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

Multilayerd Language Networks

Sanda Martinčić-Ipšić
smarti@uniri.hr

LangNet team
Ana Meštrović
Slobodan Beliga
Tajana Ban Kirigin
Tanja Miličić

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/nodes - linguistic units
  - edges/links – model their relationships

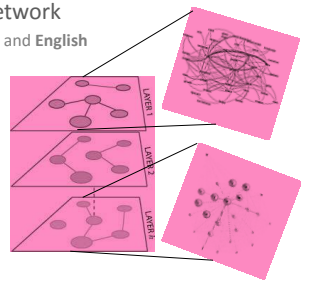
- **word level:**
  - co-occurance
  - syntax
  - semantics
  - pragmatics
- **sub-word level:**
  - morphology (morphosyntactic)
  - syllabic
  - phonetic (phonology)
  - graphemic

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

Outline

- **multilayer language** network
  - experimental results **Croatian** and **English**

- applications
  - **keyword** extraction



Word-level networks

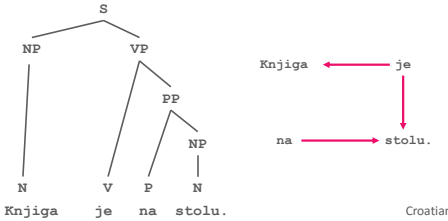
- **co-occurrence networks**
  - **directed** or undirected [ITIS 2013a]
  - **weighted** or unweighted [ITIS 2013a]
  - **stopwords** preserved [ITIS 2013a, MIPRO2014a]
  - **not lemmatized**
    - in the full variety of flective word forms

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

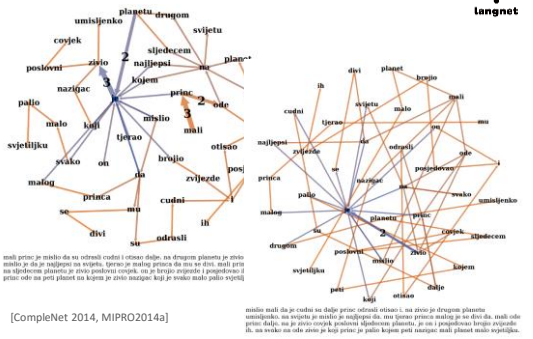
- sensitive to used **corpus**

Syntax Datasets

- **Croatian & English Dependency Treebanks**
  - parsed syntax tree
- **sentences** HR: 3.465 EN: 3.829
- **tokens** HR: 88.045 tokens EN: 94.084 tokens



Sentence-level shuffling

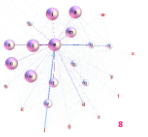


[CompleNet 2014, MIPRO2014a]

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 [Speech, 2016]
    - **phonetic** syllabification: our grapheme-to–phoneme method

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



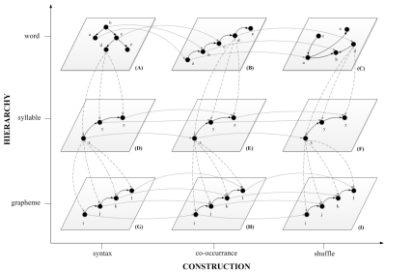
Multilayer Language Network Experiment

- datasets: **Croatian** and **English Dependency Treebanks**
- **10 networks (5 HR+ 5 EN): 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

Multilayer Network definition

- Multilayer Language Network M is a quintuple $\mathbf{M = (V_M, E_M, V, L, C)}$
  - **V** is a non-empty set of nodes;
  - C is a nonempty set of perspective elements;
  - L is a set of perspects L_i where $\{L_0, L_1; L_2\}$ is a partition of C.
    - L_0 - **language perspect**, L_1 - **hierarchy perspect** and L_2 - **construction perspect**;
  - For perspect $L_1 = \{g_1,..., g_k\}$ sequence of its elements $g_1,..., g_k$ is the subsequence of the following sequence - hierarchy:
    - discourse, sentence; phrase; syntagm; word; morphem; syllable; phoneme; grapheme
- An element of the set $L_0 \times L_1 \times L_2$ is called a **layer**;
- $\mathbf{V_M} \subseteq V \times L_0 \times L_1 \times L_2$ is the set whose elements are called **MLN-nodes;**
- $\mathbf{E_M} \subseteq V_M \times V_M$ is the set of **edges**.

Multilayer Network definition



Results: Croatian Dataset

	WORD LEVEL			SUBWORD LEVEL	
	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)

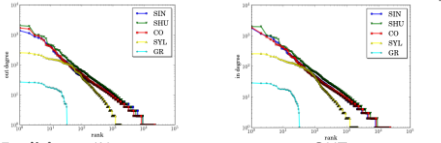
Results: English Dataset

	WORD LEVEL			SUBWORD LEVEL	
	Co-occurr	Shuffled	Syntax	Syllables	Graphemes
Number of nodes (N)	10930	10930	10930	2599	26
Number of edges (K)	50299	58920	52221	6053	333
Number of components	3	1	3	54	1
Average path length (L)	3.47	3.45	1.96	1.88	1.51
Average clustering coefficient(C)	0.286	0.295	0.153	0.057	0.0838
Transitivity	0.009	0.016	0.014	0.020	0.654
Density	0.00042	0.00049	0.00044	0.0009	0.5123

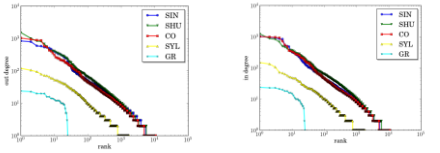
avg. path length - degree of separation between linguistic units
diameter – maximal separation
density – probability of connecting 2 units
transitivity – realized number of triangls (among possible ones)

Degree distributions

- Croatian: IN OUT



- English: IN OUT



Complex network measures

selectivity

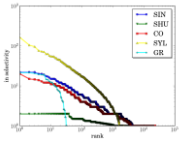
- average weight distribution on the links of the single node
- the average strength of the node

$$s_i^{in/out} = \sum_j w_{ji/ij}$$

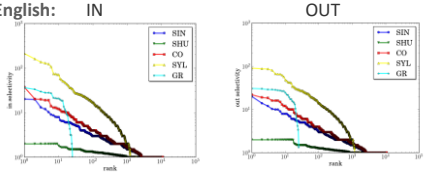
$$e_i^{in/out} = \frac{s_i^{in/out}}{k_i^{in/out}}$$

Selectivity distributions

- Croatian: IN OUT

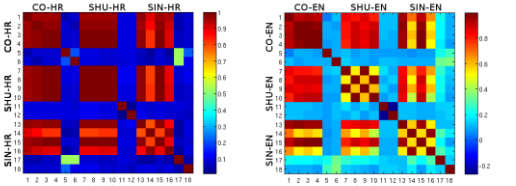


- English: IN OUT



Correlations

Croatian English



- in- & out degree; in- & out- strength; in- & out- selectivity respectively

Word-level layers overlap

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

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

- **Preserved weighted ratio:**

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

- **Preserved weighted overlap:**
  - normalized preserved weighted ratio by the number of intersected links

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

18

Results: Word layers overlap

	CROATIAN			ENGLISH		
	CO-SIN	CO-SHU	SIN-SHU	CO-SIN	CO-SHU	SIN-SHU
Jaccard	**16.72%**	5.47%	4.81%	**13.44%**	6.31%	5.34%
W	**18.96%**	6.43%	5.63%	**13.58%**	6.28%	4.82%
WO	**90.60%**	76.6%	74.6%	**90.00%**	74.72%	73.81%

19

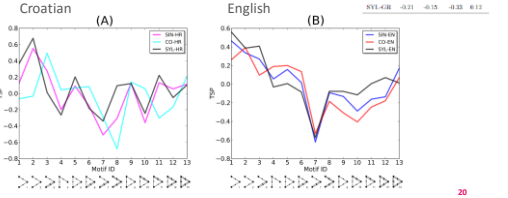
Motifs: Subword vs. Word Layers

- Pearson correlations of motif's motif frequencies and normalized triad significance profiles (TSP) of all layers

$$Z_i = \frac{N_i^{orig} - \langle N_i^{rand} \rangle}{\sigma_i^{rand}} \qquad TSP_i = \frac{Z_i}{\sqrt{\sum_i Z_i^2}}$$

	CROATIAN		ENGLISH	
	Freq.	TSP	Freq.	TSP
CO-SIN	0.96	0.42	**0.91**	**0.92**
CO-SYL	**0.96**	-0.03	**0.86**	0.71
CO-GR	-0.18	-0.36	-0.10	0.16
SIN-SYL	**0.95**	**0.83**	**0.99**	**0.91**
SIN-GR	-0.20	-0.21	-0.11	0.28
SYL-GR	-0.21	-0.15	-0.13	0.12

- Croatian English



20

Recapitulation

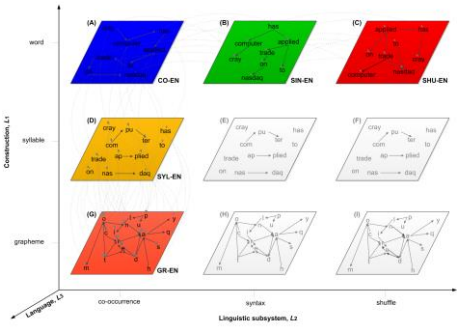
- **co-occurance:** traditional, not sufficient
- **shuffled:** reveals interesting behavior, boundary layer
- **syntax:** more credible for linguistic insights
- **syllables:** like syntax
  - syllables like morphological root
  - morphological networks should be constructed
- **graphemes:** completely different (complex network??)

21

MLN Model

- language networks can be viewed through different **perspects**:
  - **different levels** (e.g. word-level, subword-level),
  - **different construction rules** (e.g. co-occurrence, shuffle),
  - **different languages** (e.g Croatian, English)
- there is a need for a general network model that can capture all language aspects in one single framework
  - we propose an application of general multilayer networks model introduced by Kivela et al. 2014 to the multilayer language networks (MLN)

22



23

Open questions?

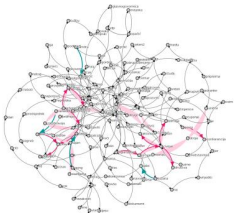
- test on **other** data sets and languages
- comments from linguists
- language network model
  - explanation of M:N relationships between layers
  - quantification of language emergent properties

- S. Martinčić-Ipšić, D. Margan, A. Meštrović. **Multilayer Network of Language: a Unified Framework for Structural Analysis of Linguistic Subsystems, Physica A**: Statistical Mechanics and its Applications, 457, 117-128, 2016,
- T. Ban Kirigin, A. Meštrović, S. Martinčić-Ipšić. **Towards a Formal Model of Language Networks, ICIST 2015**, Communications in Computer and Information Science, Springer, Vol. 538, 469-479, 2015.

24

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

Application: Keyword Extraction



25

Introduction 1/2

Keyword extraction

automatically identify a **set of terms** that best describe the document

- **identify and rank** the most representative features of the source

applications in text:

- summarization, indexing, labeling, categorization, clustering

keyword extraction is traditionally **supervised**

- based on statistical methods
- learning from **hand-annotated data sets**

26

Introduction 2/2

network-based approach

- or graph, since the number of words in isolated documents is limited
- the source (document, text) is modelled in a network
  - words are nodes and their co-occurrence is represented with links
- the **keyword extraction** can exploit: **network, subnetwork** or **node** level measures:
  - coreness, clustering coefficient
    - PageRank motivated ranking score or HITS motivated hub and authority score
  - communities
  - strength, centrality
    - **degree, betweenness, closeness** and eigenvector centrality
    - **selectivity –** the average strength of the node

27

3. Complex network analysis 1/2

strength and selectivity

$$s_i^{in/out} = \sum_j w_{ji/ij}$$

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

Masucci & Rodgers (2009)

selectivity

- average weight distribution on the links of the single node

generalized selectivity

$$ge_i^{\alpha\, in/out} = k_i^{in/out}\left(\frac{s_i^{in/out}}{k_i^{in/out}}\right)^{\alpha}$$

$\alpha \in [0,1]$ – prefers nodes with higher degree

$\alpha \in [1, +\infty]$ – prefers nodes with lower degrees

$\alpha = 1$– node strength

adapted from [Opsahl et al. (2010)]
[Beliga, Meštrović, Martinčić-Ipšić (2016)]

28

Data

- collections with manually annotated keywords by human experts
  - *HINA Croatian News Agency*
  - *Wikipedia: technical reports covering different aspects of CS*

HINA	WIKI-20
8 human experts	15 teams (2 undergrad. stud.)
60 texts	20 texts
Croatian	English
10 keywords on AVG per human	5,7 keywords on AVG per team
human consistency≈ 39.5%	team consistency ≈ 30.5%

29

HINA Dataset

1020 news articles from the Croatian News Agency
- learning: 960 documents
- testing: 60 documents - keywords manually annotated - 8 experts
- per document:
  - 60 to 800 **tokens** (318 on average)
  - 9 to 42 **keywords** (24 on average)

inter-annotator agreement
- average IIC score **39.5%** (31% - 44%)
- no predefined set of keywords list - annotators could make up their own
- in some cases annotated with keywords, which were not present in the original article (**out-of-vocabulary** words - 57%).

preprocessing:
- parsing only text and title (excluding annotations)
- cleaning diacritics and symbols (w inst. of vv, ! inst. of I, etc.)
- lematization and NSW types preserved (numbers, acronyms, abbreviations, etc.)

WIKI-20 dataset

20 technical reports covering different aspects of CS
- 20 documents: keywords manually annotated
- annotators: 15 teams of 2 senior CS undergraduate students independetly
- per document:
  - 5 terms to each report
  - controled vocabulary: Wikipedia articles
  - per document 5.7 keyphrases on average

inter-annotator agreement
- average IIC score **30.5%** (21.4% - 37.1%)

network construction:
- 20 individual networks per document
- 1 integral network for whole collection

Network construction

directed and weighted co-occurrence networks:
- **60 + 20 individual networks:** network per document **+**
- **1+1 integral network :** from all documents (collection extraction)

node: each word

links: two words are linked if they are adjacent in the sentence

weights: proportional to the overall co-occurrence frequencies of the corresponding word pairs

Python + NetworkX

Comparison of centrality measures

DATA SET	MEASURE	TOP 5			TOP 10		
		R_{avg}	P_{avg}	$F1_{avg}$	R_{avg}	P_{avg}	$F1_{avg}$
HINA	Closeness: cc_i	4.3	39.3	7.2	8.4	42.9	12.8
	Betweenness: bc_i	3.6	40.8	6.5	9.2	52.8	13.9
	in/out-degree: $dc_i^{in/out}$	23.0	38.4	17.0	62.1	25.1	**26.6**
	in/out-selectivity: $e_i^{in/out}$	14.0	35.9	**19.3**	16.8	38.0	22.1
WIKI-20	Closeness: cc_i	1.6	42.5	3.1	5.3	63.3	9.7
	Betweenness: bc_i	0.3	10.0	0.5	2.6	55.0	4.9
	in/out-degree: $dc_i^{in/out}$	0.1	5.0	0.3	2.5	57.5	4.8
	in/out-selectivity: $e_i^{in/out}$	2.3	17.8	**3.9**	8.6	18.2	**10.8**

TOP 10 ranked words:
in-degree: to be, and, in, on, which, for, but, this, self, of
betweenness: to be, and, in, on, self, this, which, for, Croatian, but
in/out selectivity: Bratislava, area, Tuesday, inland, revolution, verification, decade, Balkan, freedom, Universe

Selectivity

differentiate between two types of nodes (words)

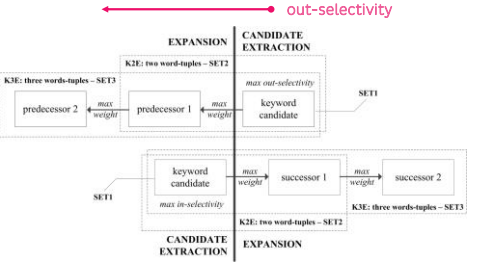
high strength and high degree values **> low selectivity**
- **closed-class words**: stop-words, conjunctions, prepositions

high strength and low degree **> high selectivity**
- **open-class words**: nouns, adjectives, verbs and
- words that are part of collocations, keyphrases, names, etc.

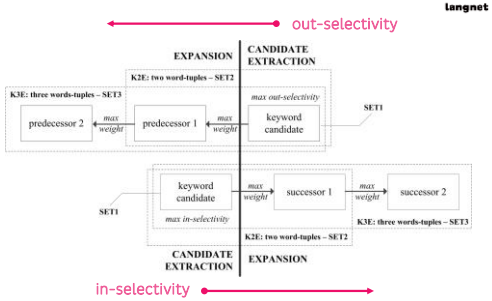
selectivity: + +

efficiently **detect semantically rich open-class words**
extract **better keyword candidates**

SBKE method



SBKE method



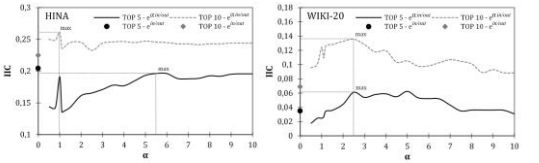
SBKE with Generalized Selectivity

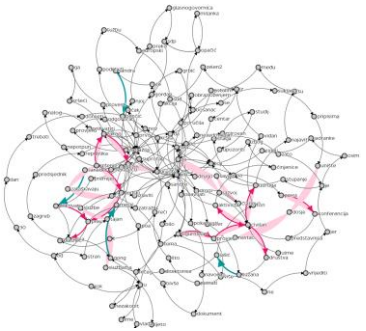
SBKE method can be easily adjusted by incorporating new measures

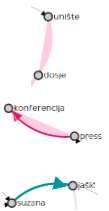
- generalized selectivity measure: $ge_i^{\alpha\ in/out}$
- α adjust the relationship between the node degree and strength – provide different set of keywords
- Find the α parameter which best fits **the experimental** settings according to IIC_O scores

SBKE with Generalized Selectivity

The performance of the first step (SET1) of the SBKE method in terms of IIC_O scores for the TOP 5 and TOP 10 keywords measured for generalized selectivity $ge_i^{\alpha\ in/out}$ with different values of parameter α (plotted as lines) and selectivity ($e^{in/out}$) values (plotted as dots on the y-axis) for HINA and WIKI-20 datasets







Evaluation - IIC

Consistency between any two annotators

- different evaluation approaches compare the obtained results against the gold standard
- in case of multiple annotations, gold standard is not clear
- **Inter-indexer consistency** (Rolling, 1981)
  - human and machine, human and human, or machine and machine
  - equivalent to the harmonic mean of *R* and *P* - **F1 score**

$$IIC = \frac{2c}{a+b}$$

- where **a** and **b** are the number of terms assigned by each annotator, and **c** is the number of terms they have in common

$$IIC_{overall} = \frac{\sum_{i=1}^{N_{doc}} \sum_{j=1}^{N_{an}} IIC_{ij}}{N_{doc} N_{an}}$$

HINA Results

Annotator	IIC_o
1	31.9%
2	33.5%
3	39.3%
4	40.2%
5	41.4%
6	41.6%
7	43.9%
8	44.4%
AVERAGE	**39.5%**

	SBKE	IIC_o	
		e	ge^{α}, α=5.5
TOP 5	SET1	20.5%	19.6%
	SET2	22.1%	21.1%
	SET3	**22.4%**	**21.9%**
		e	ge^{α}, α=1.0
TOP 10	SET1	22.5%	**26.1%**
	SET2	22.7%	19.8%
	SET3	**23.0%**	19.7%

WIKI-20 Results

Annotator	IIC_o
1	21.4%
2	24.1%
3	26.2%
4	28.7%
5	30.2%
6	30.8%
7	31.0%
8	31.2%
9	31.6%
10	31.6%
11	31.6%
12	32.4%
13	33.8%
14	35.5%
15	37.1%
AVERAGE	**30.5%**

	SBKE	IIC_o	
		e	ge^s, α=2.5
TOP 5	SET1	3.5%	6.1%
	SET2	7.5%	10.8%
	SET3	**10.3%**	**12.0%**
TOP 10	SET1	6.9%	13.6%
	SET2	11.4%	16.3%
	SET3	**12.5%**	**17.2%**

Document vs. Collection-Oriented Extraction Results

HINA	60 individual networks			integral network		
	R_{avg}	P_{avg}	$F1_{avg}$	R	P	F1
SET1	18.90	**39.35**	23.60	30.71	**35.80**	33.06
SET2	19.74	39.15	**24.76**	33.46	33.97	**33.71**
SET3	**29.55**	22.96	22.71	**60.47**	19.89	28.89

WIKI-20	20 individual networks ($e_i^{in/out}$ >1)			integral network ($e_i^{in/out}$ >1)			integral network ($e_i^{in/out}$ >2)		
	R_{avg}	P_{avg}	$F1_{avg}$	R	P	F1	R	P	F1
SET1	59.06	**13.40**	**21.48**	76.17	**19.08**	**30.51**	32.71	30.30	31.46
SET2	60.12	12.33	20.46	76.64	18.87	30.29	36.45	31.97	34.06
SET3	**62.17**	12.05	20.19	**77.50**	15.75	26.18	**36.68**	**32.04**	**34.21**

DISCUSSION – Related work

Comparison of the performance of the SBKE method in terms of $F1$ and IIC_O score with other approaches

HINA		
Method	**Approach**	**F1**
tf-idf (Ahel et al., 2009)	unsupervised	13.2%
MDL+POS (Ahel et al., 2009)	supervised	17.2%
SBKE - SET2	graph-based	**24.8%**
Method	**Approach**	**IIC_o**
SBKE - SET1, α=1.0	graph-based	**26.1%**
Humans	gold standard	39.5%

**for TOP 10 KE task*

WIKI-20		
Method	**Approach**	**IIC_o**
tf-idf (Medelyan, 2009)	unsupervised	8.3%
SBKE, α=2.5	graph-based	**12.0%**
KEA++ (Medelyan, 2009)	supervised	22.6%
Humans	gold standard	30.5%
Maui (Wang et al., 2014)	supervised	31.6%
GA (Joorabchi et al., 2013)	supervised	33.5%
Maui+ (Wang et al., 2014)	supervised	33.6%

**for TOP 5 KE task*

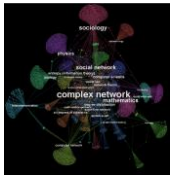
SBKE Conclusion

- **S**electivity-**B**ased **K**eyword **E**xtraction - SBKE method purely from statistical and structural information
  - the source text is reflected into the structure of the network
- **low precision vs. high recall**
  - beside keywords, **personal names and entities** - not marked as keywords
  - comparable with existing supervised and unsupervised methods
- better to longer than to shorter texts - **collection extraction task**
- Keyword annotation is **highly subjective task** - even human experts have difficulties to agree upon keyphrases
  - human annotators F2 score of **inter-annotator agreement** is 46% (SBKE 42%)
- Other graph-based (centrality) approaches
  - similar results but they incorporate linguistic knowledge in a form of different syntactic filters
  - POS tagging, stop-words filtering, noun-phrase parsing, etc.
  - generally more demanding to implement

- S. Beliga, A. Meštrović, S. Martinčić-Ipšić,"**Selectivity-Based Keyword Extraction Method**", *International Journal on Semantic Web and Information Systems*, vol. 12, No. 3, pp. 1-26, 2016, doi: 10.4018/IJSWIS.2016070101.
- S. Beliga, A. Meštrović, S. Martinčić-Ipšić. "**An Overview of Graph-Based Keyword Extraction Methods and Approaches**". *Journal of Information and Organizational Sciences*, vol. 39, No 1, pages 1-20, 2015.
- S. Beliga, A. Meštrović, S. Martinčić-Ipšić. "**Toward Selectivity-Based Keyword Extraction for Croatian News**". CEUR Proceedings of the Workshop on Surfacing the Deep and the Social Web (SDSW 2014), Vol. 1310, pp. 1-8, Riva del Garda, Trentino, Italy, 2014.

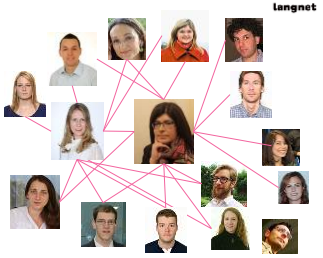
Other applications

- **Text genres differentiation**
  - Can we determine **text genre** using only network structure properties?
  - Legislation or literature? Blogs vs. literature?
- **Authorship attribution** and language detection
- Text **classification**
  - **dimensionality reduction** in **BoW model** using structural properties of text
- **Wikipedia**
  - **knowledge extraction** and linking concepts
- **Twitter**
  - **polarity** (positive/negative) detection
  - **prediction** of missing links
- **Social networks**
  - Coautorship networks analyis
  - scientometrics



LangNet team: langnet.uniri.hr/members.html

- Department of informatics:
  - Sanda Martinčić-Ipšić,
  - Ana Meštrović,
  - Slobodan Beliga,
  - Lucia Načinović Prskalo
- Students:
  - Tanja Miličić,
- Collaborators:
  - Tajana Ban Kirigin,
  - Mihaela Matešić,
  - Benedikt Perak,
  - Zoran Levnajić
- Past members:
  - Domagoj Margan, Edvin Močibob, Neven Matas, Kristina Ban, Hana Rizvić, Sabina Šišović, Luka Vretenar



* University of Rijeka grant 13.13.2.2.07

48

Sanda Martinčić-Ipšić
smarti@uniri.hr

Language Networks

Language Networks Prospects

- **universal language model**
  - include many languages
  - add perspectives /layers: semantics, cognitive representation
- **text quality evaluation:**
  - derive an **assessment model** for the evaluation of the quality of texts from complex networks parameters
  - creativity?
  - keywords extraction, summarization

50

Generalized selectivity

Additional observations

- selectivity takes into account weights which represent the frequencies of word bigrams
  - captures the importance of the node strength which is crutial for weighted networks
- degree centrality also yields optimistic results
  - combine selectivity + degree = **generalized selectivity**

$$ge_i^{\alpha\, in/out} = k_i^{in/out}\left(\frac{s_i^{in/out}}{k_i^{in/out}}\right)^{\alpha}$$

$\alpha \in [0,1]$ – prefers nodes with higher degrees

$\alpha \in [1, +\infty]$ – prefers nodes with lower degrees

$\alpha = 1$ – node strength

Beliga, Meštrović, Martinčić-Ipšić (2016)

51