Multilayer network of language: A unified framework for structural analysis of linguistic subsystems

Sanda Martinčić-Ipšić *, Domagoj Margan, Ana Meštrović

Department of Informatics, University of Rijeka, Croatia

HIGHLIGHTS

- The multilayer complex network of languages.
- The model is based on three aspects: construction, linguistic subsystem and language.
- Word-level layers: syntax and co-occurrence; subword-level: syllable and grapheme.
- We introduced: preserved weighted overlap as the measure of word-level layers similarity.
- Word vs. subword layers relations characterized by correlations of triad significance profiles.

ARTICLE INFO

ABSTRACT

Recently, the focus of complex networks' research has shifted from the analysis of isolated properties of a system toward a more realistic modeling of multiple phenomena — multilayer networks. Motivated by the prosperity of multilayer approach in social, transport or trade systems, we introduce the multilayer networks for language. The multilayer network of language is a unified framework for modeling linguistic subsystems and their structural properties enabling the exploration of their mutual interactions. Various aspects of natural language systems can be represented as complex networks, whose vertices depict linguistic units, while links model their relations. The multilayer network of language is defined by three aspects: the network construction principle, the linguistic subsystem and the language of interest. More precisely, we construct a word-level (syntax and co-occurrence) and a subword-level (syllables and graphemes) network layers, from four variations of original text (in the modeled language). The analysis and comparison of layers at the word and subword-levels are employed in order to determine the mechanism of the structural influences between linguistic units and subsystems. The obtained results suggest that there are substantial differences between the networks' structures of different language subsystems, which are hidden during the exploration of an isolated layer. The word-level layers share structural properties regardless of the language (e.g. Croatian or English), while the syllabic subword-level expresses more language dependent structural properties. The preserved weighted overlap quantifies the similarity of word-level layers in weighted and directed networks. Moreover, the analysis of motifs reveals a close topological structure of the syntactic and syllabic layers for both languages. The findings corroborate that the multilayer network framework is a powerful, consistent and systematic approach to model several linguistic subsystems simultaneously and hence to provide a more unified view on language.
© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

---

* Correspondence to: Department of Informatics, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia. Tel.: +385 51 584 714.
E-mail addresses: smarti@uniri.hr (S. Martinčić-Ipšić), dmargan@uniri.hr (D. Margan), amestrovic@uniri.hr (A. Meštrović).

1. Introduction

Recently, the field of complex networks has shifted from the analysis of isolated network (capturing and modeling one aspect of the examined system) toward the analysis of the family of complex networks simultaneously modeling different phenomena (aspects) of the examined system, or modeling interactions and relationships among different subsystems. The rise of this more realistic framework for a complex network analysis considers different layers, levels or hierarchies for different aspects of the system. In other words, multiple phenomena are characterized by multiple types of links across various levels of representations or various dimensions of relations for multiple subsystems. The multilayer network approach has been addressed in the analysis of real international trade analysis [1], social interactions in the massive online game [2], web-search queries [3], in transport and infrastructure [4–7] and in the examination of the brain's function [8]. There are variations in formal representation of the multilayer networks [4,9], multidimensional networks [3], multiplex networks [6,7,10], interdependent networks [11,5] and networks of networks [12,13]. A thorough discussion that compares, contrasts, and translates between notions of multilayer, multiplex, interdependent networks and networks of networks is in Ref. [14], which together with Ref. [15] presents a detailed overview of multilayer network theory.

Viewed as a unique, biologically-based human faculty [16], language has been recognized as the reflection of the human cognitive capacities, both in terms of its structure and its computational characteristics [17]. Studying languages at intra- and cross-linguistic levels is of paramount importance in relation to our biological, cultural, historical and social beings. Hence, human languages, besides still being our main tools of communication, reflect our history and culture. Language can be seen as a complex adaptive system [18], evolving in parallel with our society [19].

Various aspects of natural language systems can be represented as complex networks, whose vertices depict linguistic units, while links model their morphosyntactic, semantic, pragmatic, etc. interactions. Thus the language network can be constructed at various linguistic levels: syntactic, semantic, phonetic, syllabic, etc. So far there have been efforts to model the phenomena of various language subsystems and examine their unique function through complex networks. Still, the present endeavors in linguistic network research focus on isolated linguistic subsystems lacking to explain (or even explore) the mechanism of their mutual interaction, interplay or inheritance. Obtaining such findings is critical for deepening our understanding of conceptual universalities in natural languages, especially to shed light on the cognitive representation of the language in the human brain [20]. The pioneering research by Liu and Cong [21] examines the Chinese language as a multi-level system and opens a new perspective on modeling and examining languages in a multilayer complex network framework.

One of the main open questions in linguistic networks is explaining how different language subsystems mutually interact [18,22]. The complexity of any natural language is contained in the interplay among several language levels. Below the word-level, it is possible to explore the type of phonology, morphology and syllabic subsystem complexity. For example, the phonology subsystem complexity is reflected in the morphology subsystem complexity. On the word-level, the morphology subsystem complexity reflects in the complexity of the word order, syntactic rules and the ambiguity of lexis. Since the word order can be considered as the primary factor (but not the only one) that determines linguistic structure, it is important to explore the subsystems' interactions by which it is influenced.

In this research we use the multilayer network framework to explore the structural properties of various language subsystems and their mutual interactions. The multilayer network of a language is constructed for the word (co-occurrence and syntax) and subword (syllables and graphemes) language levels. The systematic exploration of layers properties is presented for the Indo-European family of languages: one representative of the Slavic group – Croatian, and one representative of the Germanic group – English. The analysis and comparison of layers are employed in order to determine structural influences and trade-offs between the subsystems of language.

Our work contributes to the field of linguistic network research by proposing the multilayer network model for language. More precisely the multilayer language network model is established on three aspects: the network construction principle, the linguistic subsystem and the language of interest. Moreover, we introduce the preserved weighted overlap as the measure of word-level layers similarity in weighted and directed networks, that expresses the quantity of weights that are shared in both layers. Finally, we propose the characterization of word vs. subword layers relationships by correlations of triad significance profiles, as a possible quantification of the inter layer relationships.

1.1. Related work

The language networks. The seed work of Dorogovtsev and Mendes [23] describes language as a selforganizing network of linked words. The observed word web structure distributions naturally emerge from the evolutionary dynamics. Masucci and Rodgers [24] investigate the topology of Orwell's 1984 within the framework of complex network theory. They exhibit local preferential attachment as growth mechanism of written language and the allocation of a set of preselected vertices that have a structural rather than a functional purpose. Choudhury and Mukherjee in Ref. [25] provide a suitable framework to model a language from three different perspectives microscopic (utterances), macroscopic (grammar rules and a vocabulary) and mesoscopic (linguistic entities—letters, words or phrases). The authors mainly present an overview of the structure and dynamics at the mesoscopic level. Sole et al. [26] review the state-of-the-art on language networks and their potential relevance to cognitive science. They also consider the intertwining of language levels related to multiple layers of complexity

in terms of the networks of connected words in order to shed light onto the relevant questions concerning language organization and its evolution. In Ref. [27] Cong and Liu provide an extensive insight into the language networks which positions human language as a multi-level system in the discipline of complex network analysis. Relationships between the system-level complexity of human language (determined by the topology of linguistic networks) and microscopic linguistic features (as the traditional concern of linguistics) are positioned within a holistic quantitative approach for linguistic inquiry, which contributes to the understanding of human language at different granularities. The same authors Liu and Cong in Ref. [21] present the pioneering empirical study which examines the Chinese language as a multilayer system. The authors construct four layers of language subsystems semantic, syntax, co-occurrence and characters (monosyllabic morphemes) from the same corpus of text. The topological properties of complex networks are used to describe and compare the organizational structure of the four language layers and to explain their relationships. The empirical results suggest that the semantic subsystem exhibits a unique status in comparison with the other three layers which are more related in terms of topological properties.

The word-level networks: co-occurrence. The construction of language networks relies on the well-established principles of modeling word interaction from the word order in a sentence or in short from their co-occurrence in text. A substantial part of reported research on language networks is dedicated to a detailed structural analysis of co-occurrence networks interpreting their topological properties in the linguistic context [23–25,28,29]. Thus, in the linguistic co-occurrence networks properties are derived directly from the word order in texts by connecting words within a window of certain size or sentence. Still, the open question is how the word order itself is reflected in topological properties of the linguistic network. One approach to address this question is to compare networks constructed from normal texts with the networks constructed from syntax dependencies in texts.

The word-level networks: syntax. The syntactic structure of language is captured through syntax dependency relations between a pair of words in a sentence: the head word – the governor of relationship and the dependent word – the modifier. Syntax dependencies between words are formally expressed by dependency grammar (e.g. a set of productions (rules) in the form of a grammar). The dependency grammar is used to parse the syntactic relationships from a sentence in the form of a syntax dependency tree. Thus, the syntax dependency treebank is the set of syntax dependency trees parsed from the sentences in a corpus. Ferrer i Cancho et al. [30], in the seed work on syntax complex networks model the syntactic dependency relationships of three languages comparatively (Czech, German, Romanian). The set of analyzed languages is extended to 7 in Ref. [31], comparing the structure of global syntactic dependency networks. The results in Refs. [31,32] show that the proportion of syntactically incorrect relationships rises from about 30% to a high 50% in a co-occurrence networks constructed with a window of size 2 and 3 respectively. In Ref. [33], based on the comparison of one syntactic dependency network and two random dependency networks of Chinese, the authors confirm small-world and scale-free properties, suggesting that scale-free architecture is of essential importance to the syntax subsystem of human language. Authors extend their work in Ref. [21] where they examine the topological properties of four layers of the Chinese language subsystems: semantic, syntax, co-occurrence and characters (monosyllabic morphemes). Liu et al. [34] and Abramov and Mehler [35] use network parameters derived from the syntax relationships for hierarchical clustering of languages, deriving the model of the genealogical similarity among 15 and 11 languages respectively. The obtained results on syntax networks suggested that a natural approach to modeling human language is considering the structure of the syntactic dependency relationships besides the simple word-order relations reflected in co-occurrence networks. Amancio et al. [36] explore the Portuguese syntax dependencies for automatic summarization of the news.

The subword-level networks: syllables. The coherent results from language networks involving units smaller than words, such as syllables [37,38], phonemes [39] or morphemes [35] are in the premature stage. Morphological networks for English and German are presented in Ref. [35] and the network properties are expressed in terms of graph entropy measures. So far, syllable networks have been constructed exclusively for Portuguese [38] and Chinese [21,37]. Syllables are a natural intermediate level in the analysis of spoken (as opposed to written) language, since they carry prosody during pronunciation. The investigation of syllables is particularly interesting for their role in language acquisition. Children begin to learn language through syllables, culminating in the development of their mental lexicons [20,40]. The model of language acquisition was recreated with humanoid robots using syllables as basic units [41] or by artificial agents [42]. Both studies witness the complexity of a language syllable system as an important factor in language acquisition.

The subword-level networks: graphemes. Language is written with a set of abstract orthographic symbols (letters of an alphabet) graphemes. Graphemes are the smallest semantically distinguishing units (the basic linguistic units) in a written language, analogous to the phonemes in spoken language. The complex networks of grapheme subsystem of language have been studied sporadically [43]. Kello and Beltz analyzed the structure of the complex network constructed from the orthographic wordform lexicon, where words are connected if one is a substring of the other. Phonemes have attracted more attention since many psycholinguistics studies regarding the representation of mental lexicon used for speech production, word recognition and language processing have been reported [44–48]. Phonetic networks are typically constructed from words in a lexicon, establishing links among phonetically similar words differing in one phoneme.

Network motifs for language. Motifs are subgraphs defined as simple building blocks of directed complex networks [49]. Motifs are used to detect the structural similarities and differences between networks on the local level. In Ref. [50] the significance profiles of motifs derive several superfamilies of networks – the language networks forms one supra family based on the triad significance profile. Biemann et al. in Ref. [51] use motifs to quantify the differences between natural

and generated language. The frequencies of three-vertex and four-vertex motifs for six languages are compared with the generated language from n-gram statistical model (n-grams are a sequence of n units from a given text). The authors show that the four-vertex motifs are directly interpretable by semantic relations of polysemy and synonymy. An initial attempt to analyze undirected triads in a multiplex network, by representing positive and negative social interactions of game players in massive online game is reported in Ref. [2].

The linguistic features of Croatian and English. A short recapitulation of the main properties of the Croatian and English languages establishes the linguistic framework needed for the comparison across languages as well as for the interpretation of insights into their structural characteristics. Croatian is a highly flective Slavic language and words can have seven different cases for singular and seven for plural, genders and numbers. The Croatian word order is mostly free, especially in non-formal writing. These features place Croatian among morphologically rich and mostly free word-order languages. English grammar has minimal inflection compared with most other Indo-European languages, therefore it is considered to be analytic. English word order is almost exclusively subject-verb-object. Both languages are characterized by an accentuation system developed on syllables.

English has been studied extensively in a complex networks framework [24–26,29,34,51], still no systematic effort explaining the effects of mutual interaction of different subsystems has been reported. So far the Croatian has been quantified in a complex networks framework based on the word co-occurrences [28,29]. The syntax relationships of Croatian as well as syllabic subword units are novelty characterized through the lenses of the analysis of complex networks in this research.

2. Methods

More details about complex networks analysis and the definition of measures can be found in Ref. [52]. Here we list a short definition of measures needed for the exploration of network layers. The network $G = (V, E)$ is a pair of a set of vertices V and a set of links E, where N is the number of vertices and K is the number of links. In weighted networks every link connecting two vertices i and j has an associated weight w_{ij}. The number of network components is denoted by ω.

For every two connected vertices i and j the number of links lying on the shortest path between them is denoted as d_{ij}, then the average path length between every two vertices i, j is $L = \sum_{i,j} \frac{d_{ij}}{N(N-1)}$. If the number of components $\omega > 1$, L is computed for the largest connected component in network. If in a directed network there is no path between two vertices, then shortest path between two vertices is assigned to 0.

For weighted networks the clustering coefficient of a vertex i is defined as the geometric average of the subgraph link weights: $c_i = \frac{1}{k_i(k_i-1)} \sum_{j,k} (\hat{w}_{ij} \hat{w}_{ik} \hat{w}_{jk})^{1/3}$, where k_i is the degree of the vertex i, and the link weights $\hat{w}_{ij}$ are normalized by the maximum weight in the network $\hat{w}_{ij} = w_{ij} / \max(w)$. The value of c_i is assigned to 0 if $k_i < 2$. The average clustering coefficient of a network is defined as the average value of the clustering coefficients of all vertices in an undirected network: $C = \frac{1}{N} \sum_i c_i$.

The transitivity of a network is the fraction of all possible triangles present in the network. Possible triangles are identified by the number of triads (two links with a shared vertex): $T = (3\#triangles)/(\#triads)$.

The in-degree and out-degree $k_i^{in/out}$ of vertex i is defined as the number of its in and out nearest neighbors. The in-strength and the out-strength $s_i^{in/out}$ of the vertex i is defined as the number of its incoming and outgoing links, that is: $s_i^{in/out} = \sum_j w_{ji/ij}$.

The in- and out-selectivity of the vertex i is then defined as proposed in Ref. [24]:

$$e_i^{in/out} = \frac{s_i^{in/out}}{k_i^{in/out}}. \tag{1}$$

The power law distribution is defined as: $P(k) \sim k^{-\gamma}$ where γ is the power-law exponent.

2.1. Network motifs analysis

Network motifs are connected and directed subgraphs (of three to up to eight vertices) occurring in complex networks at numbers that are significantly higher than those in randomized networks with the same degree distribution [49,50]. Here, we analyze only triads (all possible directed three-vertex subgraphs) by calculating their frequencies, Z-scores and triad significance profiles (*TSP*).

The scores Z_i for each triad i is calculated using equation:

$$Z_i = \frac{N_i^{orig} - \langle N_i^{rand} \rangle}{\sigma_i^{rand}}, \tag{2}$$

where N_i^{orig} is the count of appearances of the triad i in the original network, while $\langle N_i^{rand} \rangle$ and σ_i^{rand} are the average and the standard deviation of the counts of the triad i over a sample of randomly generated networks.

The triad significance profile *TSP* is the normalized vector of statistical significance scores Z_i for each triad i $TSP_i = \frac{Z_i}{\sqrt{\sum_i Z_i^2}}$.

2.2. The multilayer network

Since language networks can be viewed through different aspects: different levels (e.g. word-level, subword-level, sentence-level), different construction rules (e.g. co-occurrence, syntax, shuffle), different languages, etc. there is a need for a general network model that can capture all these aspects in one single framework. Therefore, we propose an application of general multilayer networks model introduced by Kivelä et al. in Ref. [14] to the multilayer language networks [53].

According to Ref. [14], a multilayer network can have any number d of aspects defined as a sequence $L = \{L_a\}_{a=1}^d$. There is one set of elementary layers L_a for each aspect a. In a multilayer network it is possible to construct a set of layers by assembling a set of all of the combinations of elementary layers using a Cartesian product $L_1 \times \cdots \times L_d$.

The multilayer network is a quadruplet $M = (V_M, E_M, V, L)$, where $V_M \subseteq V \times L_1 \times \cdots \times L_d$ that contains only the vertex-layer combinations in which a vertex is present in the corresponding layer, and where E_M is a set of pairs of the possible combinations of vertices and elementary layers, $E_M \subseteq V_M \times V_M$. V is a set of all vertices in all layers. Multiplex is a special case of multilayer network, which satisfies the condition that the set of vertices is shared across layers. Thus, in a multiplex network inter layer connections between different layers have 1:1 or 0:1 cardinality of relationships.

Next we present equations for the calculation of the overlap between two layers. These equations can be applied only to a multiplex network, when two layers share the same vertices (e.g. in our case it is applicable only to the construction aspects of the word-level layers in one language). In the following text we use only α and α' for the shorter notation of the one layer in the multilayer network.

Jaccard index for link overlap between two network layers α and α' is:

$$J(E_\alpha, E_{\alpha'}) = \frac{|E_\alpha \cap E_{\alpha'}|}{|E_\alpha \cup E_{\alpha'}|}. \tag{3}$$

In the same way we can calculate the Jaccard index for weighted overlap (W).

In Ref. [10] total weighted overlap $O^{(w),\alpha,\alpha'}$ between layer α and layer α' is defined as:

$$O^{(w),\alpha,\alpha'} = \sum_{i<j} \min\left(\frac{w_{ij}^\alpha}{w_{\max}^\alpha}, \frac{w_{ij}^{\alpha'}}{w_{\max}^{\alpha'}} \right) \tag{4}$$

where $w_{\max}^\alpha$ is the maximal weight in layer α.

Inspired by total weighted overlap [10] we define the preserved weighted ratio on intersected links between network layers α and α' as:

$$PW(E_\alpha, E_{\alpha'}) = \sum_{i,j} \frac{\min(w_{ij\alpha}, w_{ij\alpha'})}{\max(w_{ij\alpha}, w_{ij\alpha'})}. \tag{5}$$

Then the preserved weighted overlap (WO) is a normalized preserved weighted ratio by the number of intersected links:

$$WO(E_\alpha, E_{\alpha'}) = \frac{PW(E_\alpha, E_{\alpha'})}{|E_\alpha \cap E_{\alpha'}|}. \tag{6}$$

2.3. Croatian and English datasets

The datasets for multilayer Croatian networks are derived from the HOBS corpus—the first version of the Croatian Dependency Treebank [54]. HOBS is extracted as a part of the Croatian National Corpus [55] and annotated at the analytical layer following the Prague Dependency Treebank formalism adapted to Croatian. The corpus size is currently 3465 sentences (88,045 tokens).

The English dataset contains 3829 sentences (94,084 tokens) from the Penn Treebank corpus [56,57]. The size of the extracted Penn subset is intentionally of the same size as HOBS in order to allow for systematic comparisons across the layers, constructed as comparable corpus. Comparable corpus considers similar texts in more than one language (e.g. similar in size, dedicated to the same domain of interest, of the same genre etc.).

Multilayer Croatian (HR) and English (EN) networks are constructed from four variations of HOBS and Penn corpora: two on the word-level (syntax and co-occurrence) and two on the subword-level (syllables and graphemes). More clearly, four different realizations of the very same text in one language are used to construct the network layers (all weighted and directed) using four different relationships among the linguistic units: syntax (SIN), co-occurrence (CO), syllables (SYL) and graphemes (GR).
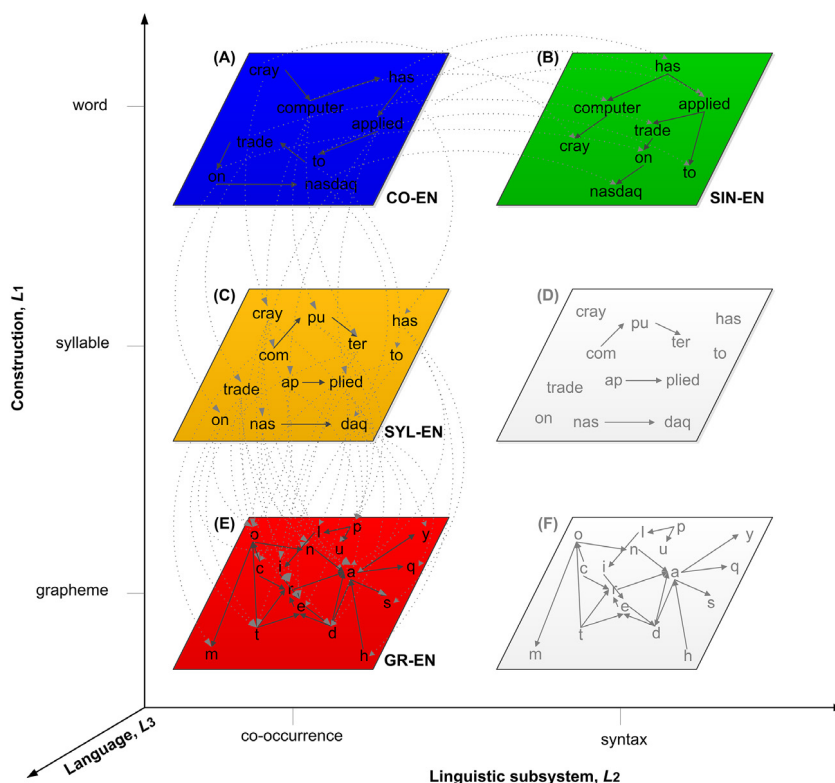
Fig. 1. The multilayer language network. Two word-level layers: (A) co-occurrence; (B) syntax and two subword-level layers: syllables (C); and graphemes (E) constructed from the English sentence "*Cray Computer has applied to trade on NASDAQ.*"; according to three aspects of multilayer network model of language: construction, linguistic subsystem and language. Note-layers (D) and (F) are gray, since they are disregarded in analysis (identical with layers (C) and (E) respectively).

2.4. Language networks construction

The language networks construction principle arises from the vary nature of text (and speech), which is always advancing in an onward direction, hence to use directed and weighted links representing relations among linguistic units [24,27,28]. The co-occurrence relation is established between two adjacent words within a sentence (CO), where the direction of link reflects the words sequencing and weight on the link reflects the frequency of words-pair mutual appearance.

The syntax relationships among word-pairs are parsed from the HOBS and Penn, as well as the text of the original sentences [56]. The sentences' boundaries are preserved, since the syntax dependency is inherent to the sentence (SIN). Thus, the sentence boundaries are considered as linkage delimiters for the co-occurrence layers as well. Fig. 1 (top part) presents the principles of word-level layers' construction for one sentence.

Next, we use the Croatian syllabification with a maximal onset algorithm to prepare the third dataset syllables [58], again from the words in the original sentences. The English syllables are obtained from the dictionary with syllabified words [59]. The process omitted words which were not contained in the syllabified dictionary. The syllable layers are constructed from the co-occurrence of syllables within words (SYL)—presented at the (C) part of Fig. 1.

Finally, we consider the set of graphemes present in words, where graphemes (GR) represent the most elementary subsystem of each language—orthographical. Since, there are some foreign words present in the used corpora we preserved the original orthographic symbols, resulting in a slightly larger number of graphemes (e.g. in Croatian foreign names contain original diacritic symbols, so we obtained q, w, x, y as Croatian graphemes as well)—(E) part of Fig. 1.

Multilayer language network for this work can be defined with the set L of three aspects: construction L_1, linguistic subsystem L_2 and language L_3, where $L_1 = \{co\text{-}occurrence, syntax\}$, $L_2 = \{word, syllable, grapheme\}$ and $L_3 = \{Croatian, English\}$. Therefore, it is possible to have 12 different layers in total ($2 \times 3 \times 2$), although not all the layers are of equal interest. More precisely, one can note that some layers are equal due to the specific construction rules. Since we connect only neighboring syllables within the word, two layers (*co-occurrence, syllable, Croatian*) and (*syntax, syllable, Croatian*) are equal. The same holds for English syllables, and for graphemes in both languages as well, as shown gray for (D) and (F) parts of Fig. 1.

It is worth noticing, that the word-level layers are forming the multiplex networks (have 1:1 inter-connections), while the connections between word and subword layers are not coupled (have N:M inter-connections).

Table 1
The standard network measures for eight layers.

	Croatian				English			
	CO	SIN	SYL	GR	CO	SIN	SYL	GR
N	23 359	23 359	2634	34	10 930	10 930	2599	26
K	71 860	70 155	18 849	491	50 299	52 221	6053	333
L	4.01	1.81	1.86	1.58	3.47	1.96	1.88	1.51
C	0.167	0.120	0.255	0.636	0.286	0.153	0.057	0.838
T	0.004	0.003	0.120	0.522	0.009	0.014	0.020	0.654
ω	2	2	17	1	3	3	54	1

Measures (N no. of vertices, K no. of links, L avg. path length, C clust. coeff., T transitivity, ω no. of components) for co-occurrence (CO), syntax (SIN), syllable (SYL) and grapheme (GR) network layers in Croatian and English.

To sum up, in total we construct eight layers: four of Croatian *(syntax, word, Croatian), (co-occurrence, word, Croatian), (co-occurrence, syllable, Croatian), (co-occurrence, grapheme, Croatian)* and four of English language *(syntax, word, English), (co-occurrence, word, English), (co-occurrence, syllable, English), (co-occurrence, grapheme, English)*, with shortened notations: SIN–HR, CO–HR, SYL–HR, GR–HR, SIN–EN, CO–EN, SYL–EN and GR–EN.

Multilayer network construction and analysis was implemented with the Python programming language using the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [60]. The frequencies and triad significant profiles of motifs are obtained with the FANMOD tool [61].

3. Results

Initially we explore the characterization of all isolated layers with the standard set of network measures (see Methods Section). The results for all eight network layers (for both languages) are in Table 1. The networks are of different sizes, however, the presented measures (L, C and T) are all normalized (by the number of nodes). More important, all networks for one language are constructed from the same text and that gives us the possibility to compare various linguistic realizations of the same source text. Even the datasets for two languages contain approximately the same number of sentences, allowing for comparisons between languages as well. The average path length (L) decrease from co-occurrence to syntax, as expected, but interestingly it is of the same range for the syntax and syllabic layer. The clustering coefficient (C) (obtained from the undirected versions of the same networks) increases on the syllabic subword-level for Croatian and decreases for English. The clustering of English CO word-level are higher than Croatian counterpart. Still, clustering coefficients of SIN layers in both languages are of the same range.

Also, the Croatian syllabic layer has the transitivity (T) higher than the corresponding word layers by one order of magnitude. The numbers of connected components in SYL layers are the highest compared with other layers, and three times higher for English than Croatian. The graphemic layers of both languages exhibit peculiar features due to the small number of vertices, or in other words, due to the high density of GR networks (0.9—HR; 1.03—EN).

3.1. Word-level layers

For the word-level layers we initially examine the distributions. Fig. 2 shows the rank distributions for in- and out-degree of word-level layers in both languages. The exploitation of the same data source per each language caused the high overlap of exposed distributions. Analogously, the in- and out-strength distributions are overlapped as well, for both languages. The power-law γ coefficients for all distributions of word-level layers are in a range between 2.14 and 2.49; thus CO and SIN layers exhibit the power-law distributions for degree and strength regardless of the language.

The potential of selectivity (in- and out-) to differentiate between different text types [62] or the ability to extract keywords [63] (identifying and ranking the most representative features of the source text) is restated in this work for the differentiation of language layers as well. Fig. 3 reveals that the rank distributions of in- and out-selectivity for all word-level layers are apart. Selectivity distributions of co-occurrence (CO) and syntax (SIN) layers are separated for both languages.

The correlation matrices in Fig. 4 show the intra (CO–CO and SIN–SIN) and inter layer (CO–SIN) correlations in terms of in- and out-degree, in- and out-strength, in- and out-selectivity distributions. The correlation values for syntax layers of both languages are lower than the corresponding values for the co-occurrence layers. Notably, the degree and strength correlation values are higher than the selectivity ones, regardless of the language and layer. Furthermore, Croatian is characterized by higher intra and inter layer correlations than English.

In order to obtain a deeper insight into word-level inter layer relationships we calculated the Jaccard overlap percentage. Since the language is modeled with weighted and directed networks, we need to consider weights. Hence, we calculate the Jaccard weighted overlap percentage (W). Then, we calculate the preserved weighted overlap (WO) to quantify the quantity of shared weights on overlapping links between word-level layers (Table 2). For both languages the percentage of preserved overlapped weights is relatively high ($\geq$90%), although slightly lower for English, bearing in mind that less than 20% (W) of the total possible weights on the total intersected links are preserved.
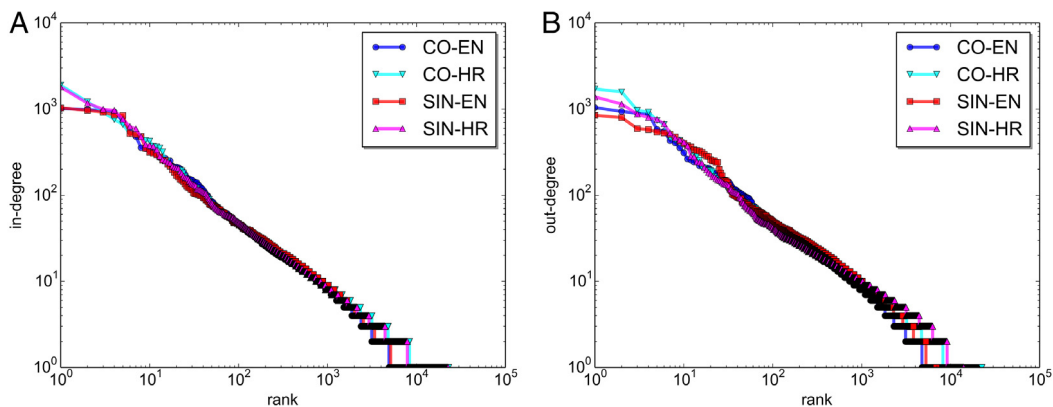
Fig. 2. The word-level layers degree rank distributions. Rank distributions of in- (A) and out- (B) degree for word-level layers: co-occurrence (CO–HR, CO–EN) and syntax (SIN–HR, SIN–EN) on log–log scale.
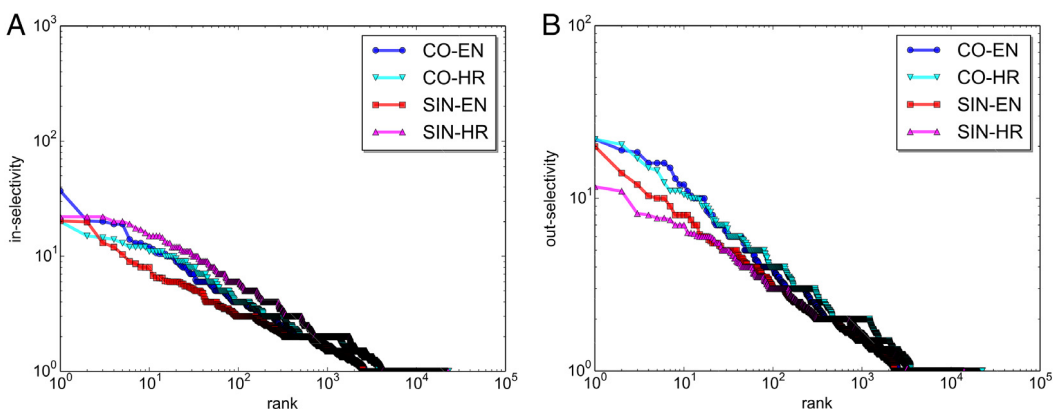

Fig. 3. The word-level layers selectivity rank distributions. Rank distributions of in- (A) and out- (B) selectivity for word-level network layers: co-occurrence (CO–HR, CO–EN) and syntax (SIN–HR, SIN–EN) on log–log scale.
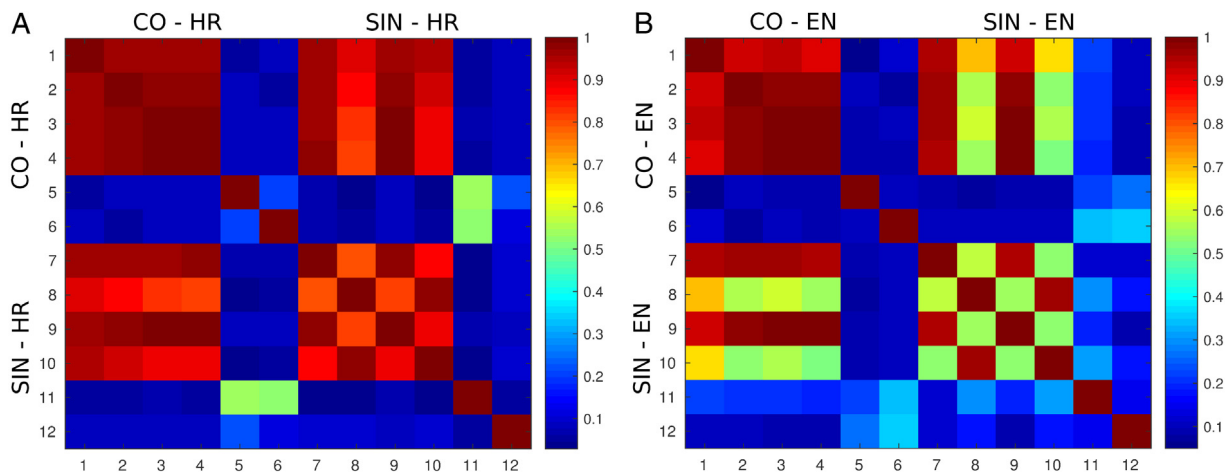

Fig. 4. Intra and inter layers correlations matrices. The correlations matrices for Croatian (A) and English (B): in-(1, 7) & out-(2, 8) degree, in-(3, 9) & out-(4, 10) strength and in-(5, 11) & out-(6, 12) selectivity respectively, presenting inter and intra layer correlations for co-occurrence (CO) and syntax (SIN) word-level layers (all p-values ≤ 0.001).

3.2. Subword-level vs. word-level layers

Subword-level layers syllabic (SYL) and graphemic (GR) in both languages exhibit the power-law γ coefficients between 1.7 and 4.42, which is broader than the observed range of the word-level layers.

Table 2
Overlap of word-level layers.

	Croatian CO–SIN	English CO–SIN
Jaccard	16.72%	13.44%
W	18.96%	13.58%
WO	90.6%	90.00%

The Jaccard overlap percentage, Jaccard weighted overlap percentage (W) and preserved weighted overlap percentage (WO) between CO–SIN word-level layers for Croatian and English.

Table 3
The Pearson correlations of triad frequencies and normalized triad significance profiles.

	Croatian		English	
	Freq.	*TSP*	Freq	*TSP*
CO–SIN	**0.95**	0.42	**0.91**	**0.92**
CO–SYL	**0.96**	−0.03	**0.86**	0.73
CO–GR	−0.18	−0.26	−0.30	0.39
SIN–SYL	**0.95**	**0.83**	**0.99**	**0.91**
SIN–GR	−0.20	−0.21	−0.31	0.28
SYL–GR	−0.21	−0.15	−0.33	0.12

The Pearson correlations of triad frequencies and normalized triad significance profiles (*TSP*) for all pairs of network layers co-occurrence (CO), syntax (SIN), syllables (SYL) and graphemes (GR) for Croatian and English (all p-values ≤ 0.001, emphasized values ≥ 0.8).
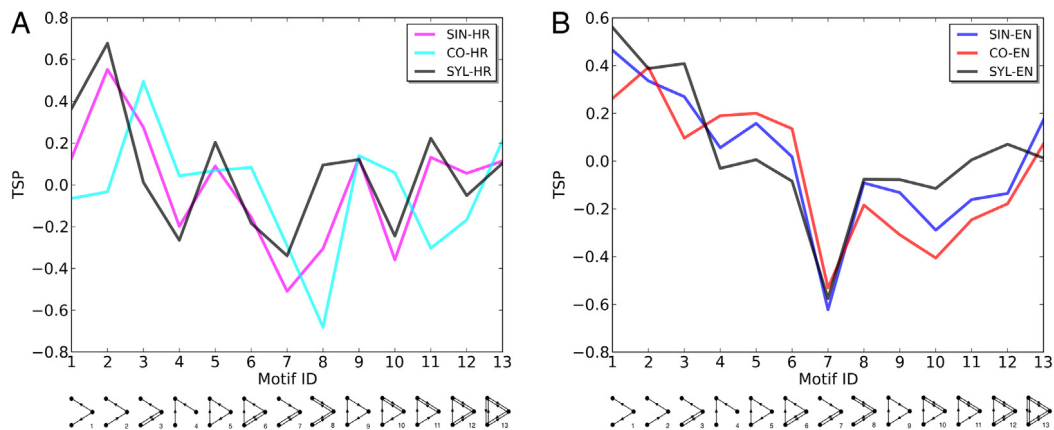

Fig. 5. The normalized triad significance profiles. The normalized triad significance profiles for 13 triadic motifs following the enumeration in Ref. [50] comparing co-occurrence (CO), syntax (SIN) and syllables (SYL) layers for Croatian (A) and English (B).

If we compare the syllabic layers of both languages, it is possible to notice some differences between Croatian and English. English syllables are characterized by distributions closer to the word-level layers distributions (γ coefficients between 1.87 and 2.14). The Croatian syllabic layer distributions reveal some deviations (γ coefficients are lower—between 1.72 and 1.94). The grapheme layers have γ coefficients between 1.7 and 4.16 for Croatian and 2.34 and 4.11 for English.

However, in the multilayer language networks it is interesting to take additional insights of the inter layer relationships, mainly to explore the relationships between word vs. subword layers. For this purpose we introduce the analysis of motifs. We exploited the motif frequencies as well as the normalized triad significance profiles (*TSP*) of all layers for the analysis. The Pearson correlations for all pairs of network layers in Table 3 highlight that motif's frequencies in all layers, with the exception of the graphemic layer are correlated. Correlations of normalized *TSP* indicate that SIN and SYL layers in both languages and additionally for English also CO and SIN layers expose similarities. In order to obtain a deeper insight the normalized significance profiles for CO–SIN–SYL layers of Croatian and English per 13 triadic motifs are compared in Fig. 5.

4. Discussion

The presented findings show that standard network measures on isolated layers exhibit no substantial differences across layers, only slight variations between word and subword-levels. Although, if we compare the structural differences across the examined languages there are indications of different principles in their organization. For instance, English is characterized by higher clustering, with the exception of the syllabic layer. The English syllabic layer has 54 components,

while Croatian has 17, which is reflected in the low clustering coefficient of English syllables. This is caused by high flectivity of Croatian, where many words share the suffix—the last syllable, which decreases the number of components, and increases the clustering coefficient. This observation raises a question, which properties will the morpheme language subsystem expose during the incorporation into a multilayer language framework?

Even a standard distribution analysis is not sufficient to take a deeper insight into the mutual influences between subsystems of language. The (in-/out-) degree and strength distributions of the word-level layers are overlapped due to the same word frequencies reflected from the same data source. Therefore, the standard approach to study the structure of linguistic networks showed no discrepancies among layers. However, the (in-/out-) selectivity values are potentially capable of quantifying differences, namely to show the potential of revealing the interplay among the layers.

In-distributions for syntax layers in both languages have higher values than the corresponding out-distributions, and generally SIN is less inter correlated than the CO layer. The inter and intra layer correlations in the multilayer language network suggest the manifestation of different governing principles in the syntax structure of the examined languages. This is primarily reflection of the rich Croatian morphology. The interesting part is that this is the first observable indication of differences between languages manifested in a multilayer analysis framework, which encouraged a deeper investigation. In addition, the selectivity distributions (regardless of side or layer or language) are not correlated, supporting the potential of selectivity as a measure capable to quantify structural differences across language subsystems. Moreover, Croatian exhibits higher correlations then English in general.

The examination of the word-level layers overlap reveals additional insights into the mutual interplay between the layers. The preserved weighted overlap (*WO*) provides a thorough insight into the intersection between network layers. It expresses the intersected quantity of weights between layers i.e. the shared frequencies of co-occurring units (e.g. bi-grams on the word level). It seems that *WO* is more appropriate to approximate the overlaps of layers in weighted networks than the commonly employed Jaccard measure. The results in Table 2 suggest that Croatian syntax is better captured through words co-occurrences than the English. The preserved weights on intersected links indicate that around 10% of the co-occurrence frequencies are not consistent with overlapped syntax dependencies. The proposed measure of preserved weighted overlap seems adequate to quantify the similarity of word-level layers in weighted and directed multilayer networks of language.

The subword layer's analysis reveals that the syllabic layer plays an important role in the manifestation of principles governing the construction of word layer, which is different for the examined languages. The graphemic layers, on the other hand, share characteristics, which are reflections of the high density of the graphemic networks (almost complete graphs in both languages).

The obtained multilayered language analysis results manifest different driving principles beneath the co-occurrence, syntactic, syllabic and graphemic layers, which was not obvious through the analysis of isolated layers. In order to obtain deeper insight into these relations we utilize the analysis of motifs, which reveal a close topological structure in the syntactic and syllabic layers of both languages. The correlations of the motifs' frequencies are more emphasized in Croatian. The triad significance profiles (*TSP*) are correlated between syntax and syllables regardless of the language, while English additionally exhibits a correlation between co-occurrence and syntax layers. It seems that the observed *TSP* correlations reflect the properties of the Croatian—the free word-order which caused different characterizations of the co-occurrence and syntax layers. Moreover, the high flectivity of Croatian is reflected in many suffixes realized by syllables. Therefore, the structure of layers also reflects the morphological properties inherent to the language, which should we examine more deeply in the future.

Our findings are in line with previous observations in language networks research. For instance, Ferrer i Cancho [32] reports that the amount of syntactically incorrect links in co-occurrence networks can increase to a high of 70%, and elaborates: "About 90% of syntactic relationships take place at a distance lower or equal than two, but word co-occurrence networks lack a linguistically precise definition of link and fail in capturing the characteristic long-distance correlations of words in sentences". This adequately explains the driving principle of the CO–SIN relationships which we have confirmed in this research. Still, an explanation of the linguistic grounding for the SIN–SYL relationships remains an open challenge.

The experimental set-up of this study is in line with the one applied in the empirical study of the Chinese language in Ref. [21], only we extend the multilayer language model to two different languages: Croatian and English. We model four layers of language: two on the word level and two on the subword level, while [21] models three on the word and one on the subword level. Our findings comply with the results reported in Ref. [21], especially regarding the importance of explaining the interplay of linguistic subsystems or as they express "The connection between the global characterization of language as a system and local language phenomena needs to be made more explicit so that the former can shed more light on the latter and vice versa".

To conclude our results strongly suggest that there are some properties which are inherent in the word-level layers and not for the subword layers; while some are inherent in the word-subword relations. More precisely, it seems that syntax and syllables exhibit influences of the same linguistic phenomena.

5. Conclusion

In this research we use the multilayer networks' framework to explore various language subsystems' interactions. Multilayer networks are constructed from four variations of the same original text: two on the word-level (syntax and co-occurrence) and two on the subword-level (syllables and graphemes). The analysis and comparison of layers at word

and subword-levels is employed in order to determine the mechanism of mutual interactions between different linguistic units.

The presented findings corroborate that the multilayer framework can meet the demands in expressing the complex structure of language. According to these results one can notice substantial differences between the networks' structures of different language layers, which are hidden during the exploration of an isolated layer, regardless of modeled language (e.g. Croatian or English). Therefore, it is important to include all language layers simultaneously in order to capture all language characteristics in the systematic exploration.

The multilayer network framework is a powerful, consistent and systematic approach to model several linguistic subsystems simultaneously and to provide a more general view on language and at the same time enable deeper insights into inter-layer relationships. The word-level layers can be represented as multiplex networks (the coupled links have 1:1 or 0:1 inter-connections), while the connections between word and subword layers are not coupled (have N:M inter-connections). Hence, defining the unified theoretical model for the multilayer language networks is essential for further endeavors in the research of linguistic networks. It is worth noticing that presented multilayer language network model can be extended with new aspects (e.g. text types, genres) and new layers (e.g. semantic, morphemes).

These findings reveal a variety of research questions which will motivate our future work in network linguistics. Hence, our future research plans involve: exploring the relationships of other languages' subsystems (i.e. morphological, phonetic, semantic), examining new languages, defining the theoretical model capable of capturing all structural variations of language subsystems' relationships and eventually explain the governing principle of mutual interactions and conceptual universalities in natural languages.

Acknowledgment

References

[1] M. Barigozzi, G. Fagiolo, D. Garlaschelli, Multinetwork of international trade: A commodity-specific analysis, Phys. Rev. E 81 (2010) 046104.
[2] M. Szell, R. Lambiotte, S. Thurner, Multirelational organization of large-scale social networks in an online world, PNAS 107 (2010) 13636–13641.
[3] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, D. Pedreschi, Foundations of multidimensional network analysis, in: Advances in Social Networks Analysis and Mining (ASONAM), 2011, pp. 485–489. http://dx.doi.org/10.1109/ASONAM.2011.103.
[4] M. Kurant, P. Thiran, Layered complex networks, Phys. Rev. Lett. 96 (2006) 138701–138705.
[5] R. Morris, M. Barthelemy, Transport on coupled spatial networks, Phys. Rev. Lett. 109 (2012) 128703.
[6] E. Estrada, J. Gómez-Gardeñes, Communicability reveals a transition to coordinated behavior in multiplex networks, Phys. Rev. E 89 (2014) 042819.
[7] A. Cardillo, J. Gómez-Gardeñes, M. Zanin, M. Romance, D. Papo, F. del Pozo, S. Boccaletti, Emergence of network features from multiplexity, Sci. Rep. 3 (2013) 1344.
[8] E. Bullmore, O. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems, Nat. Rev. Neurosci. 10 (2009) 186–198.
[9] M.D. Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. Porter, S. Gómez, A. Arenas, Mathematical formulation of multilayer networks, Phys. Rev. X 3 (2013) 041022–041037.
[10] G. Menichetti, D. Remondini, P. Panzarasa, R. Mondragón, G. Bianconi, Weighted multiplex networks, PLoS One 9 (2014) e97857.
[11] J. Gao, S. Buldyrev, H. Stanley, S. Havlin, Networks formed from interdependent networks, Nat. Phys. 8 (2012) 40–48.
[12] J. Gao, D. Li, S. Havlin, From a single network to a network of networks, Natl. Sci. Rev. (2014).
[13] G. Bianconi, S. Dorogovtsev, J. Mendes, Mutually connected component of network of networks, Phys. Rev. E 91 (2015) 012804.
[14] M. Kivelä, A. Arenas, M. Barthelemy, J. Gleeson, Y. Moreno, M. Porter, Multilayer networks, J. Complex Netw. 2 (2014) 203–271.
[15] S. Boccaletti, G. Bianconi, R. Criado, C. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadalj, Z. Wangk, M. Zaninm, The structure and dynamics of multilayer networks, Phys. Rep. 544 (2014) 1–122.
[16] N. Chomsky, Aspects of the Theory of Syntax, MIT press, 1988.
[17] R. Jackendoff, What is the human language faculty?: Two views, Language 87 (2011) 586–624.
[18] H. Brighton, K. Smith, S. Kirby, Language as an evolutionary system, Phys. Life Rev. 2 (2005) 177–226.
[19] L. Steels, Modeling the cultural evolution of language, Phys. Life Rev. 8 (2011) 339–356.
[20] G. Hickok, The functional neuroanatomy of language, Phys. Life Rev. 6 (2009) 121–143.
[21] H. Liu, J. Cong, Empirical characterization of modern Chinese as a multi-level system from the complex network approach, J. Chin. Linguist. 42 (2014) 1–38.
[22] G. Fenk-Oczlon, A. Fenk, Complexity trade-offs between the subsystems of language, in: Language Complexity: Typology, Contact, Change, John Benjamins Publishing, 2008, pp. 43–65.
[23] S. Dorogovtsev, J. Mendes, Language as an evolving word web, Proc. R. Soc. Lond. B Biol. Sci. 268 (2001) 2603–2606.
[24] A. Masucci, G. Rodgers, Network properties of written human language, Phys. Rev. E 74 (2006) 026102.
[25] M. Choudhury, A. Mukherjee, The structure and dynamics of linguistic networks, in: Dynamics On and of Complex Networks, Modeling and Simulation in Science, Engineering and Technology, Birkhäuser, Boston, 2009, pp. 145–166. http://dx.doi.org/10.1007/978-0-8176-4751-3_9.
[26] R. Solé, B. Corominas-Murtra, S. Valverde, L. Steels, Language networks: their structure, function, and evolution, Complexity 15 (2010) 20–26.
[27] J. Cong, H. Liu, Approaching human language with complex networks, Phys. Life Rev. 11 (2014) 598–618.
[28] D. Margan, S. Martinčić-Ipšić, A. Meštrović, Preliminary report on the structure of Croatian linguistic co-occurrence networks, in: Proc. of the 5th International Conference on Information Technologies and Information Society 2013, 2013, pp. 89–96.
[29] H. Liu, J. Cong, Language clustering with word co-occurrence networks based on parallel texts, Chinese Sci. Bull. 58 (2013) 1139–1144.
[30] R. Ferrer i Cancho, Patterns in syntactic dependency networks, Phys. Rev. E 69 (2004) 51915.
[31] R. Ferrer i Cancho, A. Mehler, O. Pustylnikov, A. Diaz-Guilera, Correlations in the organization of large-scale syntactic dependency networks, in: Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing, Association for Computational Linguistics, Rochester, NY, USA, 2007, pp. 65–72.
[32] R. Ferrer i Cancho, The structure of syntactic dependency networks: Insights from recent advances in network theory, in: The Problems of Quantitative Linguistics, in: Studies in Computational Intelligence, 2005, pp. 60–75.
[33] H. Liu, F. Hu, What role does syntax play in a language network? Europhys. Lett. 83 (2008) 18002.
[34] H. Liu, X. Chunshan, Can syntactic networks indicate morphological complexity of a language? Europhys. Lett. 93 (2011) 28005.
[35] O. Abramov, A. Mehler, Automatic language classification by means of syntactic dependency networks, J. Quant. Linguist. 18 (2011) 291–336.

[36] D.R. Amancio, M.G. Nunes, O.N.O. Jr., L.da F. Costa, Extractive summarization using complex networks and syntactic dependency, Physica A 391 (2012) 1855–1864.
[37] G. Peng, J.W. Minett, W.S.-Y. Wang, The networks of syllables and characters in Chinese, J. Quant. Linguist. 15 (2008) 243–255.
[38] M. Medeiros Soares, G. Corso, G. Lucena, The network of syllables in Portuguese, Physica A 355 (2005) 678–684.
[39] S. Arbesman, S. Strogatz, M. Vitevitch, The structure of phonological networks across multiple languages, Int. J. Bifurcation Chaos 20 (2010) 679–685.
[40] J. Feldman, Embodied language, best-fit analysis, and formal compositionality, Phys. Life Rev. 7 (2010) 385–410.
[41] C. Lyon, C. Nehaniv, J. Saunders, Interactive language learning by robots: the transition from babbling to word forms, PLoS One 7 (2012) e38236.
[42] P. Oudeyer, The origins of syllable systems: an operational model, in: Proceedings of the 3rd ICCS International Conference on Cognitive Science, COGSCI 2001, 2001, pp. 27–31.
[43] C. Kello, B. Beltz, Scale-free networks in phonological and orthographic wordform lexicons, in: Approaches to Phonological Complexity, Walter de Gruyter, 2009, pp. 171–190.
[44] M. Vitevitch, What can graph theory tell us about word learning and lexical retrieval? J. Speech Lang. Hear. Res. 51 (2008) 408–422.
[45] T. Gruenenfelder, D. Pisoni, The lexical restructuring hypothesis and graph theoretic analyses of networks based on random lexicons, J. Speech Lang. Hear. Res. 52 (2009) 596–609.
[46] A. Mukherjee, M. Choudhury, A. Basu, N. Ganguly, Self-organization of the sound inventories: analysis and synthesis of the occurrence and co-occurrence networks of consonants, J. Quant. Linguist. 16 (2009) 157–184.
[47] S. Arbesman, S. Strogatz, M. Vitevitch, Comparative analysis of networks of phonologically similar words in English and Spanish, Entropy 12 (2010) 327–333.
[48] M. Vitevitch, K. Chan, R. Goldstein, Insights into failed lexical retrieval from network science, Cogn. Psychol. 68 (2014) 1–32.
[49] R. Milo, S. Shen-Orr, S. Itzkovit, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (2002) 824–827.
[50] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, U. Alon, Superfamilies of evolved and designed networks, Science 303 (2004) 1538–1542.
[51] C. Biemann, S. Roos, K. Weihe, Quantifying semantics using complex network analysis, in: Proceedings of the 24th International Conference on Computational Linguistics-COLING 2012, 2012, pp. 263–278.
[52] M. Newman, Networks: An Introduction, Oxford University Press, 2010.
[53] T. Ban Kirigin, A. Meštrović, S. Martinčić-Ipšić, Towards a formal model of language networks, in: G. Dregvaite, R. Damasevicius (Eds.), Information and Software Technologies, in: Communications in Computer and Information Science, vol. 538, Springer International Publishing, 2015, pp. 469–479. http://dx.doi.org/10.1007/978-3-319-24770-0_40.
[54] HOBS, Croatian corpus, ver. 1.0, University of Zagreb, Faculty of Humanities and Social Sciences, 2013. (accessed [2015-01-20]).
[55] M. Tadić, Building the Croatian dependency treebank: the initial stages, Suvrem. Lingvist. 63 (2007) 85–92.
[56] M. Marcus, M. Marcinkiewicz, B. Santorini, Building a large annotated corpus of English: the Penn treebank, Comput. Linguist. 19 (1993) 313–330.
[57] Penn Treebank, English corpus, University of Pennsylvania, Computer and Information Science Department, 1995. http://www.nltk.org/howto/corpus.html/parsed-corpora (accessed [2015-03-25]).
[58] A. Meštrović, S. Martinčić-Ipšić, M. Matešić, Syllabification based on maximal onset principle for Croatian/Postupak automatskoga slogovanja temeljem načela najvećega pristupa i statistika slogova za hrvatski jezik, Govor Speech 32 (2015) 3–34.
[59] C. Barker, Beep pronouncing dictionary—syllabified, 2002. http://semarch.linguistics.fas.nyu.edu/barker/Syllables/index.txt (accessed [2014-03-25]).
[60] A. Hagberg, D. Schult, P. Swart, Exploring network structure, dynamics, and function using NetworkX, in: Proceedings of the 7th Python in Science Conference - SciPy 2008, 2008, pp. 11–15. URL: https://networkx.github.io/.
[61] S. Wernicke, A faster algorithm for detecting network motifs, in: Algorithms in Bioinformatics, in: Lecture Notes in Computer Science, vol. 3692, Springer, Berlin, Heidelberg, 2005, pp. 165–177. http://dx.doi.org/10.1007/11557067_14.
[62] S. Šišović, S. Martinčić-Ipšić, A. Meštrović, Comparison of the language networks from literature and blogs, in: Proceedings of the 37th International Convention on Microelectronics- MIPRO 2014, 2014, pp. 1824–1829. http://dx.doi.org/10.1109/MIPRO.2014.6859821.
[63] S. Beliga, A. Meštrović, S. Martinčić-Ipšić, Toward selectivity based keyword extraction for Croatian news, in: Proceedings of the Workshop on Surfacing the Deep and the Social Web-ISWC 2014:, 2014, pp. 1–14.