Terminology acquisition and description using ruled-based methods and lexical resources

CVETANA KRSTEV

UNIVERSITY OF BELGRADE, FACULTY OF PHILOLOGY

DEPARTMENT OF LIBRARY AND INFORMATION SCIENCES

Outline of my talk

Motivation and related work

Methodology and Design

System Architecture

Evaluation

Future work

Related work and Motivation

Approaches to Multi-word term extraction

Statistical approach

Rule-based with lexical resources

Hybrid approaches

Statistical approach (1)

The seminal work of **Kenneth Church**, **Patrick Hanks** and **Frank Smadja** in ’90s, followed by many reserachers form term extraction for many languages.

Introduced various measures, like Mutual information ***I*(*x*, *y*)**:

- If two words ***x*** and ***y*** have probabilities of occurrence ***P*(*x*)** and ***P*(*y*)**, then their mutual information is

$$I(x;y) \equiv \log_2 \frac{P(x,y)}{P(x)\ P(y)}$$

- mutual information compares the probability of observing ***x*** and ***y*** together (the joint probability) with the probabilities of observing x and y independently (chance).

Statistical approach – mutual information(2)

If there is a genuine association between ***x*** and ***y***, then the joint probability ***P*(*x*, *y*)** will be much larger than chance ***P*(*x*) *P*(*y*)**, and consequently ***I*(*x*, *y*) >> 0**.

If there is no interesting relationship between ***x*** and ***y***, then ***P*(*x*, *y*) ≈ *P*(*x*) *P*(*y*)**, and thus, ***I*(*x*, *y*) ≈ 0**.

word probabilities ***P*(*x*)** and ***P*(*y*)** are estimated by counting the number of observations of ***x*** and ***y*** in a corpus, ***f* (*x*)** and ***f*(*y*)**, and normalizing by ***N***, the size of the corpus. Joint probabilities, ***P*(*x*, *y*)**, are estimated by counting the number of times that ***x*** is followed by ***y*** in a window of ***w*** words, f_w **(*x*,*y*)**, and normalizing by ***N***.

Examples: **bread and butter**, **drink and drive**

Statistical approach – T-test (3)

T-test, a measure of dissimilarity

- Example: **powerful support** vs. **strong support**
- Comparision of ***I*(„strong“, „support“)** and ***I*(„powerful“, „support“)** cannot discard **powerful support** as a term because usually corpus used for calculation is not big enough (differences are not statistically significant).

$$t = \frac{P(powerful\ support) - P(strong\ support)}{\sqrt{\sigma^2(P(powerful\ support)) + \sigma^2(P(strong\ support))}}$$

- On the corpus of the same size this measure will show that probability of **powerful support** is significantly less likely than **strong support** measured in standard deviations.
- More on measures when we will talk about evaluation

Rule-based with lexical resources

A simple syntactic patterns are recognized in corpus that are good filters for terminology extraction.

An example:

- **((A | N)+ | ((A | N)$^{?}$(NP)-)(A | N)*)N**
- a candidate term is a multi-word noun phrase; and it either is a string of nouns and/or adjectives, ending in a noun, or it consists of two such strings, separated by a single preposition.

The prerequisit of such methods is that a copus was (at least) Part-Of-Speech tagged.

Applied in:

- A cascade of transducers on Arabic scientific and technical corpus,
- **SEJFEK**, a grammatical lexicon of about 11,000 Polish MWTs from the economical domain, where inflectional and syntactic variations are described via graph-based rules

Hybrid approaches

Linguistic rules are used to parse a text and retrieve terms

statistical measures are used for disambiguation and ranking, usually as a prerequisite for evaluation.

Applied in many systems for many languages:

- Arabic,
- Bulgarian,
- Slovene,
- Polish,
- ...

Motivation for rule-based extraction and lemmatization

MWT candidates are extracted from texts in various inflected forms

Example:

- električna energija
- električne energije
- električnoj energiji
- električnu energiju

electric power

Do we want evaluators to evaluate all different inflected forms?

Each particular form might not be frequent enough and so not included as a candidate, while all forms belonging to one lemma might have a more significant frequency.

Do we want to include all extracted forms into a term base, when this list might not be exhaustive?

Read more

Church, K. W., Hanks, P., (1990). Word association norms, mutual information, and lexicography, *Computational Linguistics*, 16, pp. 22-29.

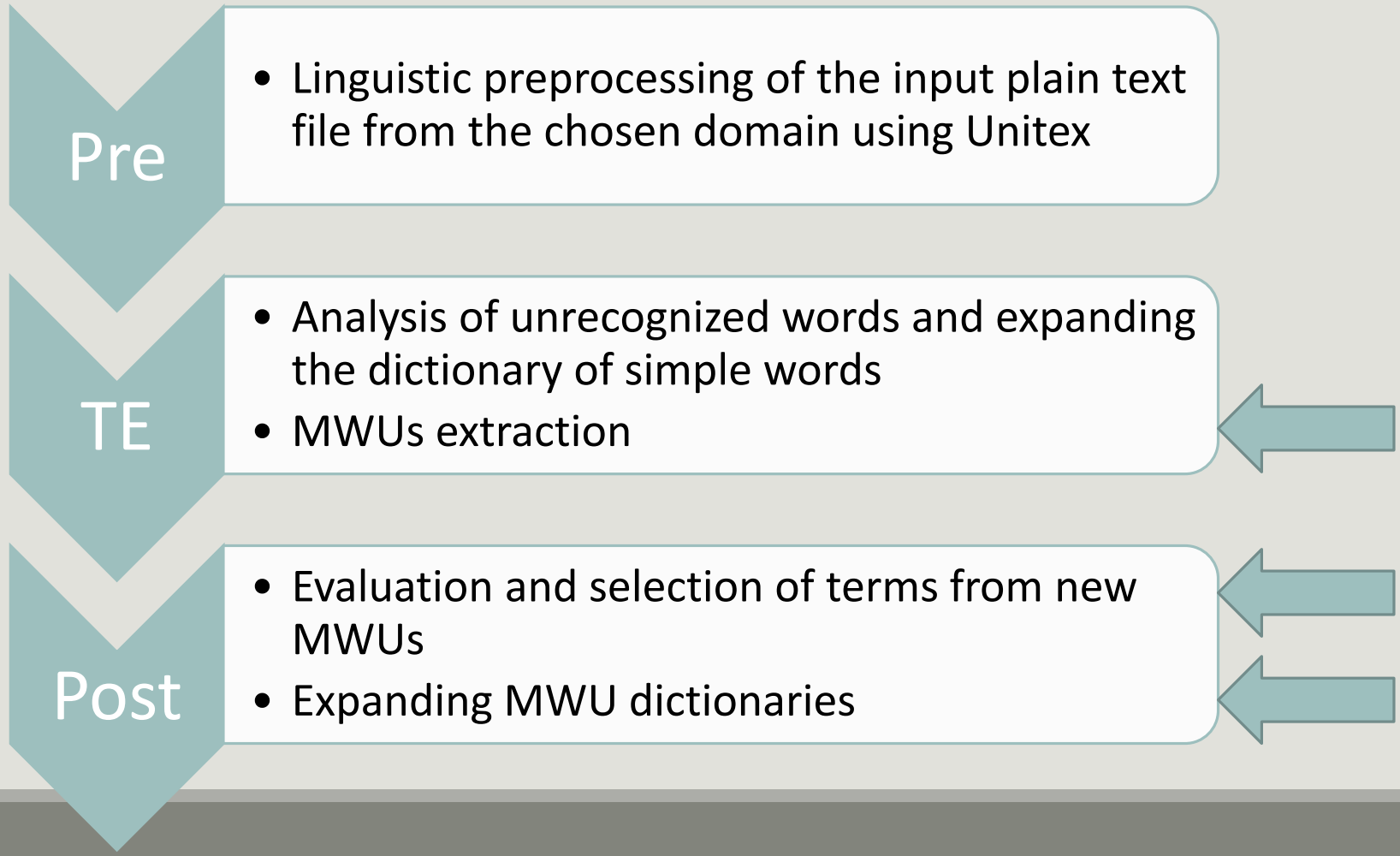
Church, K. W. Gale, W., Hanks, P., Hindle, D. (1991). *Using statistics in lexical analysis*, In U. Zernik (Ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 115--164.

Smadja, F.(1993). Retrieving Collocations from Text: Xtract, *Computational Linguistics*, 19(1), pp. 143-177.

Justeson, John S., and Slava M. Katz. "Technical terminology: some linguistic properties and an algorithm for identification in text." *Natural language engineering* 1.01 (1995): 9-27.

Methodology and Design

Process of terminology acquisition



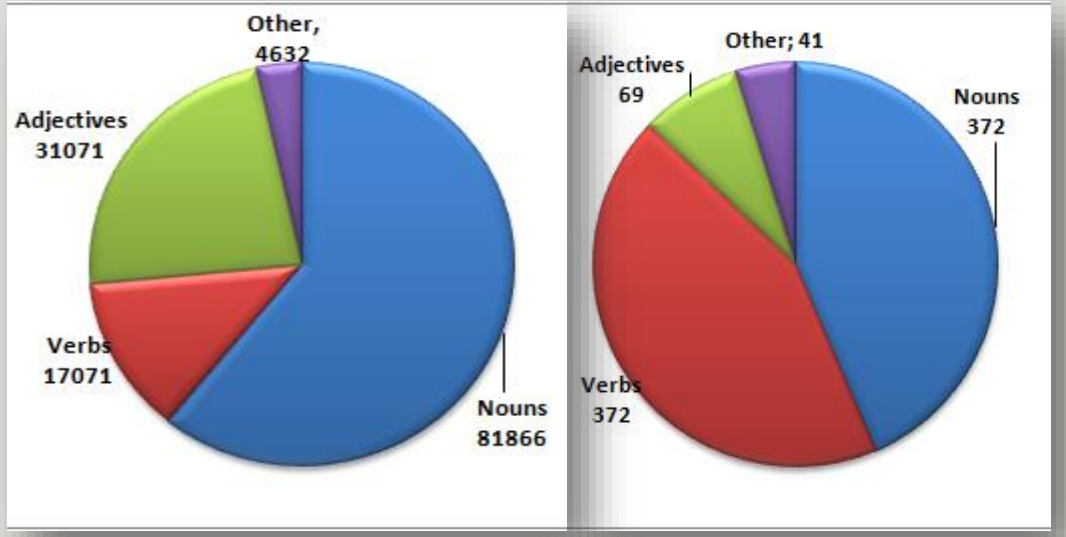
Main lexical resource – e-dictionaries

Morphological e-dictionaries of Serbian are in DELA format

DELAS > 135,000 lemmas generating more than 5 million forms (DELAF)

DELAC > 15,000 MWUs (Krstev, 2008)

POS	lemmas		FSTs	
Nouns	81,866	61%	372	44%
Verbs	17,071	13%	372	44%
Adjectives	31,071	23%	69	8%
Other	4,632	3%	41	5%
Total	134,640		854	



Structure of terms in termbases

Frequencies of terms of different lengths in samples from 3 termbases

The most frequent are MWT with two components – the frequency drops with the rise of the numebr of components

Dictionary	Term length (in number of words)				
	1	2	3	4	≥5
GeolISS Term geoliss.mre.gov.rs/term	**1436**	**2356**	**749**	**305**	**243**
RNBS rbi.nb.rs/en/home.html	**3302**	**6180**	**2062**	**806**	**415**
RudOnto rudonto.rgf.bg.ac.rs	**1004**	**1351**	**1350**	**1031**	**2341**

Multi-word unit (MWU) classes

MWUs are classified according to their syntactic structure and inflectional and other properties

Class names correspond to finite-state transducers (FST) used for inflection of MWUs belonging to that class – one class encompasses one or more FSTs.

For example, MWUs composed of a noun (**N**) preceded by an adjective (A), which agrees with a noun in gender, number, case and animateness, belong to the **AXN** class.

- **X** stands for a component that does not inflect when the MWU inflects or a separator, usually a space or a hyphen.

Class **AXN** is covered by 5 FSTs (MWU inflects in number or not, components change order in a MWU or not, a MWU change gender with a number, etc.)

Analysis of lexical resources

Nominal MWUs in Serbian :

- 14 classes account for more than 98% of all nominal MWUs

Number of components

- 4 with 2 components,
- 5 with 3 and
- 5 with 4 components

Граф	Број ВЛЈ	%
NC_AXN	5893	45,91
NC_AXN3	2169	16,90
NC_N4X	625	4,87
NC_N2X	589	4,59
NC_2XN	417	3,25
NC_N2X2	385	3,00
NC_2XN3	350	2,73
NC_2XN1	295	2,30
NC_N4X1	238	1,85
NC_AXN2X	236	1,84
NC_NXN	180	1,40
NC_AXAXN	163	1,27
NC_N6X	157	1,22
NC_AXAXN1	106	0,83
NC_NXN2	93	0,72
NC_N6X1	71	0,55
NC_NXN2m	71	0,55
NC_2XAXN	65	0,51
NC_AXN4X	64	0,50
NC_AXN2X1	42	0,33
NC_N8X	36	0,28
NC_2XN2	30	0,23

Syntactic patterns for MWTs (1)

1. **AXN** – adjective/noun that agree in all four grammatical categories; e.g. **zemni gas** ‘natural gas’
2. **2XN** – usually a prefix, an adverb derived from an adjective, a word of a foreign origine that does not inflect in a MWU, e.g. **anker-mreža** ‘anchor network’
3. **N2X** – usually noun in genitive or instrumental case; e.g. **patrona eksploziva** ‘explosive cartridge’
4. **N4X** – a noun followed by two words that do not inflect in the MWU:
   a) **NNgi** - **otkopavanje širokim čelom** ‘broad forehead excavation’
   b) **NprepNp** - **lanac sa grabuljama** ‘chain with a rake’
5. **AXN2X** – **geološko kartiranje terena** ‘geological field mapping’

Syntactic patterns for MWTs (2)

6. **NXN** – two nouns that agree in the number and the case, e.g. **bager kašikar** ‘shovel excavator’
7. **AXAXN** – **površinski istražni radovi** ‘surface exploration works’
8. **N6X** - a noun followed by three words that do not inflect in the MWU
   a) **NNgiPrepNp** - **priprema ležišta za otkopavanje** ‘deposit preparation for mining’
   b) **NNgiNgiNgi** - **istraživanje ležišta mineralnih sirovina** ‘exploration of mineral deposits’
   c) **NprepNpNgi** - **bakar sa primesama zlata** ‘copper with a sprinkling of gold’

Syntactic patterns for MWTs (3)

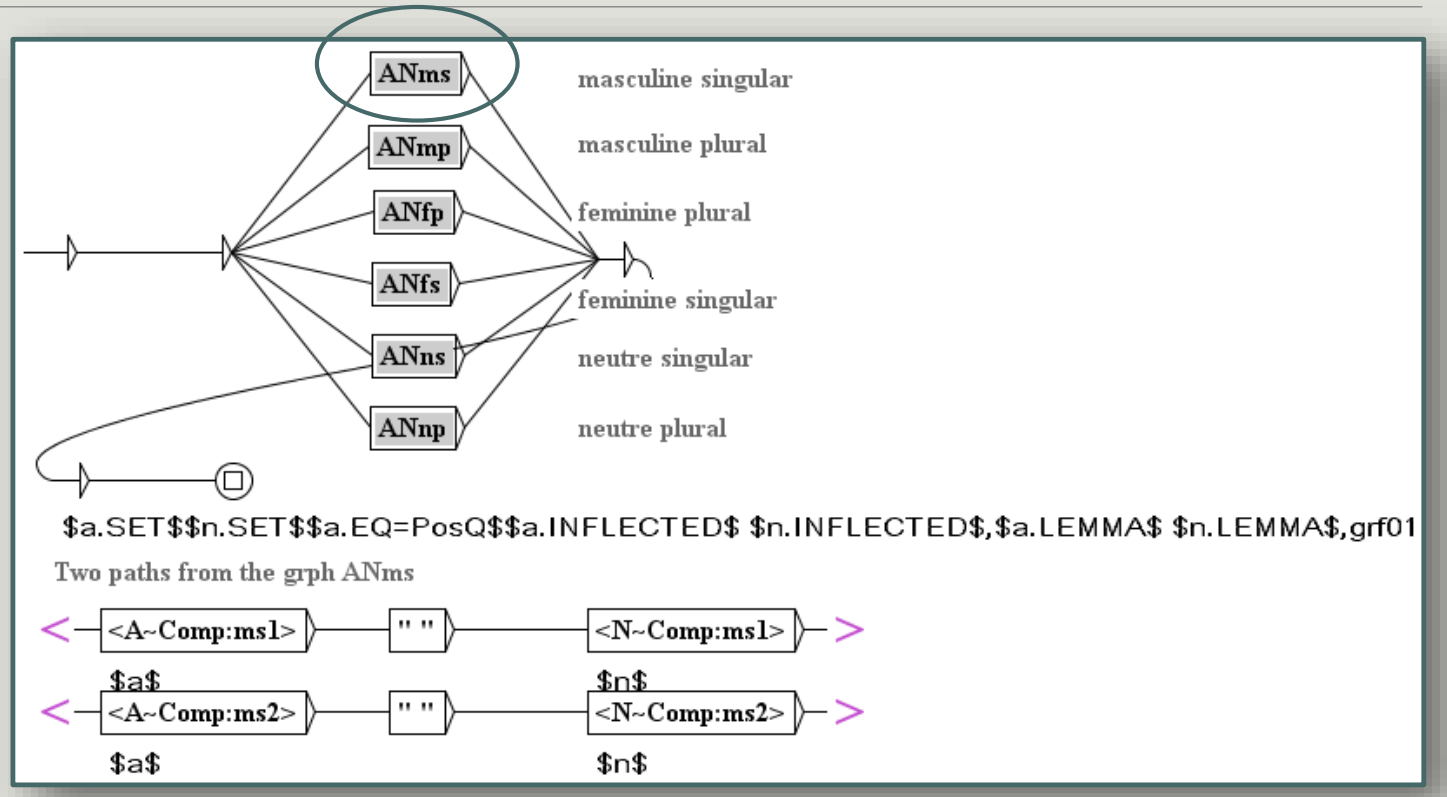
9. **AXN4X** - adjective/noun that agree in all four grammatical categories, followed by two words that do not inflect in the MWU:
   a) **ANPrepNp** - **gravitacijska koncentracija u vodi** ‘gravity concentration in water’
   b) **ANNgiNgi** - **površinska eksploatacija mineralnih sirovina** ‘surface exploitation of mineral resources’
10. **2XAXN** - e, a word of a foreign origine that does not inflect in a MWU followed by adjective/noun that agree in all four grammatical categories, e.g. **magmatsko-eruptivni masiv** ‘magmatic-igneous massif’

The example of one FST for extraction of MWTs of type AXN (1)

The FST retrieves adjective/noun expressions from untagged texts.

PoS and agreement conditions are checked against e-dictionaries.

There is one subgraph for each gender/number pair.

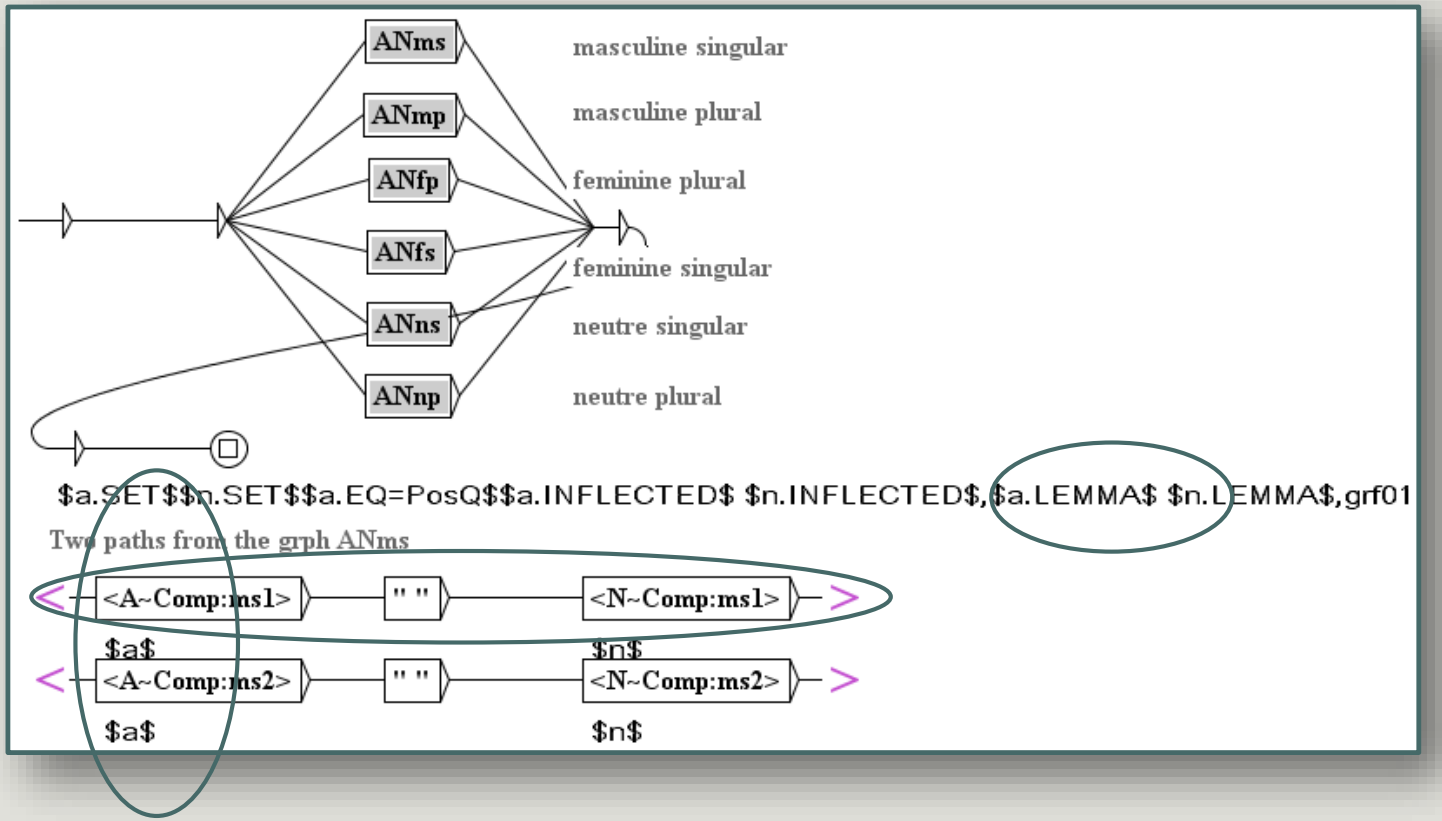


The example of one FST for extraction of MWT of type AXN (2)

Two paths from the first subgraph that illustrate the agreement between adjectives and nouns.

Recognied adjectives and nouns become values of variables, **a** and **n** respectively.

Dictionary variable used for FST output in the form **$a.LEMMA$** retrieves a lemma of recognized word form **a** thus performing the simple word lemmatization



The result of extraction FSTs – simple word lemmatization (SWL)

As a result extracted terms are obtained with additional information:

- Syntactic structure (e.g. **AXN**)
- The grammatical number of a lemma (singular or plural)
- For all MWT components their lemmas are provided (this information is obtained from e-dictionaries)

Examples:

- Extracted form: **zemnog gasa** ’natural gas’; syntactic structure: AXN; gram. number: sin; SWL: **zemni gas**
- Extracted form: **radnih uslova** ’working conditions’; syntactic structure: **AXN**; gram. number: pl; SWL: **radni uslovi**

From simple word lemmatization to multi-word lemmatization (MWL)

Simple word lemmatization do not produce always correct multi-word lemmas (approx. 50% are incorrect).

Two most frequent sources of errors:

- A noun is not in the masculine gender, and the lemma of an adjective that should agree with it is always in masculine gender.
  - SWL: **magnetski** (m) **polje** (n) – **magnetsko** (n) **polje** (n) ’magnetic field’
  - SWL: **vučni** (m) **sila** (f) – **vučna** (f) **sila** (f) ’tractive force’
- A MWL should be in plural and lemmas of its constituents are in singular, as listed in (e-)dictionaries.
  - SWL: **radni** (s) **uslov** (s) – **radni** (p) **uslovi** (p) ’working conditions’
  - SWL: **površinski** (ms) **voda** (fs) – **površinske** (fp) **vode** (fp) ’surface water’

Impact on the whole process

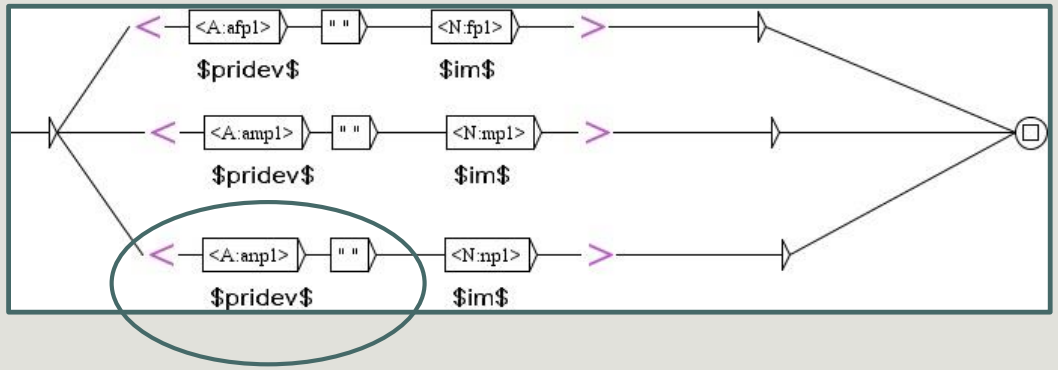
Stat. extractor confused	Stat. extractor satisfied	Stat. extractor satisfied
Evaluator confused	**Evaluator confused**	**Evaluator satisfied**
Dictonary prod. confused	**Dictonary prod. confused**	**Dictonary prod. satisfied**
magnetskog polja	magnetski polje	magnetsko polje
vučne sile	vučni sila	vučna sila
radnih uslova	radni uslov	radni uslovi
površinskih voda	površinski voda	površinske vode
Inflected forms	**SWL**	**MWL**

A FST for production of correct MWLs from SMLs

Example for MWTs with **AXN** structure

For given simple word lemmas, inflected forms are retrieved form e-dictionaries that satisfy certain agreement conditions.

Dictionary variables (e.g. **$pridev.LEMMA$**) will output correct component forms.



The ambiguity of produced MWLs (and SWLs) (1)

For a certain word combinations that satisfy syntactic conditions more than one MWU is offered by the system.

Obviously, only one offered MWL is correct.

Two sources of ambiguity:

- For certain word forms (MWT components) more than one lemma is offered by e-dictionaries;
  - E.g. The form **karata** can be lemmatized as **karat** ’carat’ and as **karta** ’map’
- Certain word forms (MWT components) can be interpreted as different grammatical forms:
  - E.g. The form **knjiga** can be the nominative singular or genitive plural of **knjiga** ’book’

The ambiguity of produced MWLs (and SWLs) (2)

There are inflected forms of MWTs that are syntactically ambiguous (parasite ambiguity)

- E.g: the genitive form **obrade podataka** of **obrada podataka** ’data processing’ (**NNg)** can be interpreted as : the genitive form **obrada podatak** (**NXN)** (incorrect)
- E.g. The genitive form **biblioteke celine** of **biblioteka celina** (**NXN**) can be interpreted as : the genitive form **biblioteka celine** (**NNg**) (incorrect)

Solution of the ambiguity problem

We want to eliminate as much as possible of incorrect and „parasite“ lemmas for one or several extracted MWTs.

In that way we facilitate the evaluation task.

To do that we use:

- Data-driven approach
- heuristics

Data-driven approach (1)

One extreme example: **obloga trake** ’belt coating’ (**NNg**) .

This MWT from the mining domain has 8 different inflected forms:

- **obloga trake**
- **obloge trake**
- **oblozi trake**
- **oblogu trake**
- **oblogo trake**
- **oblogom trake**
- **oblogama trake**

The constituents of these forms (8 for **obloga** and 1 for **traka**) can represent forms of 4 lemmas:

- Two for obloga: **obloga** ’coating’ and **oblog** ’stupe’
- Two for traka: **traka** ’belt’ and **trak** ’tentacle’

Data-drive approach – the result

MWT inflected form	Grammatical values	Struc.	First constituent	Second constituents
obloga trake	s_nom+ p_gen	**N2X**	**obloga_s_nom**	**traka_s/p_gen**
			obloga_p_gen	traka_s/p_gen
			oblog_s/p_gen	traka_s/p_gen
		NXN	oblog_s_gen	traka_s_gen
obloge trake	s_gen+ p_nom/ acc/voc	**N2X**	**obloga_s_gen+ p_nom/ acc/voc**	**traka_s/p_gen**
			oblog_p_acc	traka_s/p_gen
		NXN	oblog_p_acc	trak_p_acc
			oblog_p_acc	traka_p_acc
			obloga_p_acc	trak_p_acc
			obloga_s_gen+ p_nom/acc	traka_s_gen+ p_nom/acc
oblozi trake	s_dat/loc	**N2X**	**obloga_s_dat/ loc**	**traka_s/p_gen**
		NXN	oblog_p_nom/voc	traka_p_nom/ voc
oblogu trake	s_acc	**N2X**	**obloga_s_acc**	**traka_s/p_gen**
			oblog_s_dat/ loc	traka_s/p_gen
oblogotrake	s_voc	**N2X**	**obloga_s_voc**	**traka_s/p_gen**
oblogom trake	s_ins	**N2X**	**obloga_s_ins**	**traka_s/p_gen**
			oblog_s_ins	traka_s/p_gen
oblogama trake	p_dat/ ins/loc	**N2X**	**obloga_p_dat/ ins/loc**	**traka_s/p_gen**

The 8 forms of **obloga trake** yield 9 interpretations:

- **obloga.obloga trake.traka (NNg-sin) – correct**
- **obloge.obloga trake.traka (NNg-plu)**
- **oblog.oblog trake.traka (NNg-sin)**
- **oblozi.oblog trake.traka (NNg-plu)**
- **obloga.obloga traka.traka (NXN-sin)**
- **obloge.obloga trake.traka (NXN-plu)**
- **oblog.oblog traka.traka (NXN-sin)**
- **oblozi.oblog trake.traka (NXN-plu)**
- **oblozi.oblog trak.trak (NXN-plu)**

Distribution among various forms is uneven: from one to up to 6 possibilities

Data-drive approach – disambiguation

MWT inflected form	Grammatical values	Struc.	First constituent	Second constituents
obloga trake	s_nom+ p_gen	**N2X**	**obloga_s_nom**	**traka_s/p_gen**
			obloga_p_gen	traka_s/p_gen
			oblog_s/p_gen	traka_s/p_gen
		NXN	oblog_s_gen	traka_s_gen
obloge trake	s_gen+ p_nom/ acc/voc	**N2X**	**obloga_s_gen+ p_nom/ acc/voc**	**traka_s/p_gen**
			oblog_p_acc	traka_s/p_gen
		NXN	oblog_p_acc	trak_p_acc
			oblog_p_acc	traka_p_acc
			obloga_p_acc	trak_p_acc
			obloga_s_gen+ p_nom/acc	traka_s_gen+ p_nom/acc
oblozi trake	s_dat/loc	**N2X**	**obloga_s_dat/ loc**	**traka_s/p_gen**
		NXN	oblog_p_nom/voc	traka_p_nom/ voc
oblogu trake	s_acc	**N2X**	**obloga_s_acc**	**traka_s/p_gen**
			oblog_s_dat/ loc	traka_s/p_gen
oblogotrake	s_voc	**N2X**	**obloga_s_voc**	**traka_s/p_gen**
oblogom trake	s_ins	**N2X**	**obloga_s_ins**	**traka_s/p_gen**
			oblog_s_ins	traka_s/p_gen
oblogama trake	p_dat/ ins/loc	**N2X**	**obloga_p_dat/ ins/loc**	**traka_s/p_gen**

Our approach works under assumption that there are no MWT forms that can have two different correct lemmas (which is almost true)

Two forms are discriminative:

- **oblogo trake** (vocative/sin)
- **oblogama trake** (dat,ins,loc/plu)

If these two forms were extracted from the corpus (among some other forms), all parasite lemmas would be rejected.

Data-drive approach – uncomplete rejection

MWT inflected form	Grammatical values	Struc.	First constituent	Second constituents
obloga trake	s_nom+ p_gen	N2X	obloga_s_nom	traka_s/p_gen
			obloga_p_gen	traka_s/p_gen
			oblog_s/p_gen	traka_s/p_gen
		NXN	oblog_s_gen	traka_s_gen
obloge trake	s_gen+ p_nom/ acc/voc	N2X	obloga_s_gen+ p_nom/ acc/voc	traka_s/p_gen
			oblog_p_acc	traka_s/p_gen
		NXN	oblog_p_acc	trak_p_acc
			oblog_p_acc	traka_p_acc
			obloga_p_acc	trak_p_acc
			obloga_s_gen+ p_nom/acc	traka_s_gen+ p_nom/acc
oblozi trake	s_dat/loc	N2X	obloga_s_dat/ loc	traka_s/p_gen
		NXN	oblog_p_nom/voc	traka_p_nom/ voc
oblogu trake	s_acc	N2X	obloga_s_acc	traka_s/p_gen
			oblog_s_dat/ loc	traka_s/p_gen
oblogom trake	s_ins	N2X	obloga_s_ins	traka_s/p_gen
			oblog_s_ins	traka_s/p_gen

If these two forms

- **oblogo trake** (vocative/sin)
- **oblogama trake** (dat,ins,loc/plu)

were NOT extracted from the corpus, maybe not all parasite lemmas would be rejected, but at least some.

For instance, the parasite lema **oblozi.oblog trak.trak (NXN-plu)** is offered as a possibility only for the form **obloge trake**.

If any other form is retrieved as well this one will be rejected.

Data-driven approach – a more realistic example

4 different forms were retrieved for the MWT **električna energija** 'electric power'

Only one of them offers a MWL in plural – **električne energije**

It is rejected

Graph	Num	Recognized form	Frequency	Temporary lemma	Lemma
grf01	plu	električne energije	85	električni energija	električne energije
grf01	sin	električna energija	10	električni energija	**električna energija**
		električne energije	85		
		električnom energijom	8		
		električnu energijo	5		

Heuristic approach (1)

MWT inflected form	Grammatical values	Struc.	First constituent	Second constituents
obloga trake	s_nom+ p_gen	N2X	obloga_s_nom	traka_s/p_gen
			obloga_p_gen	traka_s/p_gen
			oblog_s/p_gen	traka_s/p_gen
		NXN	oblog_s_gen	traka_s_gen

If after applying the data-drive approach some parasite approach still remain we apply the heuristic approach.

For instance if only the form **obloga trake** is extracted from the corpus, three parasite MWL remain (nothing could be rejected by data-drive approach)

- **obloge.obloga trake.traka (NNg-plu)**
- **oblog.oblog trake.traka (NNg-sin)**
- **oblog.oblog traka.traka (NXN-sin)**

Heuristic approach (2)

MWT inflected form	Grammatical values	Struc.	First constituent	Second constituents
obloga traka	s_nom+ p_gen	N2X	obloga_s_nom	traka_s/p_gen
			obloga_p_gen	traka_s/p_gen
			oblog_s/p_gen	traka_s/p_gen
		NXN	oblog_s_gen	traka_s_gen

oblog.oblog traka.traka (NXN-sin)

- This MWL is rejected because MWT of syntactic structure NXN are much less frequent then those having NNg structure (evidence from e-dictionaries)

obloge.obloga trake.traka (NNg-plu)

- This MWL is rejected because MWL in plural are much less frequent then those in singular (evidence from e-dictionaries)

oblog.oblog trake.traka (NNg-sin)

- This parasite MWL remains – it will be rejected in the evaluation phase.

Production of complete MWT lemma and its inflected forms

After extraction, multi-word lemmatization, and disambiguation a compete MWT lemma can be produced automatically

This MWT lemma enables automatic production of its ALL inflected forms (not just those extracted from the corpus) and this forms can by included in a term-base.

DELAC	električna(električni.A2:aefs1g) energija(energija.N600:fs1q),NC_AXN
DELACF	električnoj energiji,električna energija.N:fs7q
	električne energije,električna energija.N:fp1q
	električnih energija,električna energija.N:fp2q
	električnim energijama,električna energija.N:fp3q

Read more

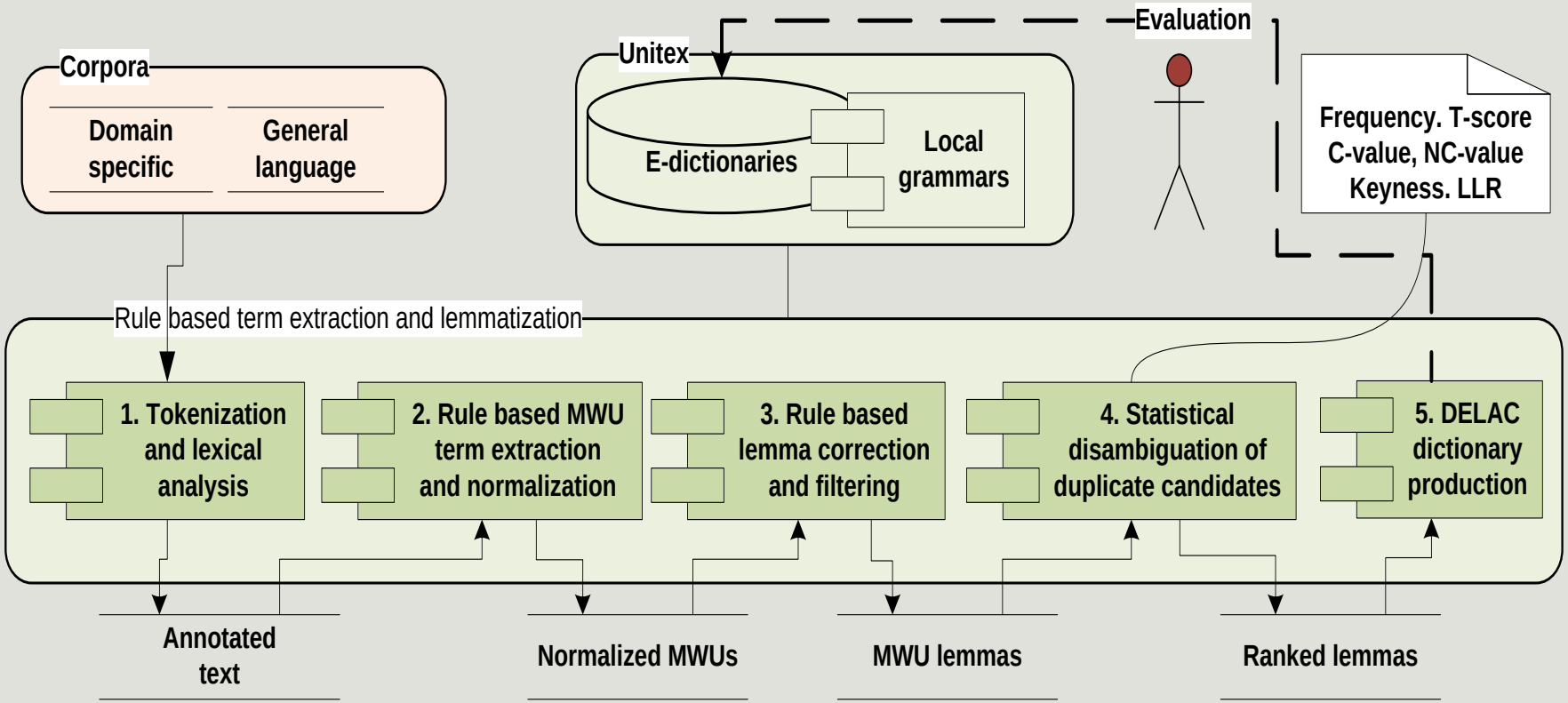
Courtois, B., Silberztein, M. (1990). Dictionnaires électroniques du français. Larousse, Paris.

Cvetana Krstev, *Processing of Serbian – Automata, Texts and Electronic dictionaries* Faculty of Philology, University of Belgrade, Belgrade, 2008.

Krstev, C., Stanković R., Obradović I., Lazić B. Terminology acquisition and description using lexical resources and local grammars, *Terminology & Artificial Intelligence 2015*, Granada

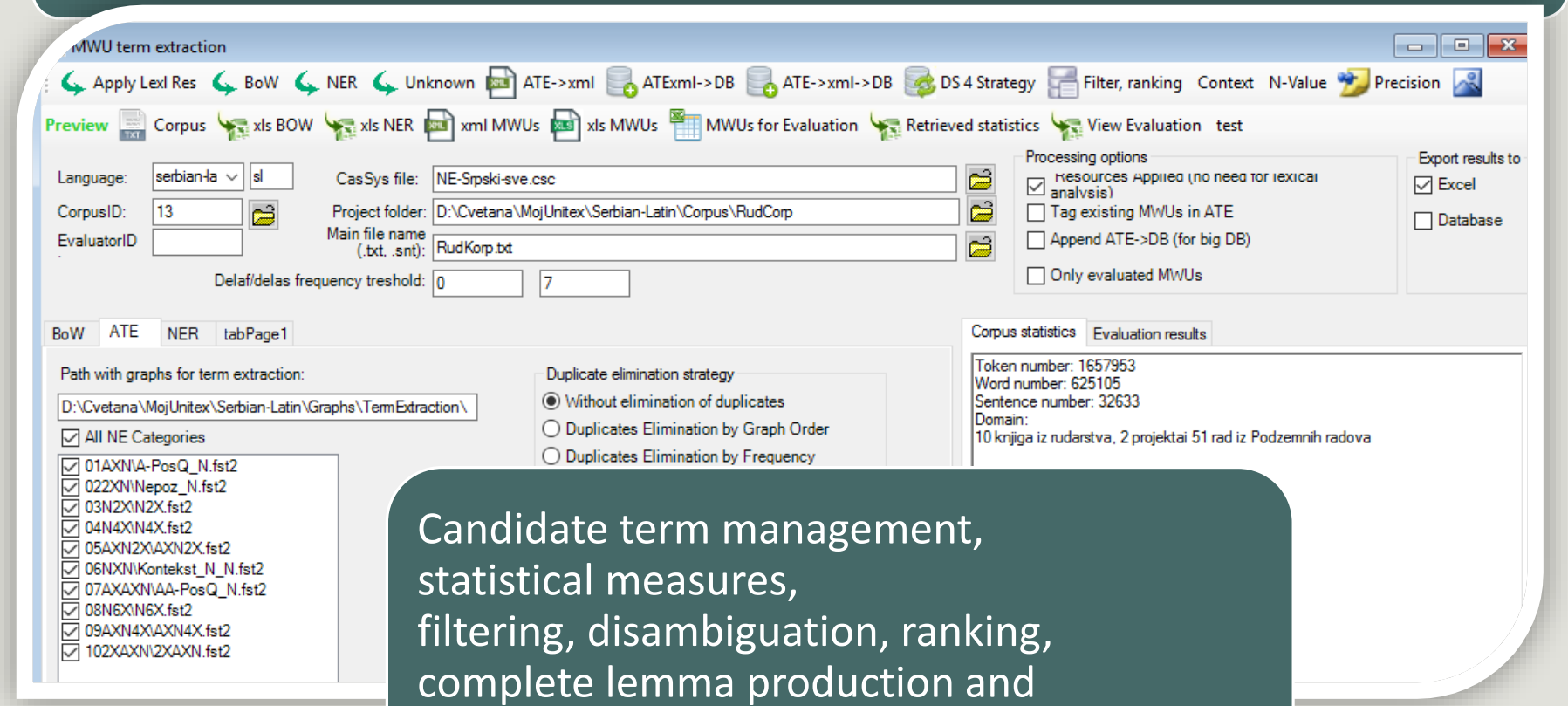
System Architecture

Architecture of the system for MWU extraction



Software solution for MWTE

Graphs integrated into LeXimir with metadata corpus management for specific domains (mathematics, geology, energetics, library science,…)



Candidate term management,
statistical measures,
filtering, disambiguation, ranking,
complete lemma production and
evaluation management

Evaluation

MWT extraction for mining – results of extraction by FSTs

Graph		sin		plu	
		Lemma	Forms	Lemma	Forms
grf01	AXN	570	11,845	481	8,200
grf02	2XN	6	55	5	53
grf03	N2X	701	11,330	613	6,931
grf04a	N4X	148	2,113	107	1,359
grf04b	N4X	122	2,002	106	1,171
grf05	AXN2X	63	800	33	205
grf06	NXN	299	4,249	134	1,751
grf07	AXAXN	5	37	4	36
grf08a	N6X	6	77	6	46
grf08b	N6X	8	103	8	72
grf08c	N6X	9	103	7	81
grf09a	AXN4X	17	195	10	124
grf09b	AXN4X	9	86	7	47
grf10	2XAXN	8	62	6	100
Total		**1,971**	**33,057**	**1,527**	**20,176**

Total (plu+sin)	**3,498**	**53,233**

Terms (forms and lemmas) extracted by graphs with numbers of lemmas that passed the frequency threshold 7

Evaluation on a corpus:

10 textbooks, 2 projects and 51 journal articles from the mining domain

32,633 sentences and 625,105 simple word forms

MWT extraction for mining – results of the rejection of ambiguous forms

		sin		plu	
Graph		Lemma	Forms	Lemma	Forms
grf01	AXN	568	11,828	47	740
grf02	2XN	6	55		
grf03	N2X	668	10,993	8	34
grf04a	N4X	143	2,065	2	22
grf04b	N4X	122	2,002	1	9
grf05	AXN2X	63	800	3	28
grf06	NXN	265	3,903		
grf07	AXAXN	5	37	1	8
grf08a	N6X	6	77		
grf08b	N6X	6	83	1	7
grf08c	N6X	9	103		
grf09a	AXN4X	17	195		
grf09b	AXN4X	8	78		
grf10	2XAXN	8	62	1	24
Total		1,894	32,281	64	872

Total (plu+sin)	**1,958**	**33,153**

Number of lemmas that are passed for manual evaluation

Evaluation measures

Evaluation measures used in our system:

- Frequency
- CValue
- Tscore
- LLR
- Keyness

CValue (1)

A measure known as **CValue** for extracting complex term was proposed by Frantzi in 1997;

The measure is based upon the claim that a substring of a term candidate is a candidate itself given that it demonstrates adequate independence from the longer version it appears in.

For example, **E. coli food poisoning**, **E. coli and food poisoning** are acceptable as valid complex term candidates. However, **E. coli food** is not.

Therefore, some measures are required to gauge the strength of word combinations to decide whether two word sequences should be merged or not.

CValue (2)

$$Cvalue(a) = \begin{cases} \log_2 |a| \cdot f_a & \text{if } |a| = g \\ \log_2 |a| \cdot (f_a - \frac{\sum_{l \in L_a} f_l}{|L_a|}) & \text{otherwise} \end{cases}$$

where $|a|$ is the number of words in a, L_a is the set of longer term candidates that contain a, g is the longest n-gram considered, f_a is the frequency of occurrence of a, and $a \notin L_a$.

One can observe that longer candidates tend to gain higher weights due to the inclusion of $\log_2 |a|$.

TScore

The T-Score is used to measure the adhesion between two words in a corpus. It was defined by the following formula:

$$TS(w_i, w_j) = \frac{P(w_i, w_j) - P(w_i).P(w_j)}{\sqrt{\frac{P(w_i, w_j)}{N}}}$$

$P(w_i,w_j)$ is the probability of bi-gram w_iw_j in the corpus, $P(w)$ is the probability of word w in the corpus, and N is the total number of words in the corpus.

The adhesion is a type of unithood feature since it is used to evaluate the intrinsic strength between two words of a term.

Log-likelihood - LLR (1)

The general idea is related to comparison of weighted frequencies in the two corpora: if a word appears much more frequently in the domain corpus, it is a probable term.

Log-likelihood comparison gives better results than more traditional ***tf-idf*** based comparison

In our case we computed the frequencies of bi-grams.

As the general corpus we used a 22 million word excerpt from the **Corpus of Contemporary Serbian (SrpKor – http://www.korpus.matf.bg.ac.rs)**.

Log-likelihood – LLR (2)

The weight of bi-grams was calculated using the following formula:

$$G = 2 * \left(\left(freq_{domain} * \log \left(\frac{freq_{domain}}{freq\_Expected_{domain}} \right) \right) + \left(freq_{general} * \log \left(\frac{freq_{general}}{freq\_Expected_{general}} \right) \right) \right)$$

Where $freq_{domain}$ and $freq_{general}$ are real frequencies in the domain corpus and in the reference corpus.

Log-likelihood – LLR (3)

freq_Expected $_{domain}$ and ***freq_Expected*** $_{general}$ are expected frequencies in the domain corpus and in the reference corpus. They are calculated according to the following formulae:

$$freq\_Expected_{domain} = size_{domain} * \frac{freq_{domain} + freq_{general}}{size_{domain} + size_{general}}$$

$$freq\_Expected_{general} = size_{general} * \frac{freq_{domain} + freq_{general}}{size_{domain} + size_{general}}$$

As the result of application of this formula, all words in the domain corpus are assigned weights.

Keyness

A simple method for identifying keywords of one corpus vs another. That includes a variable which allow the user to focus on higher or lower frequency words.

The keyness score of a word is calculated according to the following formula:

$$\frac{fpm_{\text{focus}} + n}{fpm_{\text{ref}} + n}$$

where fpm_{focus} is the normalized (per million) frequency of the word in the focus corpus, fpm_{ref} is the normalized (per million) frequency of the word in the reference corpus, n is the simple math (smoothing) parameter ($n = 1$ is the default value).

Statistical measures for MWT

				Measures					Rank				
Graph	Num	Lemma	Eliminates lemma	Freq	CValue	TScore	LLR	Keyness	Freq	CValue	TScore	LLR	Keyness
grf01	sin	mineralna sirovina	mineralne sirovine	736	736	27129	5144	707	1	1	1	1	1
grf01	sin	površinski kop	površinski kopovi	305	230	17464	1925	161	2	4	2	3	17
grf01	sin	toplotni tok		258	249	16062	1803	335	4	3	4	4	3
grf01	sin	toplotna provodljivost	toplotne provodljivosti	236	222	15362	1649	312	5	5	5	5	4
grf03	sin	kvalitet uglja	kvaliteti uglja	297	289	17234	2076	375	3	2	3	2	2
grf03	sin	nivo buke	nivo buka;nivoi buke	118	118	10863	825	171	15	18	15	13	11
grf01	sin	površinska povreda	površinske povrede	116	104	10770	811	169	17	27	17	15	13
grf09a	sin	površinska povreda u predelu	površinske povrede u predelu	24	48	4899	168	39	272	90	272	262	223
grf05	sin	površinska povreda potkolenice	površinske povrede potkolenic	8	13	2828	56	14	1660	867	1660	1544	1521
grf01	sin	električna energija	električne energije	121	121	11000	128	4	13	15	13	383	1739
grf01	sin	električna provodljivost	električne provodljivosti	46	46	6782	321	72	97	94	97	87	80
grf03	sin	vibracija šake	vibracije šake	54	34	7348	377	83	74	171	74	66	62
grf04b	sin	vibracija šake ruke		54	63	7348	377	83	75	56	75	67	63
grf08b	sin	sindrom vibracija šake ruke	sindromi vibracija šake ruke	16	32	4000	112	26	539	190	539	514	496
grf04b	sin	sindrom vibracija šake	sindromi vibracija šake	16	0	4000	112	26	551	1690	550	496	489
grf04b	sin	merenje vibracija šake	merenja vibracija šake	15	0	3873	105	25	577	1691	575	557	53
grf08b	sin	merenje vibracija šake ruke	merenja vibracija šake ruke	15	30	3873	105	25	584	204	584	547	

Interpretation of statistical results

površinska povreda ‘surface injury’ was ranked lower by **CValue (27)** than by **Freq (17)** because it is part of two other MWUs with three components.

Consequently, **površinska povreda potkolenice** ‘surface lower leg injury’ was ranked higher by **CValue (867)** than by **Freq (1660)**, and as this term is more common in the mining corpus than in SrpKor, it was ranked higher by **LLR (1544)** and **Keyness (1521)** than by **Freq (1660)**.

The oposite example is **električna energija** ‘electrical energy’ that is more frequently mentioned in reference corpus (a lot of news articles), and thus in terms of domain specificity it is being pushed to the bottom of the list (**Keyness=1739**).

Nivo buke ’noise level’ is better ranked by **LLR (13)** and **Keyness (11)** compared to **Freq (15)** due to a multitude of books in the field of occupational safety in mines, and the fact that the term ‘noise level’ is characteristic for mining.

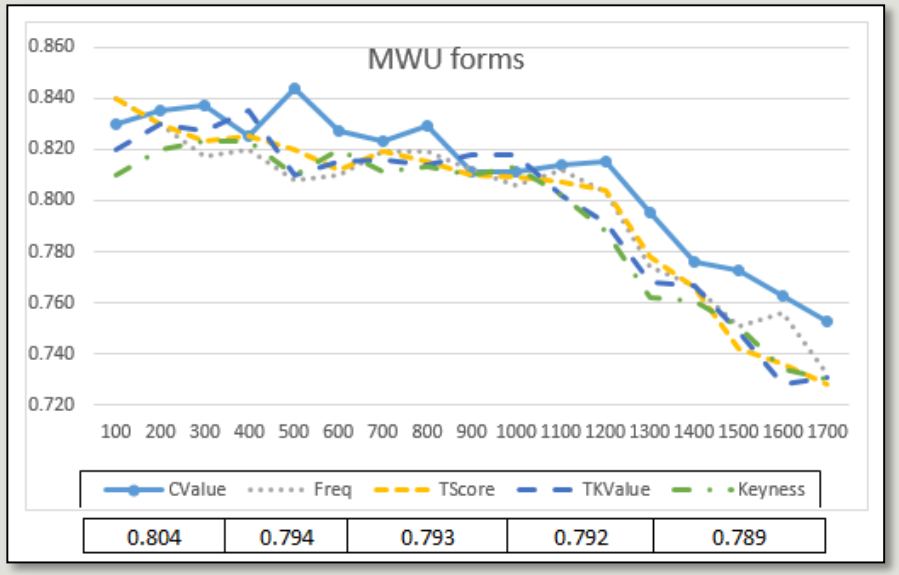
MWT extraction for mining – evaluation process

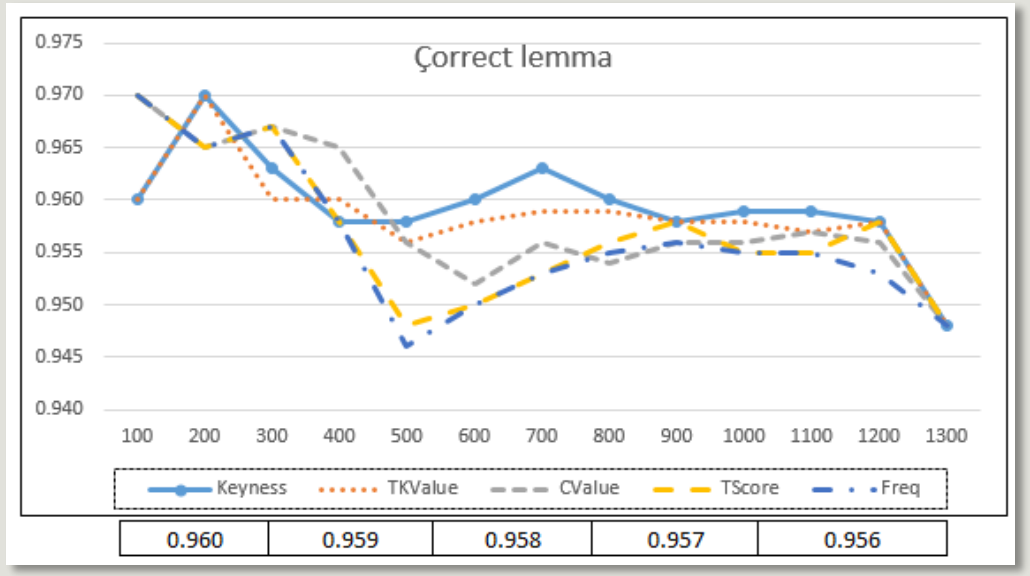
			MWU OK		Lemma OK	
		Graph	No	%	No	%
plu	grf01	AXN	42	89	8	19
	grf03	N2X	7	88	3	43
	grf04a	N4X	1	50		0
	grf04b	N4X	1	100		0
	grf05	AXN2X	3	100		0
	grf07	AXAXN	0	0		0
	grf08b	N6X	1	100		0
	grf10	2XAXN	1	100		0
sin	grf01	AXN	553	97	552	100
	grf02	2XN	6	100	5	83
	grf03	N2X	608	91	599	99
	grf04a	N4X	102	71	97	95
	grf04b	N4X	111	91	106	95
	grf05	AXN2X	58	92	58	100
	grf06	NXN	29	11	13	45
	grf07	AXAXN	4	80	4	100
	grf08a	N6X	5	83	5	100
	grf08b	N6X	6	100	5	83
	grf08c	N6X	7	78	7	100
	grf09a	AXN4X	10	59	10	100
	grf09b	AXN4X	7	88	7	100
	grf10	2XAXN	8	100	8	100
	Total		1570	83	1487	

Number of lemmas after manual evaluation

The evaluator performed two tasks: (a) checking for each lemma whether they actually represent a MWU, and (b) verifying for each proposed lemma whether it is a correct lemma

Evaluation results (1)

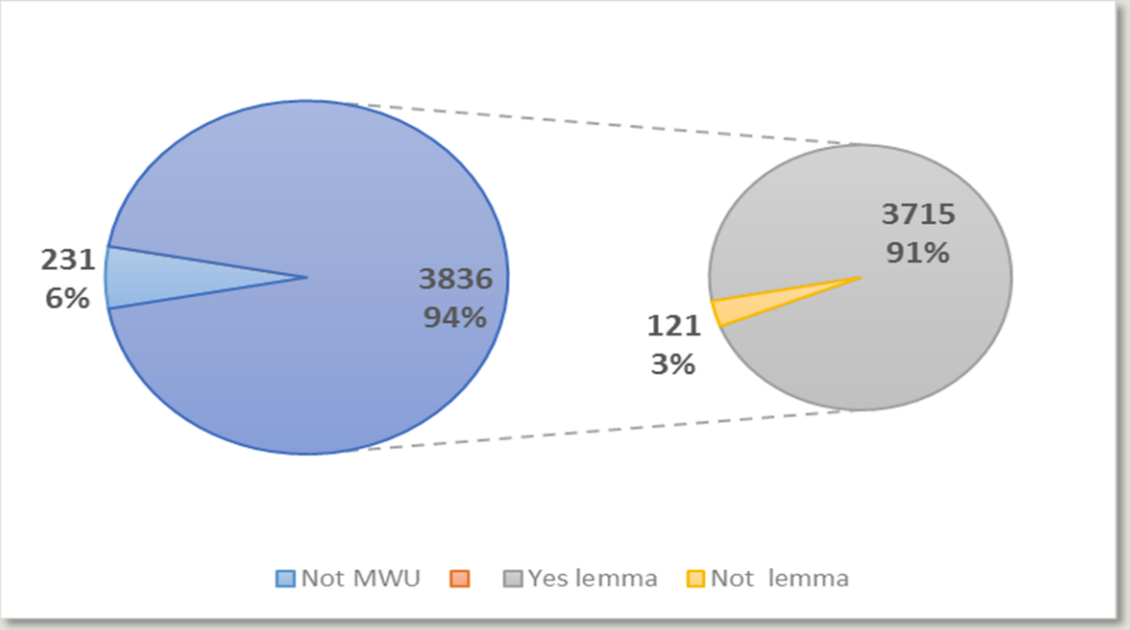




The precision of retrieval was calculated for each of these tasks for groups of hundreds ranked by measures:

Frequency, **C-Value**, **T-Score**, and **Keyness**. Mean average precision given at the bottom shows that all measures gave comparable results.

Evaluation results (2)



Out of 4067 distinct forms, **3836 (94%)** were evaluated as proper MWUs and **231 (6%)** were removed as notbeing proper MWUs.
Among proper MWUs there were **3715 (97%)** with a correct lemma and **121 (3%)** with an incorrect lemma.

Read more

K. Frantzi. 1997. Incorporating context information for the extraction of terms. *In Proceedings of the 35th Annual Meeting on Association for Computational Linguistics*. Spain.

C. Manning and H. Schuetze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press Cambridge, Massachusetts.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), pp.61--74.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography* 1(1), pp.7--36.

Conclusion and future work

Concluding remarks (1)

•Terminology extraction and lemmatization for Serbian	• based on e-dictionaries and local grammars • 14(+9) graphs were developed • most frequent syntactic structures
Evaluation of terminology extraction reports success	• first from library and information science • then from a mining corpus • other domains : urbanism, management, electric distribution
Automatic generation of a complete lemma for e-dictionary of Serbian	• a challenging task • previously not been tackled • **the most important contribution**

Concluding remarks (2)

When determining the lemma

- a number of possible candidates are generated
- using a system of rules and the frequency, incorrect suggestions are discarded automatically
- and less probable remain ranked

Evaluation of the results

- detailed
- performed manually
- presented by tables and graphs

Results

- Presented TE solution will speed up the development of e-dictionaries,
- Approach can be applied to the extraction of MWUs from general lexica
- Expanding the e-dictionaries will further improve systems for information retrieval, information extraction, query expansion etc.

Future Work

•Term extraction	Implementation	Development
• different domain corpora • **bilingual aligned corpora**	• different measures for ranking • **Annotate named entities before term extraction**	• new extraction FSTs for syntactic structures of MWTs • **additional strategies to avoid offering of incorrect lemmas**

Finalization of the web application