	68,762	7.88	0.074	0.318	0.00040	0.00010	0.339	09	0.0107	22.913	2.994	−0.193	0.291	47
Bulgarian	5.41	0.728	2.449	1.854	23,655	58,746	4.97	0.061	0.185	0.00020	0.00009	0.306	17	0.0034	5.091	3.323	−0.189	0.503	496
Slovenian	7.58	0.716	2.343	1.791	28,669	83,470	5.82	0.037	0.122	0.00020	0.00008	0.286	11	0.0105	8.593	3.558	−0.117	0.396	62
Russian	7.51	0.719	2.334	4.502	29,333	81,405	5.55	0.045	0.123	0.00010	0.00008	0.285	10	0.0057	5.415	3.576	−0.112	0.428	57
Ukrainian	4.41	0.765	2.345	2.629	29,363	78,155	5.32	0.054	0.147	0.00018	0.00008	0.289	15	0.0066	5.654	3.543	−0.159	0.438	36
Czech	4.71	0.765	2.387	1.878	31,486	83,320	5.29	0.041	0.101	0.00016	0.00008	0.274	12	0.0057	4.298	3.726	−0.086	0.438	64
Slovak	7.07	0.733	2.288	2.305	32,542	87,625	5.39	0.029	0.086	0.00016	0.00008	0.270	13	0.0075	4.896	3.775	−0.081	0.431	65
Croatian	7.31	0.716	2.317	2.003	27,369	63,826	4.66	0.039	0.144	0.00017	0.00010	0.267	14	0.0040	2.693	3.819	−0.134	0.550	132
Polish	5.92	0.734	2.390	3.155	27,390	72,721	5.31	0.048	0.122	0.00019	0.00009	0.277	16	0.0082	5.123	3.678	−0.130	0.470	70

lower $\langle k \rangle$ compared to all other languages in the dataset, while the English language has the higher one.

In addition to the network clustering coefficient (C), a measure of the degree to which nodes in a graph tend to cluster together, we calculate the square clustering (C_4) which is the quotient between the number of squares and the total number of possible squares [14].

Similar to the concept of clustering (C) is the concept of transitivity ($trans$) [24] of the network. Moreover, both C and $trans$ depend on the number of triangles (cliques of three nodes) in the network, so we have also included these features ($trans$ and η_∇, respectively) as part of our set of metrics. Another important feature of networks is the average path length (ℓ) between two nodes which is also included in our list. Croatian has the longest value for $\ell = 3.81$ steps, while the shortest one was English with $\ell = 2.99$. This is likely because morphological languages like most of Slavic languages tend to have long sentences than analytic languages like English and Dutch [1]. The diameter of the network D is the largest shortest path and another important feature we included here. Note that at this point, the idea is to have an exhaustive list of features that could represent a language.

Related to community detection algorithms is the modularity of the network given by Q which is designed to measure the strength of a division of a network into groups; a measure the community structure [8]. The value of Q for all 20 networks was calculated using the approach proposed by Newman [19]. Based on this metric, Croatian has the largest modularity value of 0.550, while the lowest value was 0.291 scored by English.

Centrality measures are used to identify the important nodes within a network; here we used degree centrality (C_d) which is highly correlated to $\langle k \rangle$, betweenness (C_b), and closeness (C_c) as defined by Borgatti [5]. However since we want a network feature, we represent the average of all these values given by $\langle C_d \rangle$, $\langle C_b \rangle$, and $\langle C_c \rangle$. Last, we compute the degree assortativity of the network which is given by r [18].

After all the analysis, we had a 19-dimension feature vector for each language as depicted in Table 1. This vector is used in clustering the networks, but we will also try to identify the significant features and reduce the dimension.

4 Results and Discussion

In order to compare the tree resulted from the hierarchical clustering with the ground truth tree (Fig. 1), we measured the quality of the alignment of the two trees by calculating the entanglement function. Entanglement is a measure between 1 (full entanglement) and 0 (no entanglement) which corresponds to a good alignment. We took all the possible combination of the 19 parameters in the matrix, for each combination, we constructed a tree and compared it with the ground truth in order to find the entanglement. Table 2 contains the best 10 entanglement from all combinations. Furthermore, we can evoke which features have high impact on the results like

Table 2 Best 10 Entanglement with its combinations

Entanglement	κ	β	α_d	α_s	n	m	$\langle k \rangle$	C_4	C	$\langle C_d \rangle$	$\langle C_b \rangle$	$\langle C_c \rangle$	D	$trans$	$\eta_\triangledown$	ℓ	r	Q	com
0.0602616					✓		✓	✓				✓		✓		✓	✓		
0.0602616					✓		✓	✓				✓		✓		✓	✓		✓
0.0604673			✓			✓	✓							✓	✓	✓	✓	✓	✓
0.0604673			✓		✓	✓	✓					✓		✓	✓		✓	✓	✓
0.0604673			✓		✓	✓	✓							✓	✓	✓	✓	✓	✓
0.0604673			✓			✓	✓							✓		✓	✓	✓	✓
0.0604673			✓			✓								✓	✓	✓	✓	✓	✓
0.0653795			✓		✓				✓	✓		✓		✓		✓	✓		✓
0.0663400			✓				✓		✓		✓	✓	✓		✓		✓		
0.0687276					✓		✓	✓						✓		✓	✓		✓

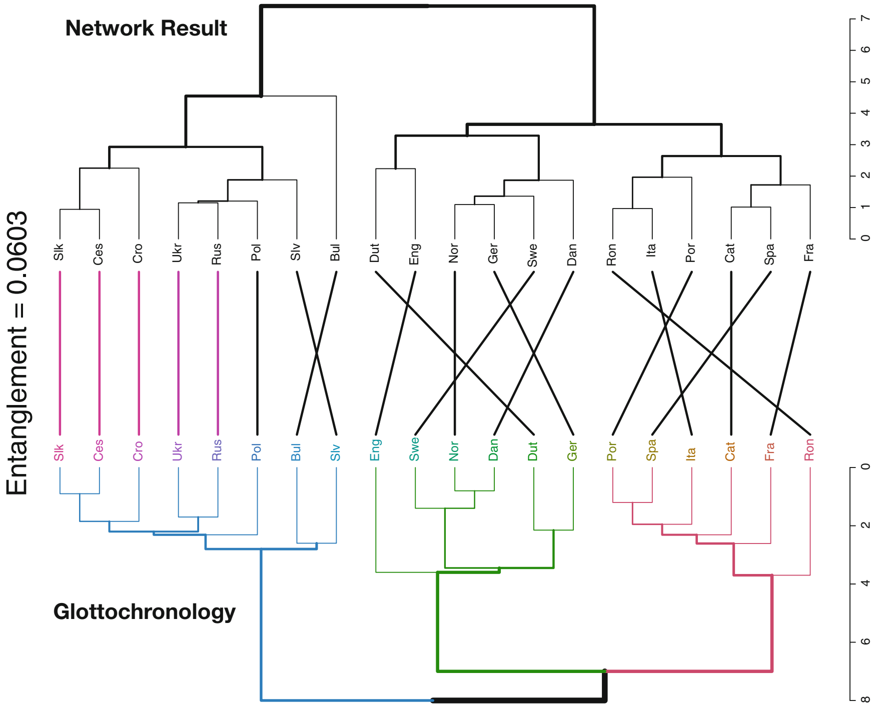
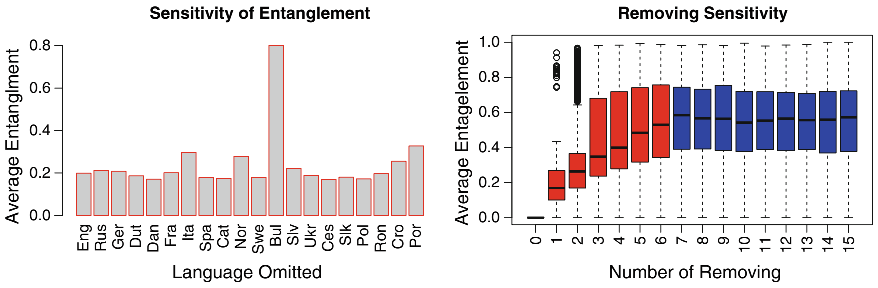


Fig. 2 Entanglement between two trees using the best entanglement case

$trans$, r, ℓ, and $\langle k \rangle$ which they appeared in the most cases. In contrast, there are some parameters useless for this work like Heaps' law parameters and $\langle C_b \rangle$ (Table 2).

The best combination between all the cases has the entanglement value of 0.06 (first case in Table 2), this case has only seven parameters, which are the smallest combination parameters that give better values (Fig. 2). The hierarchical clustering was not only able to distinguish the Slavic languages from the non-Slavic language but also to capture the branches relation and distances for this sub-family with one exception which is the Bulgarian language (discussed later). Moreover, it was ambidextrous to recognize the Germanic from Romance languages with some differences in the branches relation like Germany with Norwegian instead of the Dutch language.

In order to check the consistency of result, we tested the sensitivity of removing languages. First, we remove one language each time and calculate the average entanglement for all cases. Secondly, we remove two languages and calculate the average entanglement, and so on (Fig. 3b). The average entanglement increased until the sixth language removed and then started to be constant at a high level, which means that the topology of the tree is completely destroyed and the removal of more languages does not affect the result.



(a) The effect of language removal to the average entanglement.

(b) Average entanglement sensitivity as a function of languages removal.

Fig. 3 Entanglement sensitivity as a function of removing languages

To test for certain language impact on the average entanglement and tree topology, we removed one language each time and recalculated the average entanglement. The language with high average entanglement in Fig. 3a means the most effective language on the tree topology. In our languages set, when we removed the Bulgarian language which occupied a whole branch in the network result cluster, the average entanglement became very high (0.79) which means the branches relation is very tangled. The unpredictable behavior of the Bulgarian language may be due to several reasons; first, the number of unique words (nodes) is less than others Slavic languages. Also, words in the Bulgarian language are most likely to connect with another word several times which describes the reason why the language has a number of connections less than all other language networks in the dataset. On the other hand, several important dissimilarities exist between the Bulgarian language and other Slavic languages. For instance, Bulgarian is an analytic language and its unique morphological features tend toward the Balkan family of languages. The Bulgarian language roots back to the Proto-Slavic branch of the Indo-European language family which have common features with the Indo-Iranian languages, more specifically, the Germanic family, but it was much similar to the Baltic family of languages. Finally, a lot of the words in the Bulgarian language were borrowed from the Turkish and Greek languages [20].

5 Conclusion

In this study, we used the topological measurements extracted from word co-occurrence networks of 20 Indo-European languages along Heaps' law parameters to construct the hierarchical cluster that represents the chronological distance between those languages. The comparison that we made of our results with the glottochronological classification based on the lexical distance between word fluctuation among different languages shows a strong agreement between the two methods. In order to support this finding, we test the tolerance of the cluster against languages variation. We did this by removing one language a time and calculate the entanglement. Also, we extracted the best features that give the lowest entanglement; these features we believe they best describe the chronological difference between languages. The results we get from this work open the door for many future works; for instance, we could expand our study to include languages from different main families. Also, it is possible to apply our method to find the closest translation of document to the original text in order to assets the quality of translation.

References

1. Abramov, O., Mehler, A.: Automatic language classification by means of syntactic dependency networks. J. Quant. Linguist. **18**(4), 291–336 (2011)
2. Al Rozz, Y., Hamoodat, H., Menezes, R.: Characterization of written languages using structural features from common corpora. In: Workshop on Complex Networks CompleNet, pp. 161–173. Springer, Berlin (2017)
3. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. **2008**(10), P10008 (2008)
4. Bolhuis, J.J., Tattersall, I., Chomsky, N., Berwick, R.C.: How could language have evolved? PLoS Biol. **12**(8), e1001934 (2014)
5. Borgatti, S.P.: Centrality and network flow. Soc. Netw. **27**(1), 55–71 (2005)
6. Bosu, A., Carver, J.C.: How do social interaction networks influence peer impressions formation? a case study. In: IFIP International Conference on Open Source Systems, pp. 31–40. Springer, Berlin (2014)
7. Campbell, L.: American Indian Languages: The Historical Linguistics of Native America. Oxford University Press, Oxford (2000)
8. de Arruda, H.F.: Costa, L.da F., Amancio, D.R.: Topic segmentation via community detection in complex networks. Chaos Interdiscip. J. Nonlinear Sci. **26**(6), 063120 (2016)
9. Goldhahn, D., Eckart, T., Quasthoff, U.: Building large monolingual dictionaries at the leipzig corpora collection: from 100 to 200 languages. In: LREC, pp. 759–765 (2012)
10. Gordon, R.G., Grimes, B.F., et al.: Ethnologue: Languages of the World, vol. 15. SIL International, Dallas (2005)
11. Gray, R.D., Atkinson, Q.D.: LangUage-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin, vol. 426. Nature Publishing Group, London (2003)
12. Gray, R.D., Atkinson, Q.D., Greenhill, S.J.: Language evolution and human history: what a difference a date makes. Philos. Trans. R. Soc. Lond. B Biol. Sci. **366**(1567), 1090–1100 (2011)
13. Herdan, G.: Type-Token Mathematics, vol. 4. Mouton, Berlin (1960)
14. Lind, P.G., Gonzalez, M.C., Herrmann, H.J.: Cycles and clustering in bipartite networks. Phys. Rev. E **72**(5), 056127 (2005)
15. Liu, H., Xu, C.: Can syntactic networks indicate morphological complexity of a language? EPL (Europhys. Lett.) **93**(2), 28005 (2011)
16. Lü, L., Zhang, Z.-K., Zhou, T.: Zipf's law leads to heaps' law: analyzing their relation in finite-size systems. PloS One **5**(12), e14139 (2010)
17. McWhorter, J.H.: The Story of Human Language. Teaching Company (2004)
18. Newman, M.E.J.: Assortative mixing in networks. Phys. Rev. Lett. **89**(20), 208701 (2002)
19. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**(23), 8577–8582 (2006)
20. Osenova, P.: Bulgarian. Revue belge de philologie et d'histoire **88**(3), 643–668 (2010)
21. Petroni, F., Serva, M.: Language distance and tree reconstruction. J. Stat. Mech. Theory Exp. **2008**(08), P08012 (2008)

22. Renfrew, C., McMahon, A., Trask, R.L.: Time depth in historical linguistics. The Macdonald Institute for Archaelogical Research (2000)
23. Rowe, B.M., Levine, D.P.: A Concise Introduction to Linguistics. Routledge, Abingdon-on-Thames (2015)
24. Schank, T., Wagner, D.: Approximating clustering-coefficient and transitivity. Universität Karlsruhe, Fakultät für Informatik (2004)
25. Van der Loo, M.P.J.: The stringdist package for approximate string matching. R J. 2 (2014)