Network Motifs Analysis of Croatian Literature

Ana Meštrović
amestrovic@uniri.hr
Sanda Martinčić-Ipšić
smarti@uniri.hr
Hana Rizvić
hrizvic@uniri.hr

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

- Network motifs are used for the network analysis on the meso-scale level
- Network motifs in language networks:
  - quantify the differences between a natural and a generated language [Biemann at al. 2012]
  - may reveal structural properties of the language
- Our motivation:
  - to determine whether the local structure of the Croatian language networks share the same properties as other language networks
Network motifs

- significantly overrepresented connected and directed subgraphs in the graph (network)
- may contain up to 8 vertices (only 3-vertex and 4-vertex motifs analyzed so far)

13 types of 3-vertex connected subgraphs
199 types of 4-vertex connected subgraphs
Network motifs

• The **Z-score** of the subgraph $H$ in graph $G$:

$$Z(H) = \frac{F_G(H) - \mu_R(H)}{\sigma_R(H)}$$

• The higher the Z-score is, the more significant motif

• The **significance profile** (SP) is the vector of Z-scores normalised to length 1:

$$SP_i = \frac{Z_i}{\sqrt{\sum Z_i^2}}$$
Dataset

- 4 books and 1 forum
  - *Mama Leone (ML), The Return of Philip Latinowicz (PL), The Picture of Dorian Gray, (DG), Bones, (BO) and forum Narodne novine*
  - we constructed 5 directed co-occurrence networks

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of words</th>
<th>Number of vertices (N)</th>
<th>Number of edges (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>86,043</td>
<td>12,416</td>
<td>52,012</td>
</tr>
<tr>
<td>PL</td>
<td>28,301</td>
<td>9,166</td>
<td>22,344</td>
</tr>
<tr>
<td>DG</td>
<td>75,142</td>
<td>14,120</td>
<td>47,823</td>
</tr>
<tr>
<td>BO</td>
<td>199,188</td>
<td>25,020</td>
<td>106,999</td>
</tr>
<tr>
<td>NN</td>
<td>146,731</td>
<td>13,036</td>
<td>55,661</td>
</tr>
</tbody>
</table>
Network motif analysis

- we analyse 3-vertex subgraphs (triads) in language networks
- FANMOD tool:
  - results in terms of: Z-scores, $p$-values and frequencies

Experiment -analyzed:
- the frequencies of the triads
- triad significance profile
Results I

- The frequencies of the triads for 5 datasets

![Graph showing triad frequencies for IDs 1, 5, 9, 11, and 13]
Results: Triad significance profile

- significantly overrepresented triads: ID₃#1, ID₃#3, ID₃#10 and ID₃#13

Milo at al. 2004
Results III

• motifs: 10 and 13 – overrepresented

• frequency of all triangular motifs different in Croatian than in English:
  - jako ga voli / really loves it
  - voli ga jako / loves him very much
  - jako voli ga / loves it
  - ga voli jako / loves him very much
  - ga jako voli / he loves
  - voli jako ga / really likes it
Conclusion and future work

- Croatian language networks have similar triad significance profiles for different texts
  - still some differences with other reported languages
  - triads ID3#10 and ID3#13 which are overrepresented

- That may be caused by the free word-order nature of Croatian language.

- Motif-based analysis of the language networks is sensitive to the word order and syntax rules.
Conclusion and future work

- perform motif-based analysis of language networks for different languages on comparable corpora
- analyse syntax networks and sub-word level networks
- analyse the presence of the four-vertex motifs in language networks
  - if they can be interpreted by the semantic relations (polysemy, synonymy) [Biemann et al. 2012]
Network Motifs Analysis of Croatian Literature

Ana Meštrović
amestrovic@uniri.hr
Sanda Martinčić-Ipšić
smarti@uniri.hr
Hana Rizvić
hrizvic@uniri.hr

ITIS 2014 – November 5–7 in Šmarješke toplice, Slovenia
Network Motifs Analysis of Croatian Literature

langnet

Ana Meštrović
amestrovic@uniri.hr

Department of Informatics, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia