Characterization of Written Languages using Structural Features from Common Corpora

Younis Al Rozz, Harith Hamoodat, and Ronaldo Menezes

BioComplex Laboratory, School of Computing
Florida Institute of Technology, Melbourne, USA
yyounis2013@my.fit.edu
hhamdon2013@my.fit.edu
rmenezes@cs.fit.edu

Abstract. *For more than 5,000 years, we have been communicating using some form of written language. For many scholars, the advent of written language contributed to the development of societies because it enabled knowledge to be passed to future generations without considerable loss of information or ambiguity. Today, it is estimated that we use about 7,000 languages to communicate, but the majority of these do not have a written form; in fact, there are no reliable estimates of how many written languages exist today. There are three main families of written languages: Afro-Asiatic, Indo-European, and Turkic. These families of languages are based on historical family-trees. However, with the amount of data available today, one can start looking at language classification using regularities extracted from corpora of text. This paper focus on regularities of 10 languages from the mentioned families. In order to find features for these languages we use (1) Heaps' law, which models the number of distinct words in a corpus as a function of the total number of words in the same corpora, and (2) structural properties of networks created from word co-occurrence in large corpora for different languages. Using clustering approaches we show that despite differences from years of being used in separate countries, the clustering still seem to respect some historical organization of families.*

Keywords: *co-occurrence networks, language classification, Heaps' Law, clustering.*

1 Introduction

The development of society cannot be said to be caused by the advent of writing but writing is certainly linked to modern life as it only appeared around 5,000 years ago. According to Coulmas [15], writing is the most important "sign system" ever invented. It is quite difficult to imagine our society thriving without books, research articles, instruction manuals, lecture notes, etc. The importance of writing is even recognized by many cultures and often its invention is attributed to divine intervention such as god Ganesh in India, or the god Thoth in ancient Egypt.

Writing enables the transmission of information between many generations without any loss of information; it broadens the range of communication of individuals. Today, it is estimated that humans use about 7,000 languages to communicate[1], although this

[1] https://www.ethnologue.com

number is in decline as languages become extinct. Moreover, the majority of these do not have a written form; in fact, there are no reliable estimates of how many written languages exist today. Linguists have been studying languages and how they should be organized for a long time [11] , however most classifications are based on historical or phonetic approaches. There are many families of languages, and few are well known such as: Uralic, Afro-Asiatic, Indo-European, and Turkic. Fig. 1 shows a sample of the Indo-European set of languages.

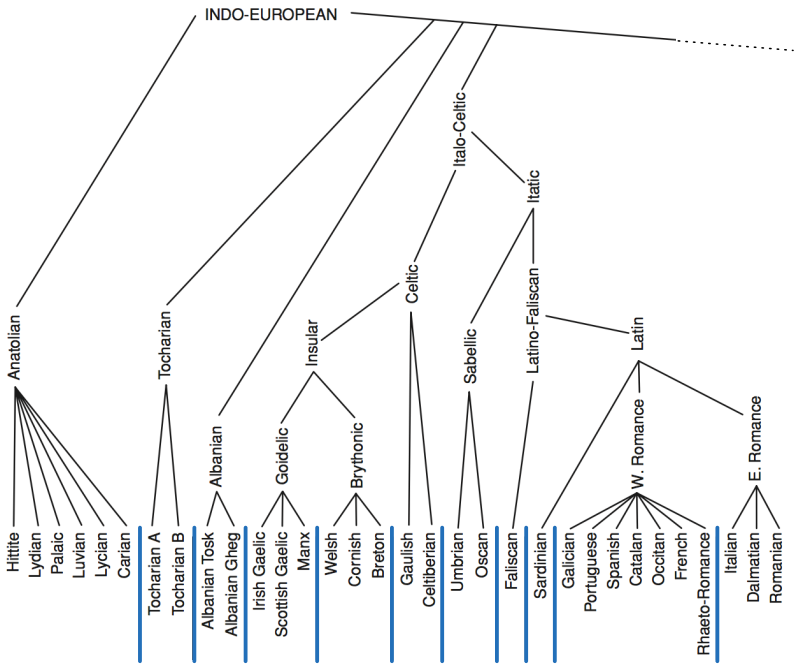

Fig. 1: Part of the family tree of Indo-European languages (adapted from [11]).

The advances in Network Science and Natural Language Processing (NLP) in recent years has motivated researchers to utilize both disciplines together in language classification.

Nowadays studies can be done quantitatively and not only qualitatively. It is quite common to have data regarding any subject of interest. In the context of text analysis, the studies range from discovering language structure [30], classification of languages into families [24, 23, 6, 19], word tagging problems [10], machine translation [2], summarization systems [3], to the improvement of search engines and information retrieval (IR) [28]. Although we review a few of the related work in Section 2, an interested reader can find a deeper analysis of the literature in [30].

The understanding of structural language similarities can lead to metrics to evaluate the quality of one's writing, translations, and even classification of literary styles. It is quite possible that different styles present different writing structures. In this work, we show that even without semantic analysis of the text itself, and focusing solely on the structure built from syntax, we can reveal that characteristics of many languages are common. More specifically, we used statistical measures of a word co-occurrence networks as well as regularities extracted from parameters of Heaps' law to classify 10 world languages. The classification process was performed using two methods: K-Means, and Hierarchical Clustering.

2 Related work

Many researchers have investigated the possibility of using statistical and mathematical modeling to understand regularities in written languages. Chouldhury and Mukherjeee [13] discuss many ways in which networks can be created from text but they all fall into two main categories: lexical networks and word co-occurrence networks. The first category is concerned with cognitive systems and Psycholinguistics studies [7] and can be further classified into phonological [4], semantic [32], and orthographic networks [14]. Phonological networks can be a network of phonemes [27] or syllables [29]. The second type of language network can be further categorized into co-location [25] and syntactic dependency networks [22].

The attempt to use language structure as a classification tool is not entirely new. In fact, Song [31] discussed the concept of *linguistic typology* as a field which looks at the comparison of languages (search for similarities and differences) across all levels of language structure such as syntax, semantics, morphology, and phonology. Three types of linguistic typology exist [8]: qualitative, quantitative, and theoretical.

Liu and Xu constructed syntactic networks for 15 languages using word and lemma form. They analyzed seven network parameters to classify languages and found that word-formed networks are better than lemma networks in classifying languages [24].

Liu and Cong [23] created co-occurrence networks from a text in 14 different languages and used complex network parameters for their classification using hierarchical clustering. Ban et al. [6] built a co-occurrence network using text from five books for three languages and used network measures to find the similarity and differences between those three languages. Gao et al. [19] constructed six directed and weighted word co-occurrence networks based on 100 reports from the United Nations. Then they compared the network measures but they did not perform any clustering.

3 Methodology

3.1 Data Curation and Model

The data was collected from the Leipzig Corpora Collection [20]. The languages chosen for this work were English, Arabic, Russian, Italian, Spanish, French, German, Turkish, Dutch, and Danish; they were chosen to represent three main language families, namely Afro-Asiatic, Indo-European, and Turkic. The text corpus for each language was constructed from Wikipedia and news pages to ensure some vocabulary diversity

and a good representation for each language. The size of the corpus for each language is consistently made of one million sentences. The entire text was converted to lower case, then punctuation and special characters were removed. This work looks at language structure for meaningful words and sequences; stop words (e.g. prepositions, articles, etc.) were removed from the text. These so-called functional words can skew the statistical representation of the words in particular in the context of network science (described later).

3.2 Feature Extraction

One of the best-known characteristics of vocabulary is the Heaps' law (also known as Herdan's law) introduced in the 1960s [21] which describes the vocabulary growth in texts [18]. The law is defined as:

$$V_R(n) = Kn^{\beta}, \tag{1}$$

where V_R is the number of vocabulary words in the text of size n, and K and β are parameters determined experimentally.

Heaps' law represents the vocabulary richness of a certain language, a large text corpus of 10 million words was used for the fitting of the Heaps' law parameters Fig. 2. These parameters are used as a part of the features vector that will be used to characterize the 10 languages used in this work.

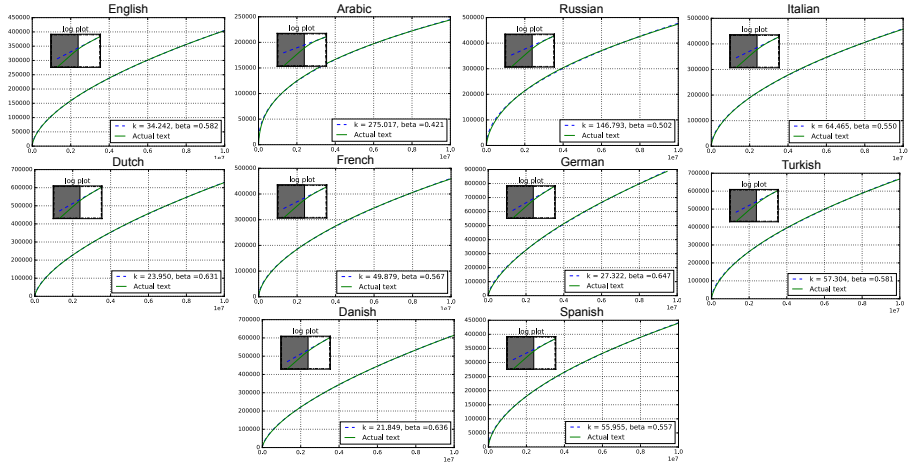

Fig. 2: Fitting of Heaps' law for the 10 languages used in this study (and the value of K and β respectively.

Table 1 shows the values of K and β for the fitting in Fig. 2. For English, the values of K are expected to be between 10 and 100 and the values β between 0.4 and 0.6. Our results agree with this expectation but the values of K for Arabic and Russian is greater than 100.

Table 1: From top to bottom and from left to right the languages in Fig. 2. The values of K and β from Equation 1 is shown

	English	Arabic	Russian	Italian	Dutch	French	German	Turkish	Danish	Spanish
K	34.24	275.01	146.79	64.46	23.95	49.87	27.32	57.30	21.84	55.95
β	0.58	0.42	0.50	0.55	0.63	0.56	0.64	0.58	0.63	0.55

Table 2: Size of the word co-occurrence networks for all 10 languages.

	English	Arabic	Russian	Italian	Dutch	French	German	Turkish	Danish	Spanish
n	18,986	29,995	37,341	31,361	30,475	30,248	39,098	34,945	30,329	29,999
m	77,989	81,046	93,587	94,494	94,427	94,611	95,774	89,385	88,985	94,919

After the fitting of Heaps' law to our corpora, we set to create co-occurrence word networks. Our networks are simple and link words in each corpus if they are adjacent to each other in text. Hence, nodes represent unique words and edges represent the connection between each two consecutive words. The edges' weights represent the frequency in which the two words appear next to each other. Table 2 shows the size of each network in terms of number of nodes n and number of edges m.

The generation of the networks gives us the structure and the values for n and m. Note however from Table 2 that for all languages the values of n and m are very similar which indicates they are not good features to let us characterize the languages. However, there are other structural characteristics that can be computed from the networks.

The average degree $\langle k \rangle$ is generally provided as an information item. These networks tend to display a power-law degree distribution and the average degree does not represent the distribution well. The highest average degree was 8.21 for English and the lowest was 4.89 for German. The reason for this is because the German language's vocabulary is much bigger than that of English [9].

The clustering coefficient of a network (C) is given by the average clustering of the clustering coefficients of each node (C_i) which (informally) captures the extend to which the neighbors of a node i are connected between themselve, this cab be calculated using the equation below:

$$C_i = \frac{2E_i}{k_i(k_i - 1)}, \tag{2}$$

where, E_i is the number of links that exist between the neighbors of node i, and the denominator number of possible links that could exist between nodes i.

Russian and Arabic have the lowest clustering coefficient: 0.012 and 0.019 respectively; English and Danish score the highest: 0.047 and 0.041 respectively. This is due to the fact that Russian and Arabic are morphological languages, which means that they have more word forms than analytic languages such as English and Danish [1].

Another vital characteristic for networks analysis is the average path length. We know that social networks have high C and low average path length (ℓ) computed as:

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}, \tag{3}$$

where d_{ij} is the distance between nodes i and j. Russian has the longest value for ℓ with 4.91 steps, while the shortest one was 3.82 for English. Again, this happens because morphological languages like the Russian and Arabic tend to have a longer path than analytic languages like English and Dutch [1].

Networks can be divided into consistent groups of nodes called communities [16] whose density of edges within the community is higher than outside it. There are many algorithms in the literature proposed to find these communities but one of the classical ways is to calculate the modularity of the network (Q). We computed the value of Q for all 10 networks using the approach proposed by Newman [26]. Based on this metrics, Russian has the largest modularity value of 0.481, while the lowest value was 0.379 scored by English.

The last two parameters, α_d and α_s were obtained by fitting functions to weighted degree distribution of the network and size distribution of communities of words. As shown later in Table 3, the values of α_d are quite close to what is expected for real-world networks ($2 \leq \alpha \leq 3$). We believe the reason for the lower exponent values was the removal of the functional words. Fig. 3 shows that a power law function (i.e. $P(k) \sim k^{\alpha}$, where k represents the node degrees) has the best fit when compared to other common functions of real-world networks.

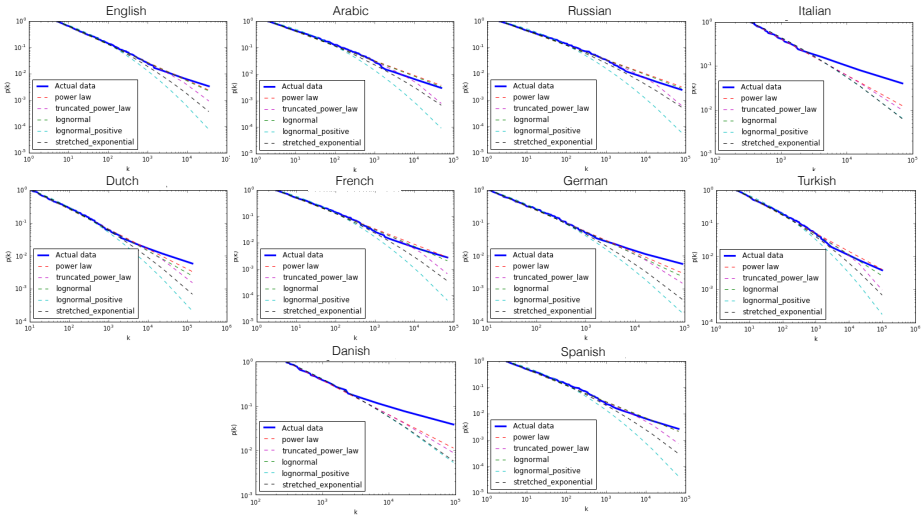

Fig. 3: Fitting of the degree distribution.

Similarly, the α_s value for the distribution of community size shows a good fit with a power law function, which is expected also in real-world networks with community structure; according to Arenas et al. [5] the distribution of community sizes in real network appear to have a power law form $P(s) \sim s^{\alpha}$. Both exponents have been used as part of the feature vector representing the languages. Figure 4 shows the fitting for the community size for all 10 languages and Table 3 shows the values for α_s.

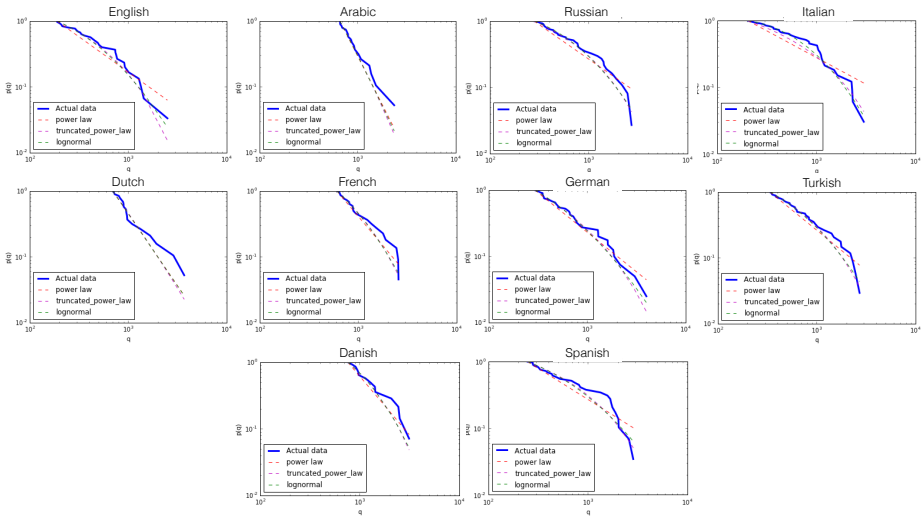

Fig. 4: Fitting of the size distribution in the power law package.

For each of the networks we built, we generated random networks with the same size and using the Erdös-Rényi model. The purpose was to analyze the clustering of our word networks in comparison with a random network. The average clustering coefficient values for the random networks were much smaller than those in the word networks. For example, in Italian, the average clustering coefficient for our network is 0.022 while in the random network was 0.00019. Also, the average path length (ℓ) for the 10 languages was between 3.8 and 4.9 which means our networks appear to be small-world [33].

After all the analysis we had an 8-dimension feature vector for each language as depicted in Table 3. In the next section, we will use these vectors to do a clustering of the languages leading to a classification of them based on their structural similarities.

Table 3: Each line in this table represent 8-dimension feature vector for the language shown in the first column.

Languages	β	K	$\langle k \rangle$	C	ℓ	Q	α_d	α_s
English	0.582	34.242	8.215	0.047	3.824	0.379	1.827	2.070
Arabic	0.421	275.017	5.404	0.019	4.454	0.466	1.508	3.937
Russian	0.502	146.793	5.012	0.012	4.910	0.481	1.660	2.037
Italian	0.550	64.465	6.026	0.022	4.280	0.405	1.751	1.800
Dutch	0.631	23.950	6.197	0.026	4.194	0.388	1.725	3.186
French	0.567	49.879	6.255	0.023	4.213	0.385	1.745	2.774
German	0.647	27.322	4.899	0.023	4.471	0.464	1.689	2.194
Turkish	0.581	57.304	5.115	0.023	4.430	0.471	1.716	2.223
Danish	0.636	21.849	5.868	0.041	4.200	0.438	1.740	2.761
Spanish	0.557	55.955	6.328	0.023	4.239	0.389	1.730	1.934

4 Results and Discussion

We have executed clustering using two known algorithms: K-Means and Hierarchical Clustering. Recall that the purpose of this work is to classify languages according to the features extracted from Heaps' law and network properties.

4.1 K-means clustering

K-Means is a fast and widely-used clustering algorithm that works by minimizing the sum-of-squares distance of the data points within the cluster. The number of clusters must be specified in advance, so two methods were used to find the optimal number of clusters. The first one is the silhouette method; it provides a visual aid in determining the number of clusters. The silhouette coefficient which ranged between -1 and 1 indicates the closeness of each data point in a cluster to other points in the neighboring clusters. After that, we used the elbow method to validate the number of clusters found in the silhouette method.

Due the high dimensionality of the feature vectors, we run a Principle Component Analysis (PCA) to reduce the dimensionality of the features vector to two dimensions so that the resulting K-Means clusters can be visualized. We also wanted to independently check whether the parameters extracted from the Heaps' law were providing extra information to the clustering of the feature vectors. The silhouette method was applied with and without the two Heaps' law parameters (K and β). In the first case, the optimal number of clusters was three. When the Heaps' parameters were added, the silhouette plot suggests a number of clusters between four and five as a good choice (Fig. 5). These results indicate the importance of the Heaps' parameters to the process of the language classification.

The elbow method was used to validate the optimal number of clusters found by the silhouette method. The elbow plot suggests an optimal number of three clusters when the two Heaps' parameters are not considered, which agreed with the results of the silhouette method. The result of the K-means clustering for this case was that Italian, Spanish, German, Russian, and Turkish clustered together. The second cluster contains French, Danish, Dutch, and English, while Arabic appeared in its own cluster.[2] When adding the parameters of the Heaps' law, the elbow of the curve indicates an optimal number of four clusters (Fig. 6(b)). In this case, Italian, Spanish, French, Danish, and Dutch were clustered together. The second cluster contains Russian, German, and Turkish, while English and Arabic separated into their own clusters (Fig. 6(a)), which also supports the results of the silhouette method indicating the importance of Heaps' parameters to the classification process and the fact that the complete set of parameters offers a higher granularity for the clustering. These results match, to a certain degree, the linguistic typology classification of languages into genetic families as the Arabic language belongs to the Afro-Asiatic family, while the rest of the languages belong to the Indo-European Family.

An interesting finding from the clustering process is Turkish, which belongs to the Turkic family, was clustered with the Indo-European Family. As the aim of this work is

[2] We again decided to show the charts only for the case with the Heaps' parameters due to space restrictions.

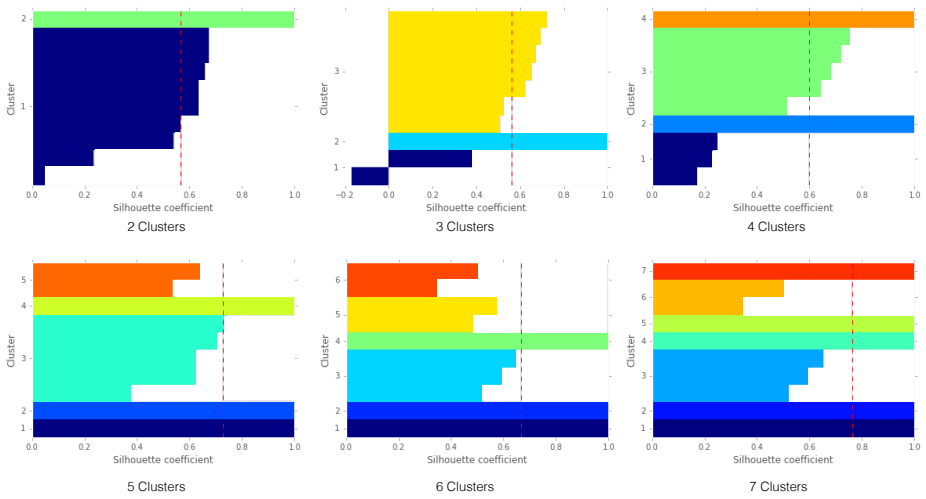

Fig. 5: Silhouette analysis on K-Means clustering where the value of the Heaps' law parameters were included after the PCA. The same analysis has been done for the case without the use of the Heaps' law parameters which we did not include a picture due to space limitations in the paper.

to classify languages based on lexical rather than syntactical perspective, the removal of the functional words (stop words) has affected the structure of the languages networks [12]. This in turn has reduced the syntactic barriers between languages belonging to different families. The addition of the Heaps' law parameters enforced the separation of the languages based on their vocabulary richness and lexical structure represented by the network statistics.

In light of the previous assumption, the development of languages seen in the modern age, caused by the effects of technology, globalization, and migration among other factors, has had on effect on languages classification. For the case of the Turkish language, as of the year 2011, three million Turkish people were living in Germany, representing 3.6% of the German population [17].

4.2 Hierarchical clustering

The results of K-means clustering can only classify languages from the top level of the family tree. To find the relationships between languages in a more structured way we applied a hierarchical clustering to the language feature vectors. In this case, we decided to also test whether the heaps' law features alone would provide a similar classification to the classification based on network features alone. Fig. 7(b) show the classification using only the Heaps' parameters while Fig. 7(c) shows the same results using only network parameters. Although both classifications have interesting characteristics that resemble traditional language classifications, the combination of both features in Fig. 7(a) yields a classification that appears to be enhanced. For instance, the distance between the Turkish and German languages was increased.

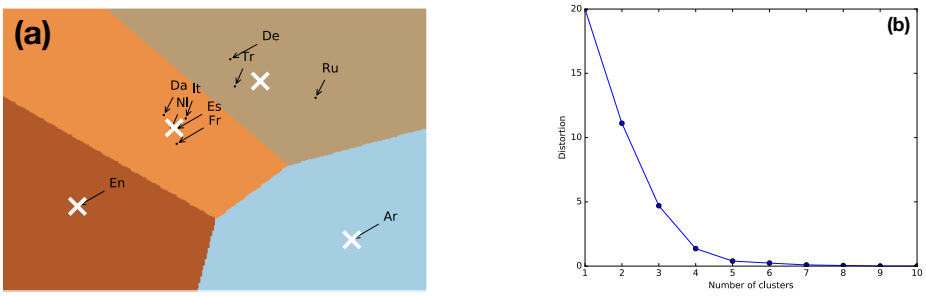

Fig. 6: (a) K-means clustering after PCA and using Heaps' law parameters and network parameters. (b) The elbow rule shows that four clusters appear to be the best choice for the K-means.

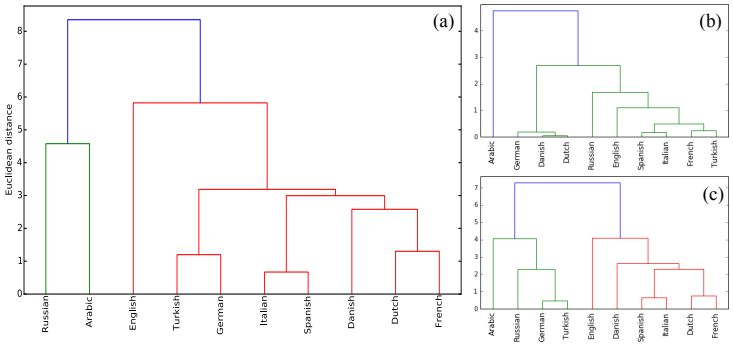

Fig. 7: Hierarchical clustering of the 10 languages used in our study. (a) Shows the classification using the network parameters as well as the Heaps' law parameters while (b) shows the classification using Heaps' law parameters and (c) network parameters separately.

5 Conclusion

The understanding of languages and their characterization has again become a topic of interest for the scientific community. Studies using large amounts of data may be able to provide a different view of how languages relate to one another and see possible trends or influences of one over the other.

In our study, we look at the possibility of characterizing written language solely from the point of view of structural features. We concentrated on two class of features: Heaps' law, which looks at richness of vocabulary in a language, and Network Science features extracted from the construction of word co-occurrence networks. In the process of extracting network features, we also demonstrated that these networks exhibit both scale-free and small-world properties.

We used K-Means and Hierarchical Clustering together with the silhouette and elbow methods to identify the optimal number of language clusters to the dataset we have.

We showed that the hierarchical clustering distinguish relationships between languages sub-families, while K-Means clusters languages based on their main genetic families (Proto-Families). We also showed that the Heaps' law parameters enhanced the classification process by distinguishing languages based on their vocabulary richness.

Following this work, we would like to go deeper in the characterization of of languages by augmenting the number of languages we use from 10 to around 30 or 40 languages. The difficulty is to find good corpora that includes this number of languages. Also, we believe structural analysis of written language could be used in identification of literary styles or even author analysis. It would be interesting to perform a similar analysis for several languages and understand if authors have a structural fingerprint in their writing style that can be identified and whether this fingerprint resist the translations of their texts.

References

1. Olga Abramov and Alexander Mehler. Automatic language classification by means of syntactic dependency networks. *Journal of Quantitative Linguistics*, 18(4):291–336, 2011.
2. Diego R Amancio, Lucas Antiqueira, Thiago AS Pardo, Luciano da F. Costa, Osvaldo N Oliveira Jr, and Maria GV Nunes. Complex networks analysis of manual and machine translations. *International Journal of Modern Physics C*, 19(04):583–598, 2008.
3. Lucas Antiqueira, Osvaldo N Oliveira, Luciano da Fontoura Costa, and Maria das Graças Volpe Nunes. A complex network approach to text summarization. *Information Sciences*, 179(5):584–599, 2009.
4. Samuel Arbesman, Steven H Strogatz, and Michael S Vitevitch. The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, 20(03):679–685, 2010.
5. Alex Arenas, Leon Danon, Albert Diaz-Guilera, Pablo M Gleiser, and Roger Guimera. Community analysis in social networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):373–380, 2004.
6. Kristina Ban, Ana Meštrović, and A Martinčić-ipšić. Initial comparison of linguistic networks measures for parallel texts. In *5th International Conference on Information Technologies and Information Society (ITIS), 97104*. Citeseer, 2013.
7. Nicole M Beckage and Eliana Colunga. Language networks as models of cognition: Understanding cognition through language. In *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, pages 3–28. Springer, 2016.
8. Balthasar Bickel. Typology in the 21st century: major current developments. *Linguistic Typology*, 11(1):239–251, 2007.
9. Christian Biemann, Stefan Bordag, Gerhard Heyer, Uwe Quasthoff, and Christian Wolff. Language-independent methods for compiling monolingual lexical data. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 217–228. Springer, 2004.
10. Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4):543–565, 1995.
11. Lyle Campbell and William J Poser. Language classification. *History and method. Cambridge*, 2008.
12. Xinying Chen and Haitao Liu. Function nodes in chinese syntactic networks. In *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, pages 187–201. Springer, 2016.
13. Monojit Choudhury and Animesh Mukherjee. The structure and dynamics of linguistic networks. In *Dynamics on and of Complex Networks*, pages 145–166. Springer, 2009.

14. Monojit Choudhury, Markose Thomas, Animesh Mukherjee, Anupam Basu, and Niloy Ganguly. How difficult is it to develop a perfect spell-checker? a cross-linguistic analysis through complex network approach. *TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, page 81, 2007.
15. Florian Coulmas. *The writing systems of the world*. B. Blackwell, 1989.
16. Henrique F. de Arruda, Luciano da F. Costa, and Diego R. Amancio. Topic segmentation via community detection in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(6):063120, 2016.
17. Deutschland and Statistisches Bundesamt Deutschland. *Statistisches Jahrbuch Deutschland und Internationales*. Statistisches Bundesamt, 2012.
18. Francesc Font-Clos, Gemma Boleda, and Álvaro Corral. A scaling law beyond zipf's law and its relation to heaps' law. *New Journal of Physics*, 15(9):093033, 2013.
19. Yuyang Gao, Wei Liang, Yuming Shi, and Qiuling Huang. Comparison of directed and weighted co-occurrence networks of six languages. *Physica A: Statistical Mechanics and its Applications*, 393:579–589, 2014.
20. Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, pages 759–765, 2012.
21. Gustav Herdan. *Type-token mathematics*, volume 4. Mouton, 1960.
22. Ramon Ferrer i Cancho. The structure of syntactic dependency networks: insights from recent advances in network theory. *Problems of quantitative linguistics*, pages 60–75, 2005.
23. HaiTao Liu and Jin Cong. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10):1139–1144, 2013.
24. Haitao Liu and Chunshan Xu. Can syntactic networks indicate morphological complexity of a language? *EPL (Europhysics Letters)*, 93(2):28005, 2011.
25. Nuno Mamede, José Correia, and Jorge Baptista. Syntax deep explorer. In *Computational Processing of the Portuguese Language: 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings*, volume 9727, page 189. Springer, 2016.
26. Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.
27. Cynthia SQ Siew. Community structure in the phonological network. *Frontiers in psychology*, 4:553, 2013.
28. Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.
29. M Medeiros Soares, G Corso, and LS Lucena. The network of syllables in portuguese. *Physica A: Statistical Mechanics and its Applications*, 355(2):678–684, 2005.
30. Ricard V Solé, Bernat Corominas-Murtra, Sergi Valverde, and Luc Steels. Language networks: Their structure, function, and evolution. *Complexity*, 15(6):20–26, 2010.
31. Jae Jung Song. *The Oxford Handbook of Linguistic Typology*. Oxford University Press, 2010.
32. Mark Steyvers and Joshua B Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78, 2005.
33. Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.