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

• 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

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

• word level:
  – co-occurrence
  – 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

• 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
- ...

ITIS 2014
Outline

- experimental results
- Croatian **multilayered** networks

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

  - **sub-word** level
    - syllables
    - graphemes
Experiment: Multilayered language networks

word level

(a) shuffled  (b) syntax  (c) co-occurrence  (d) aggregated
Word-level networks

• 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)

Knjiga je na stolu.

CC – BY – NC – SA

ITIS 2014
mali princ je mislio da su odrasli cudni i otisao dalje, na drugom planetu je zivio umislenko mislio je da je najljepši na svijetu. tjerao je malog princa da mu se divi. mali princ ode da na sljedećem 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.

mislilo mali da je cudni su dalje princ odrasli otisao i. na zivio je drugom planetu umislenko, na svijetu je mislio je najljepši da. mu tjerao princa malog je se divi da. mali o princ dalje. na je zivio covjek poslovni sljedećem planetu, je on i posjedovao brojio zvijezdih. na svako na ode zivio je koji princa je palio kojem peti nazigac mali planet malo svjetiljki.
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

- 5 networks: directed and weighted
  - not lemmatized, stopwords included
- Word level: sentence boundaries
  - co-occurrence – 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

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

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

Average clustering coefficient (C)

- density
- SYLL vs SYN 21x
- GRA vs SYN 161x
Word-level overlap

- **Jaccard overlap**
  - between two network layers $\alpha$ and $\alpha'$:

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

- Co-occurr – Syntax: 16.72%
- Co-occurr – Shuffle: 5.47%
- Syntax – Shuffle: 4.81%

ITIS 2014
Degree distributions

![Graphs showing degree distributions for different ranks, with logarithmic scales for in-degree and out-degree. The graph on the left shows in-degree distribution for SIN, SHU, CO, SYL, and GR, while the graph on the right shows out-degree distribution for the same categories.]
Motifs results

ITIS 2014
Triad significance profile

Milo at al. 2004
Recap.

• **co-occurrence**: 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 representations?
  • keywords extraction, summarization
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
An Overview of Language Networks: Case of Croatian

KEEP CALM

AND ASK A QUESTION?