langnet

Department of Informatics, University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
Tel.: + 385 51 584 700 Fax: + 385 51 584 749
www.langnet.uniri.hr

An Overview of Language Networks: Case of Croatian

Sanda Martinčić-Ipšić
smarti@uniri.hr
Ana Meštrović
Domagoj Margan
Slobodan Beliga
Hana Rizvić
Sabina Šišović

ITIS 2014 – November 5-7 in Šmarješke toplice, Slovenia

Language

- main tool of **communication**
- **reflects** our history and culture
- **evolving** in parallel with our society
  - can be seen as a **complex adaptive system**

- written (as well as spoken) language can be modeled via **complex networks**
  - the lingual units (words) are represented by **vertices** and their linguistic interactions by **links**
  - allows systematic quantitative analyses

Language networks

langnet

- model the various **language subsystems** (levels)
  - examine unique **function** through complex networks
  - examine various linguistic units
- deepening the understanding of conceptual **similarities, differences and universalities** in natural languages
  - cognitive representation of the language in the human brain

- establish a bridge:
  - linguistics, complex networks science, computer science and natural language processing

Language networks – levels

langnet

- various language **subsystems** – represented as complex networks
  - vertices – linguistic units
  - links – model their relationships

- **word level:**
  - co-occurance
  - syntax
  - semantics
  - pragmatics

- **sub-word level:**
  - morphology (morphosyntactic)
  - syllabic
  - phonetic (phonology)
  - graphemic

- present: **focus on isolated linguistic subsystems**
  - lacking to explain (or even explore) the mechanism of their mutual interaction, interplay or inheritance

Croatian language networks

langnet

- model the phenomena of various Croatian language subsystems and examine their functions through complex networks
- relying on the well-established principles by modeling **interactions of linguistic units at each and across levels**
  - word order in a sentence, co-occurrence, syntax...
  - sub word units, ...

- **goal:** systematic investigation of Croatian language networks
  - Croatian limited resources and technologies (CESAR – METANET)

Croatian at glance

- a highly flective Slavic language
  - 14 different cases
  - 7 for singular and
  - 7 for plural
- three genders and two numbers
- mostly free word order
- morphologically rich language
  - Tadić: up to 614 different word forms
    - 404 - nouns
    - 155 - adjectives
    - …

NOUN:
kuć-a
kuć-e
kuć-i
kuć-om
kuć-o
kuć-ama
s-kuć-iti
kuć-ica
…

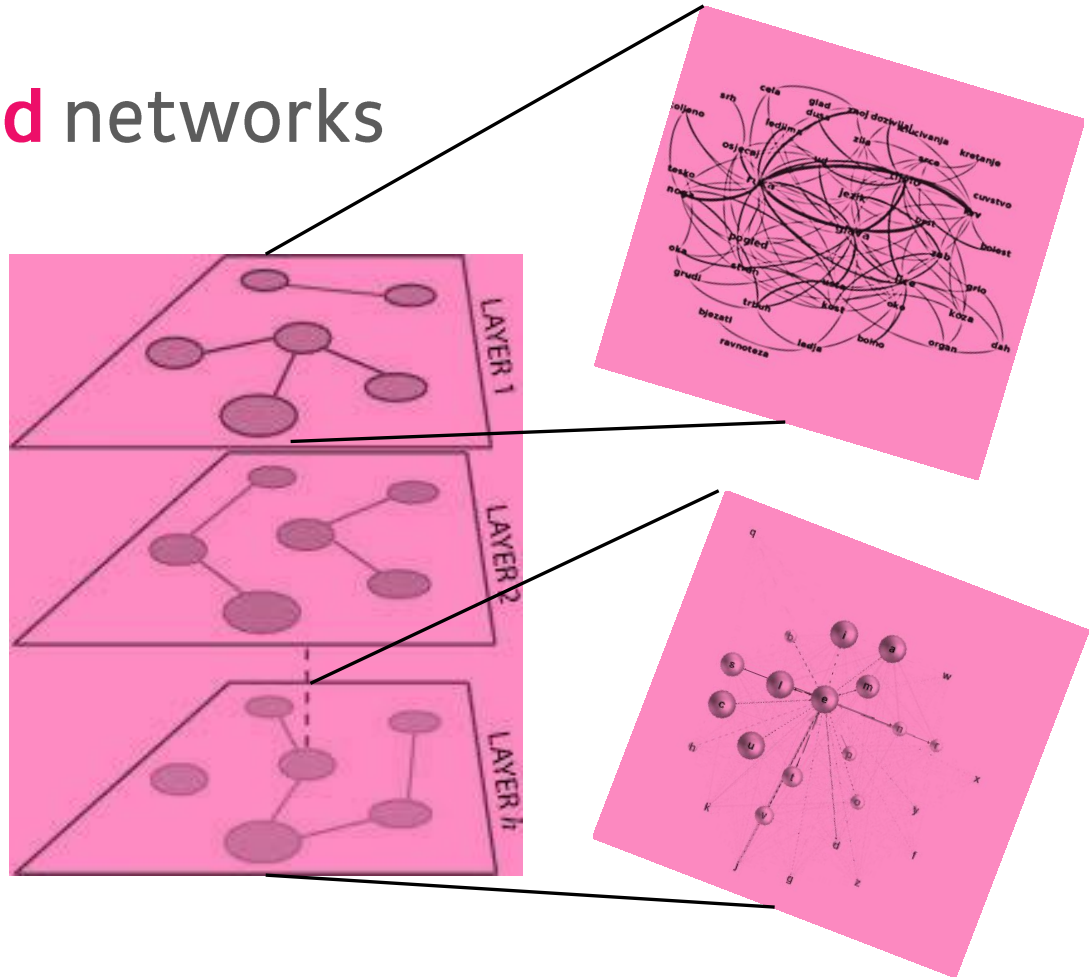
VERB:
gleda-ti
gleda-m
gleda-š
gleda-amo
gleda-te
gleda-ju
gleda-smo
gleda-še
gleda-hu
o-gleda-ti
…

Outline

- experimental results
- Croatian **multilayered** networks

- **word** level
  - co-occurrence
  - shuffle
  - syntax

- **sub-word** level
  - syllables
  - graphemes

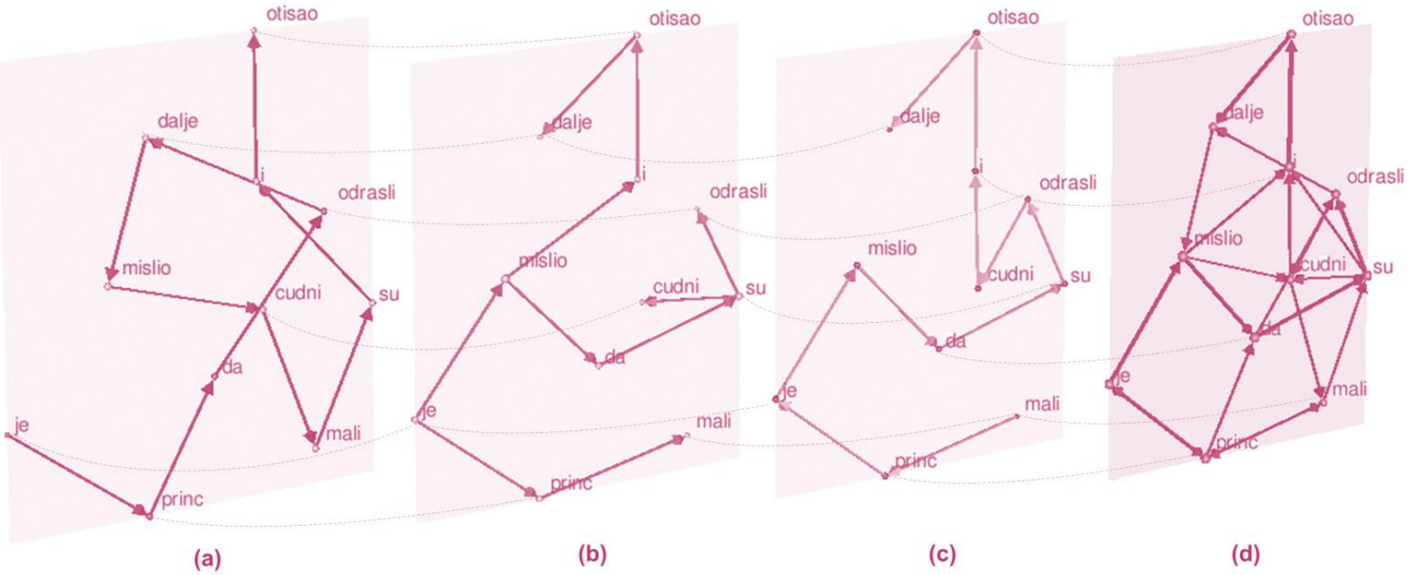


langnet

Department of Informatics, University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
Tel.: + 385 51 584 700 Fax: + 385 51 584 749
www.langnet.uniri.hr

Experiment: Multilayered langauge networks

word level

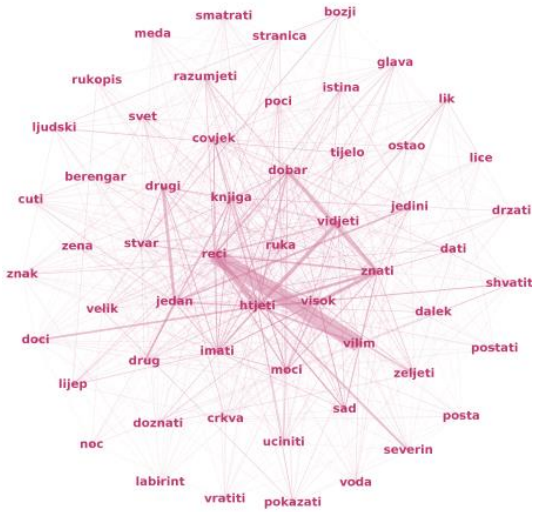


shuffled | syntax | co-occurrence | aggregated

Word-level networks

langnet

- **co-occurrence networks**
  - **directed** or undirected [ITIS 2013a]
  - **weighted** or unweighted [ITIS 2013a]
  - **stopwords** preserved [ITIS 2013a, MIPRO2014a]
  - **not lemmatized**
    - in the full variety of flective word forms

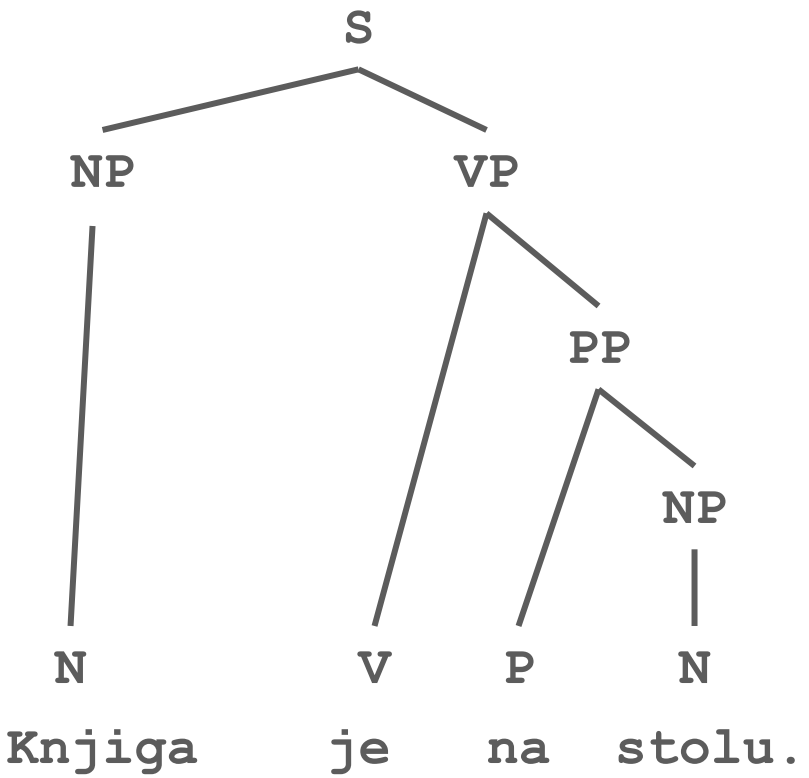


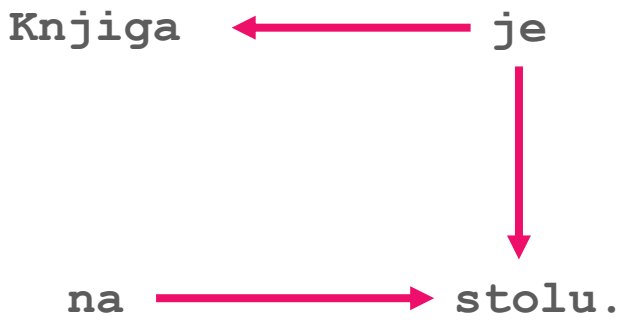
- size of the **co-occurrence window: 2** [ITIS 2013a]
- within boundaries: **words and sentences** [ITIS 2013a, CompleNet 2014]

- sensitive to used **corpus**

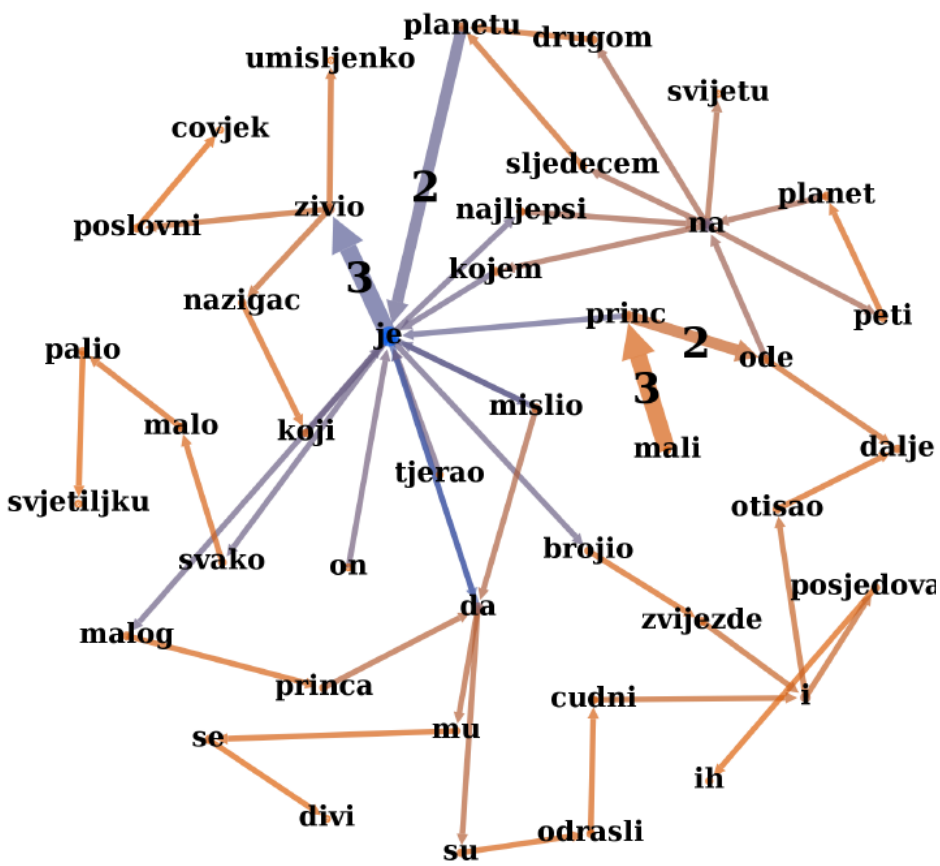
Syntax Dataset

- **Croatian Dependency Treebank** (Agić at al.)
  - parsed syntax tree
  - 3.465 sentences (88.045 tokens)

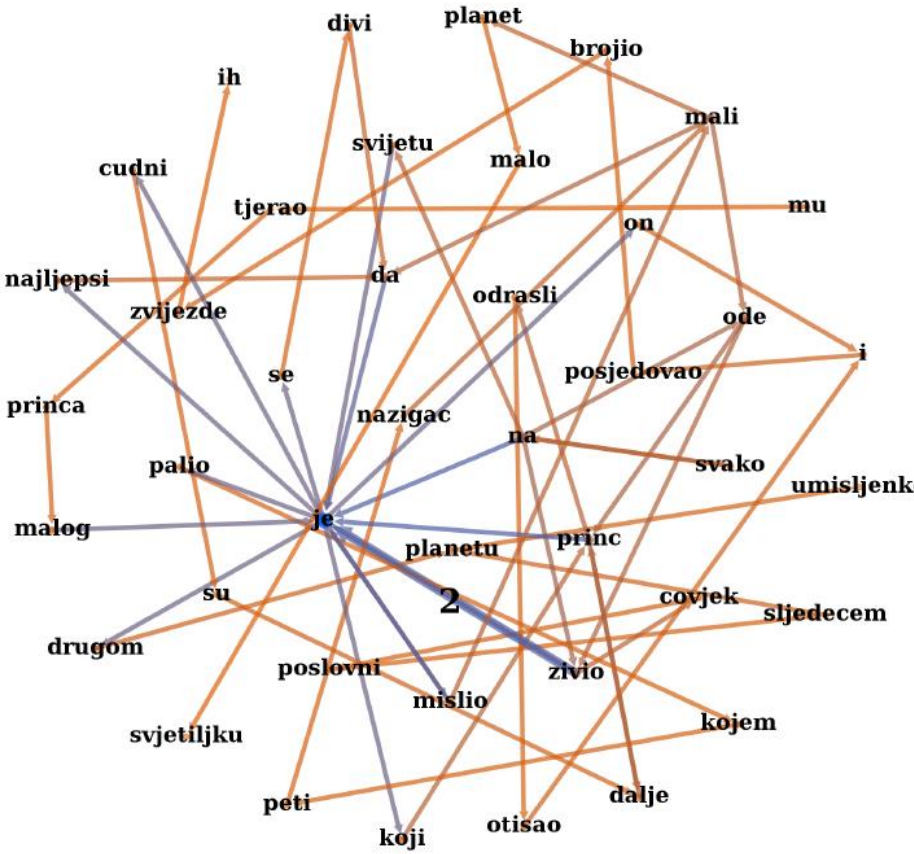




Sentence-level shuffling



mali princ je mislio da su odrasli cudni i otisao dalje. na drugom planetu je zivio umisljer
mislio je da je najljepsi na svijetu. tjerao je malog princa da mu se divi. mali princ ode da
na sljedecem planetu je zivio poslovni covjek. on je brojio zvijezde i posjedovao ih. mali
princ ode na peti planet na kojem je zivio nazigac koji je svako malo palio svjetiljku.



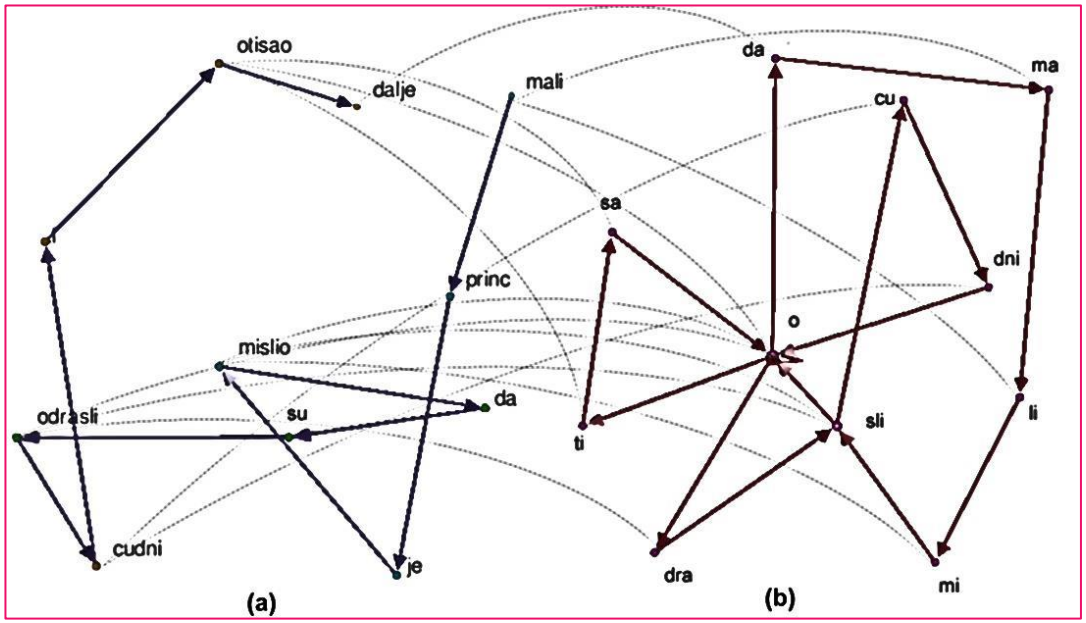
mislio mali da je cudni su dalje princ odrasli otisao i. na zivio je drugom planetu
umisljenko. na svijetu je mislio je najljepsi da. mu tjerao princa malog je se divi da. mali o
princ dalje. na je zivio covjek poslovni sljedecem planetu. je on i posjedovao brojio zvijezd
ih. na svako na ode zivio je koji princ je palio kojem peti nazigac mali planet malo svjetiljk

[CompleNet 2014, MIPRO2014a]

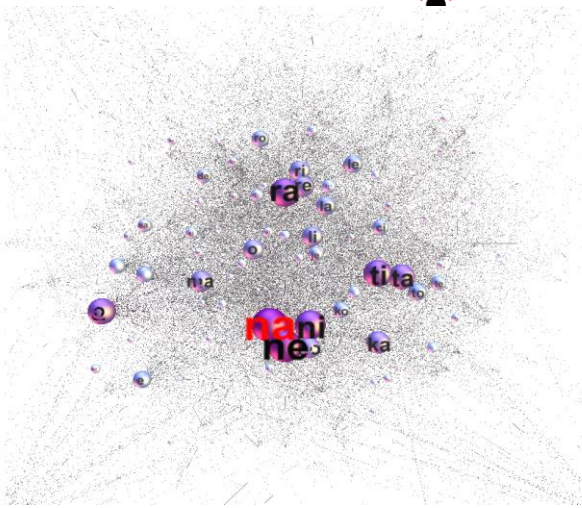
langnet

Department of Informatics, University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
Tel.: + 385 51 584 700 Fax: + 385 51 584 749
www.langnet.uniri.hr

Sub-word level networks

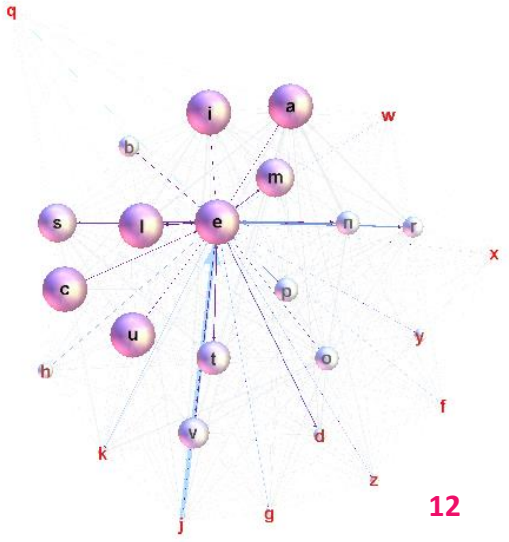


Subword-level networks



- **syllables** network [MIPRO2013]
  - syllables that co-occur in the same word
    - also syllables across words – toward speech
  - Croatian has two possible syllabifications
  - phonological and phonetic
    - **phonological** syllabification: our algorithm
    - **phonetic** syllabification: our grapheme-to-phoneme method

- **graphemes** network
  - graphemes that co-occur in the same word



Experiment

langnet

- 5 networks: directed and weighted
  - not lemmatized, stopwords included
- word level: sentence boundaries
  - **co-occurrance** – window size 2
  - **shuffle**
  - **syntax**
- subword level: words boundaries
  - **syllables** from words in original sentences
  - **graphemes** from words in original sentences

- same dataset: Croatian Dependency Treebank

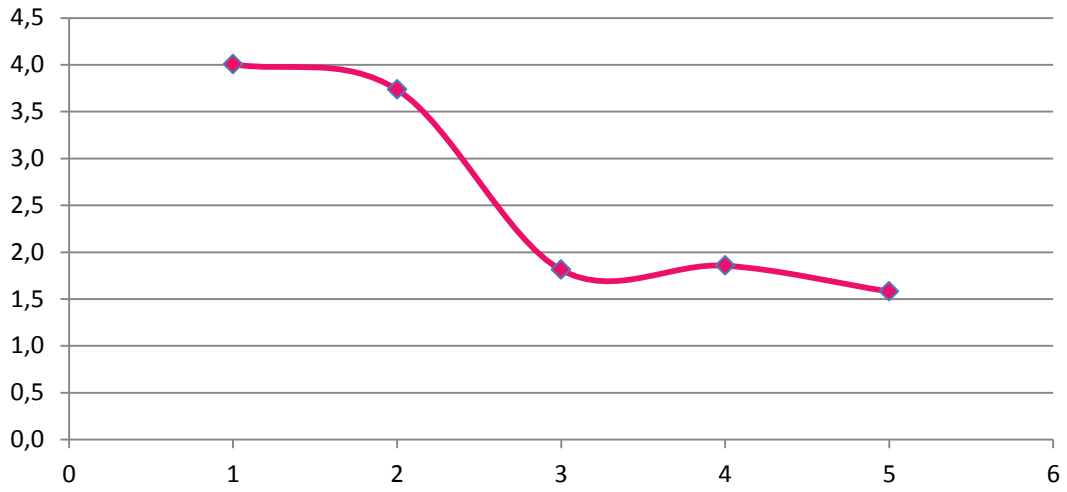
Results

langnet

	Original co-occurr	Shuffled	Syntax	Syllables	Graphemes
Number of nodes (N)	23359	23359	23359	2634	34
Number of edges (K)	71860	86214	70155	18849	491
Number of components	2	2	2	17	1
Average path length (L)	4,01	3,74	1,81	1,86	1,58
Diameter (D)	16	17	12	8	3
Average clustering coefficient(C)	0,17	0,19	0,12	0,26	0,64
Transitivity	0,004	0,013	0,003	0,120	0,522
Density	0,00013	0,00016	0,00013	0,00272	0,43761

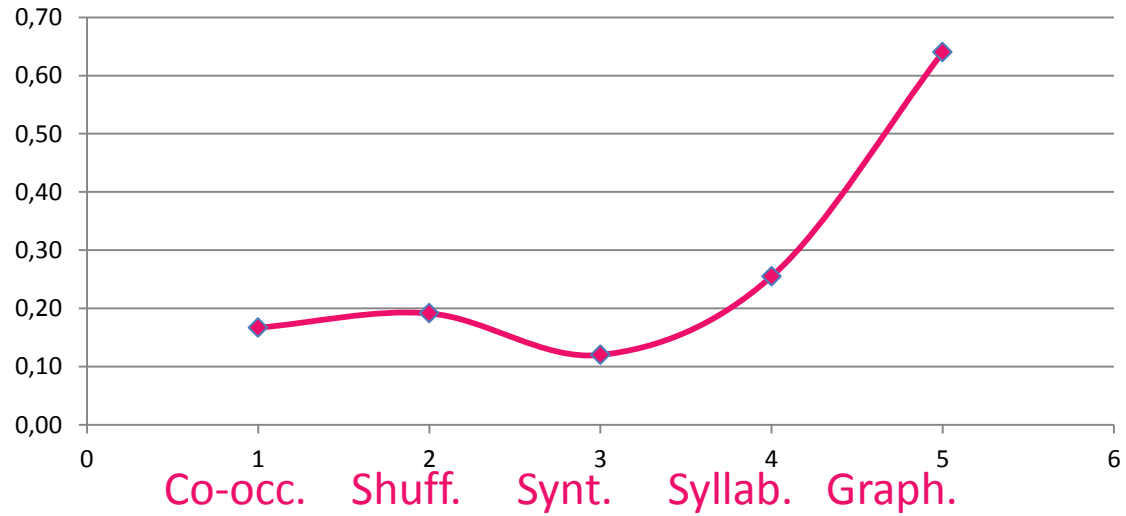
- avg. path length - degree of separation between linguistic units
- diameter – maximal separation
- density – probability of connecting 2 units
- transitivity – realized number of triangls (among possible ones)

Average path length (L)



Co-occ. Shuff. Synt. Syllab. Graph.

Average clustering coefficient (C)



Co-occ. Shuff. Synt. Syllab. Graph.

- **density**
- SYLL vs SYN 21x
- GRA vs SYN 161x

Word-level overlap

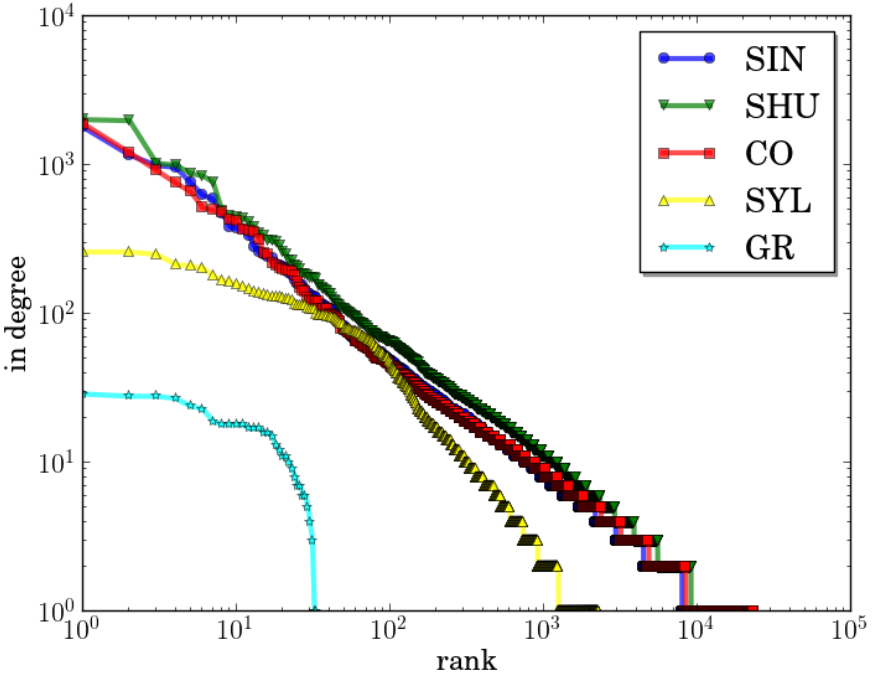
langnet

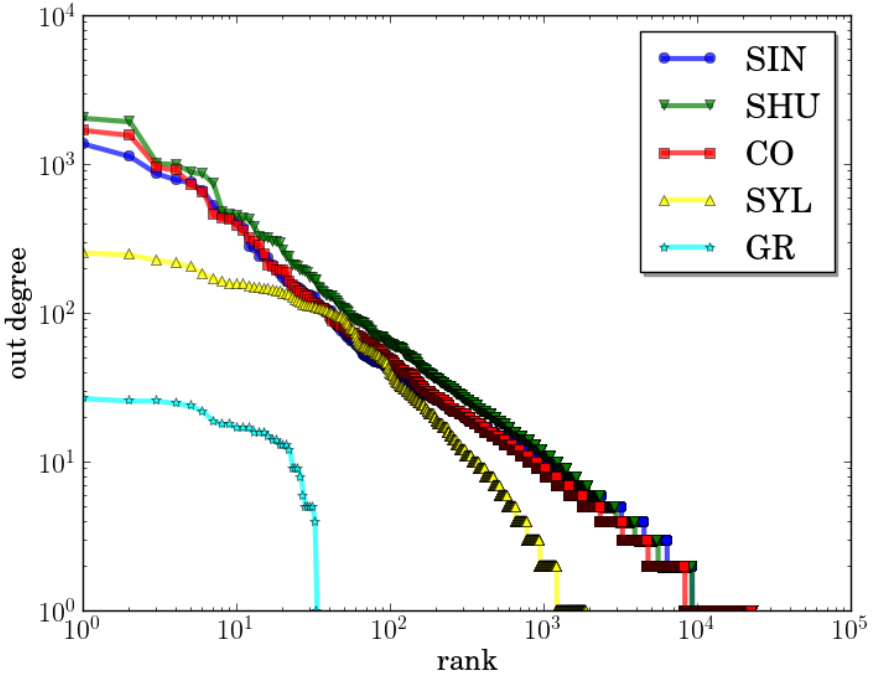
- **Jaccard overlap**
  - between two network layers α and α':

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

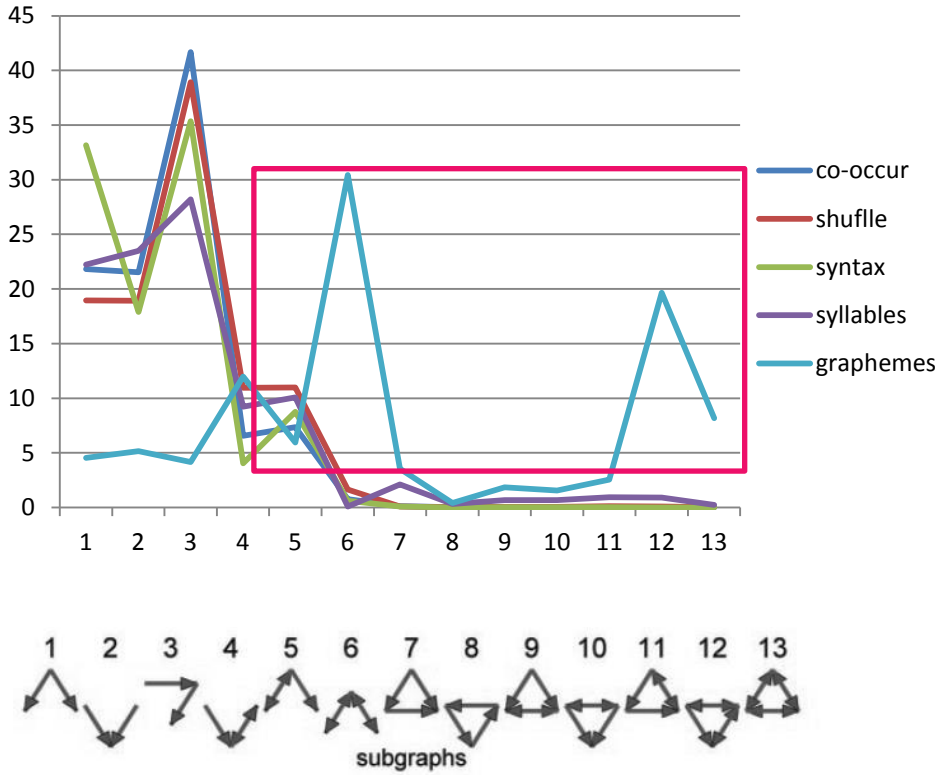
- **Co-ocurr – Syntax:** **16.72 %**
- **Co-ocurr – Shuffle:** **5.47 %**
- **Syntax – Shuffle:** **4.81 %**

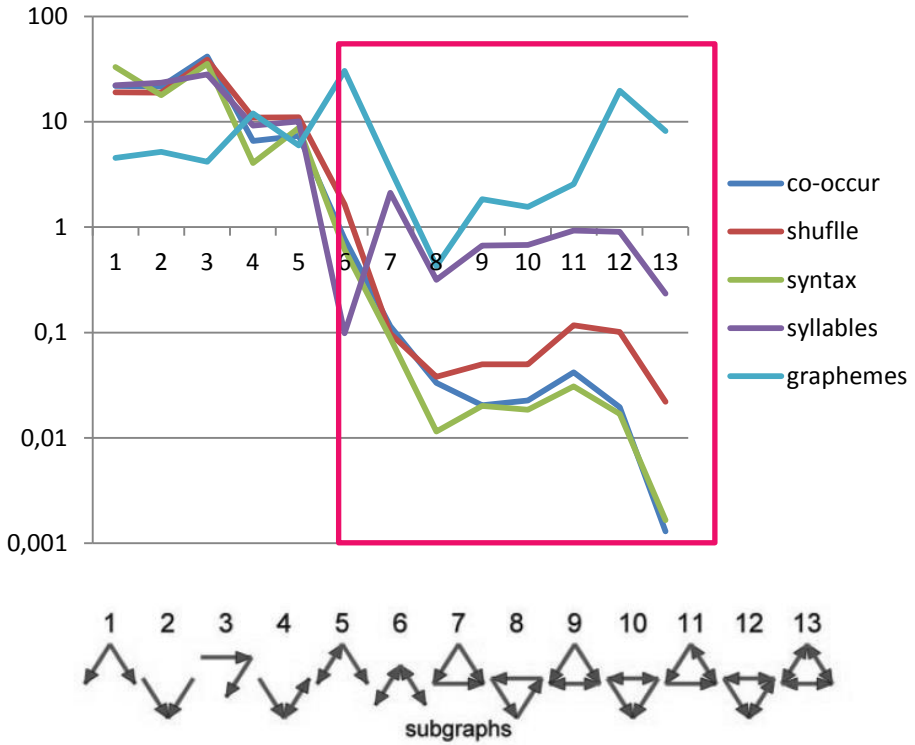
Degree distributions



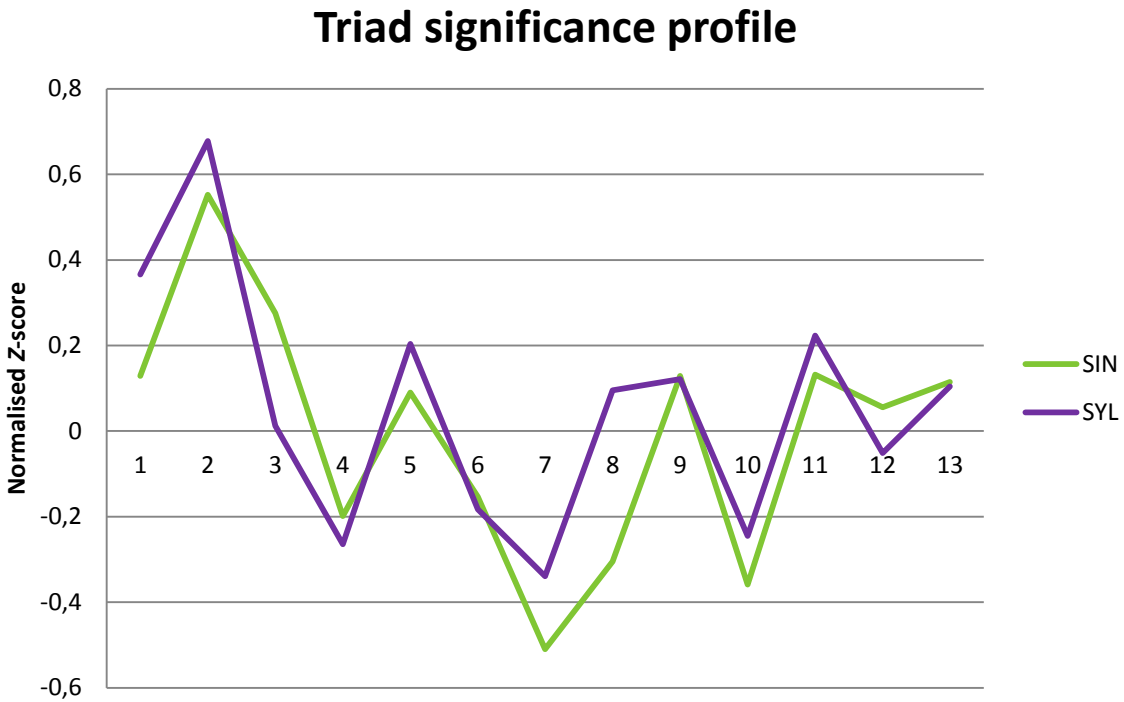


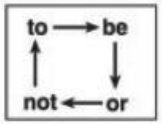
Motifs results

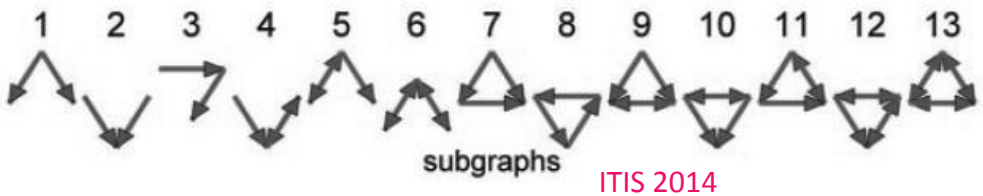




Triad significance profile







Milo at al. 2004

Recap.

langnet

- **co-occurance:** traditional, not sufficient
- **shuffled:** reveals interesting behavior
- **syntax:** more credible for linguistic insights
- **syllables:** like syntax
  - syllables like morphological root
  - morphological networks should be constructed
- **graphemes:** completely different (complex network??)
- **motifs:** preliminary results for Croatian different from other languages
  - morphologically rich – highly flective
  - free word order

Open questions?

- test on **other** data sets and languages
- still **need** comments from linguists
- overall language network model
  - Formalization of model
    - which formalism? – multiplex is not sufficient?
  - anything …. everybody ☺

LangNet plan

- **across language levels:**
  - **subword level**: phonemes, syllables, morphemes
  - **word level**: words co-occurrence, syntax dependencies
- **across languages:**
  - comparative analysis with **other languages** English, Italian, Slovene,...
- **interaction across language levels:**
  - how different language **subsystems mutually interact**
- **text quality evaluation:**
  - derive an **assessment model** for the evaluation of the quality of Croatian texts from complex networks parameters
  - creativity? cognitive represantations?
  - keywords extraction, summarization

resources

methodology

applications

langnet

Department of Informatics, University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
Tel.: + 385 51 584 700 Fax: + 385 51 584 749
www.langnet.uniri.hr

An Overview of Language Networks: Case of Croatian

Sanda Martinčić-Ipšić
smarti@uniri.hr
Ana Meštrović
Domagoj Margan
Slobodan Beliga
Hana Rizvić
Sabina Šišović

ITIS 2014 – November 5-7 in Šmarješke toplice, Slovenia

Sanda Martinčić-Ipšić
smarti@uniri.hr

An Overview of Language Networks: Case of Croatian

KEEP
CALM
AND ASK
A QUESTION
?