University of Belgrade, Faculty of Mining and Geology

Serbian Unitex Day

Extraction of Bilingual Terminology

using Graphs, Dictionaries and GIZA++

Ranka Stanković
ranka@rgf.rs

Branislava Šandrih
branislava.sandrih@fil.bg.ac.rs

March 11, 2019

Outline

An Approach for Bilingual Terminology Extraction
- Introduction
- An approach

Unitex: Deus ex machina
- Extraction System
- Results

BilTE Web Application
- System Overview
- Results
- Practical Demo

Outline

Introduction
Motivation

- Rapid development of terminology in many research and technological fields

Introduction
Motivation

- Rapid development of terminology in many research and technological fields
- Challenge is to produce and maintain up-to-date terminology resources for under-resourced language such as is Serbian

Introduction
Motivation

- Rapid development of terminology in many research and technological fields
- Challenge is to produce and maintain up-to-date terminology resources for under-resourced language such as is Serbian
- Can existing NLP resources, methods and tools for Serbian help in the automatic compilation of bilingual terminology lexicon?

Introduction
Unitex

- aligned corpora;

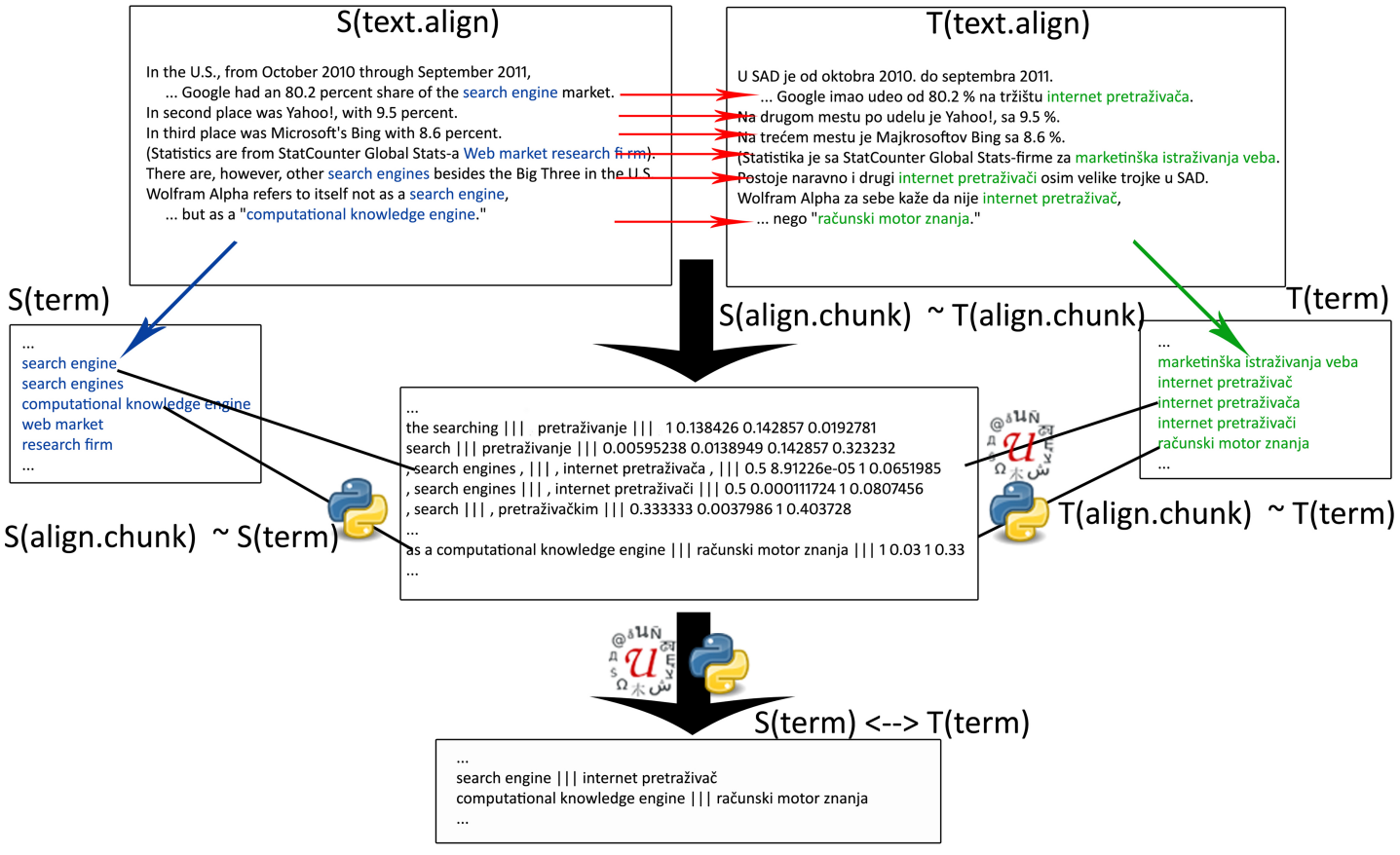
Introduction
Unitex

- aligned corpora;
- electronic dictionaries of simple and multi-word units for Unitex;

Introduction
Unitex

- aligned corpora;
- electronic dictionaries of simple and multi-word units for Unitex;
- shallow parsers for Unitex.

An approach
System Workflow



Outline

Unitex: Deus ex machina
MWT Extraction

Hybrid approach is proposed for MWT extraction:

MWUs extraction based on linguistic rules - syntactic patterns using a cascade of Unitex transducers to parse documents and retrieve candidate terms and statistical measures, mainly for filtering

Unitex: Deus ex machina

MWT Extraction

Hybrid approach is proposed for MWT extraction:

MWUs extraction based on linguistic rules - syntactic patterns using a cascade of Unitex transducers to parse documents and retrieve candidate terms and statistical measures, mainly for filtering

MWT candidates are extracted from texts in various inflected forms, lemmatization of extracted MWT candidates, linking to one normalized or head-word form, important for highly-inflected languages, such as Serbian and other Slavic languages

Unitex: Deus ex machina

Methodology and Design: lexical resources

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

Unitex: Deus ex machina

Methodology and Design: lexical resources

- MWUs are classified according to their syntactic structure and inflectional and other properties;
- Class names correspond to FSTs used for inflection of MWUs belonging to that class;

Unitex: Deus ex machina

Methodology and Design: lexical resources

- MWUs are classified according to their syntactic structure and inflectional and other properties;
- Class names correspond to FSTs used for inflection of MWUs belonging to that class;
- For example, MWUs composed of a noun (**N**) followed by an adjective (**A**), which agrees with a noun in gender, number, case and animateness, belong to the **AXN** class;

Unitex: Deus ex machina

Methodology and Design: lexical resources

- MWUs are classified according to their syntactic structure and inflectional and other properties;
- Class names correspond to FSTs used for inflection of MWUs belonging to that class;
- For example, MWUs composed of a noun (**N**) followed 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.

Unitex: Deus ex machina

Nominal MWUs in Serbian

- 17100 nominal MWUs +1000 other MWUs;
- 14 classes account for more than 98% of all nominal MWUs.

Unitex: Deus ex machina

Number of components

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

Unitex: Deus ex machina

Rank	MWU Class	Nubee of MWI	%	Cumulative %
1	NC_AXN	10556	61.7	61.7
2	NC_2XN	1405	8.2	69.9
3	NC_N2X	1310	7.7	77.6
4	NC_N4X	1242	7.3	84.9
5	NC_AXAXN	557	3.3	88.1
6	NC_AXN2X	439	2.6	90.7
7	NC_NXN	396	2.3	93.0
8	NC_N6X	352	2.1	95.1
9	NC_AXN4X	173	1.0	96.1
10	NC_2XAXN	104	0.6	96.7
11	NC_N8X	93	0.5	97.2
12	NC_AXN6X	78	0.5	97.7
13	NC_4XN	54	0.3	98.0
14	NC_NXA	40	0.2	98.2
15	NC_N10X	34	0.2	98.4
16	NC_AXN8X	32	0.2	98.6
17	NC_A3XN	30	0.2	98.8
18	NC_N3XN	27	0.2	98.9
19	NC_AXNXN	27	0.2	99.1
20	NC_AXAXAXN	19	0.1	99.2
21	NC_AXAXN4X	18	0.1	99.3
22	NC_AXAXNXN	13	0.1	99.4
23	NC_AXAXN2X	11	0.1	99.5
24	NC_NXAXN	10	0.1	99.5
25	NC_2XN4	8	0.0	99.6

AXN – agree in all four grammatical categories

bistar um $\longleftrightarrow$ clear mind

Unitex: Deus ex machina

Methodology and Design: FST graph extraction types

AXN – agree in all four grammatical categories

bistar um $\longleftrightarrow$ clear mind

N2X – usually noun in genitive or instrumental case

igra reči $\longleftrightarrow$ wordplay

N4X – a noun followed by two words that do not inflect in the MWU

NprepNp učenje na daljinu $\longleftrightarrow$ distant learning (noun followed by a prepositional phrase)

NNgiNgi softver otvorenog koda $\longleftrightarrow$ open source software (noun followed by two adjectives/nouns in the genitive/instrumental case)

NXN – agrees with it in number and case

veb stranica $\longleftrightarrow$ web page

NXN – agrees with it in number and case

veb stranica $\longleftrightarrow$ web page

N6X - a noun followed by three words that do not inflect in the MWU

NNgiPrepNp konverzija teksta u govor $\longleftrightarrow$ text-to-speech conversion

NNgiNgiNgi cifra arapskog brojnog sistema $\longleftrightarrow$ digit of Arabic number system

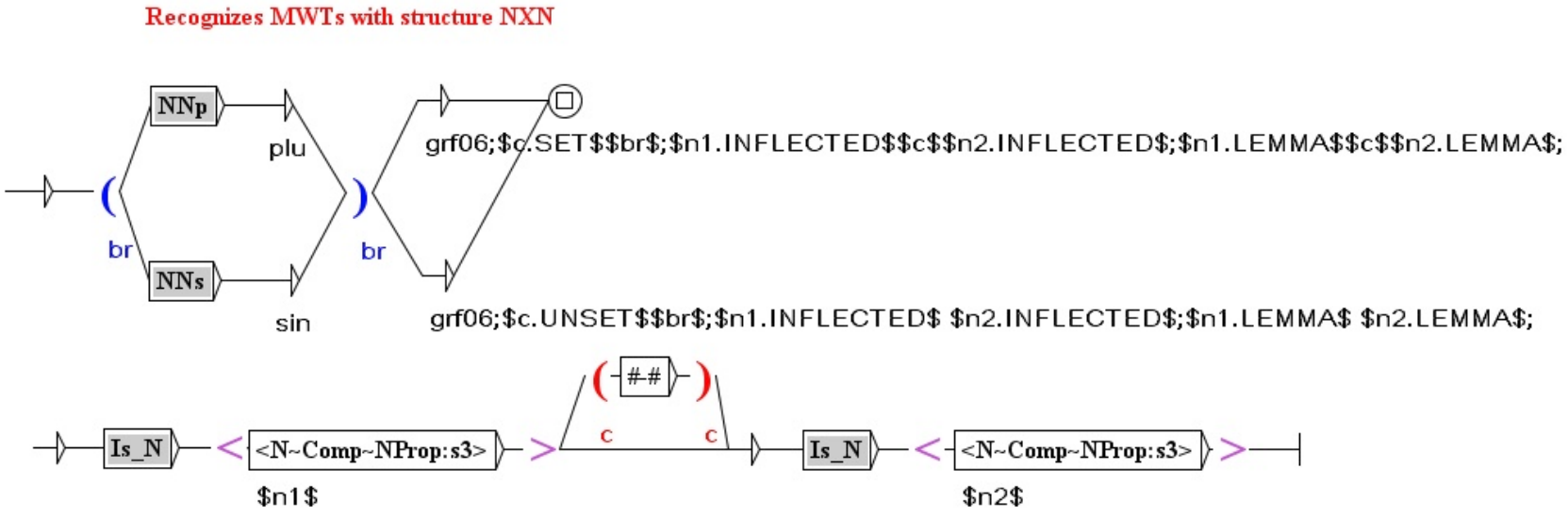
Unitex: Deus ex machina

FST for extraction of type NXN

- FST graph for extraction of NXN type MWUs: Two subgraphs, NNp and NNs, recognize possible singular and plural forms
- Dictionary variables $n1.LEMMA$ and $n2.LEMMA$ at FST output perform normalization, that is, simple word lemmatization by retrieving lemmas for the recognized word forms $n1$ and $n2$

Unitex: Deus ex machina

FST for extraction of type NXN



kupus ribanac, kupusa ribanca, kupusu ribancu, kupuse ribanče, kupusom ribancem, kupusi ribanci, kupusima ribancima, kupusima ribancima, . . .

Unitex: Deus ex machina

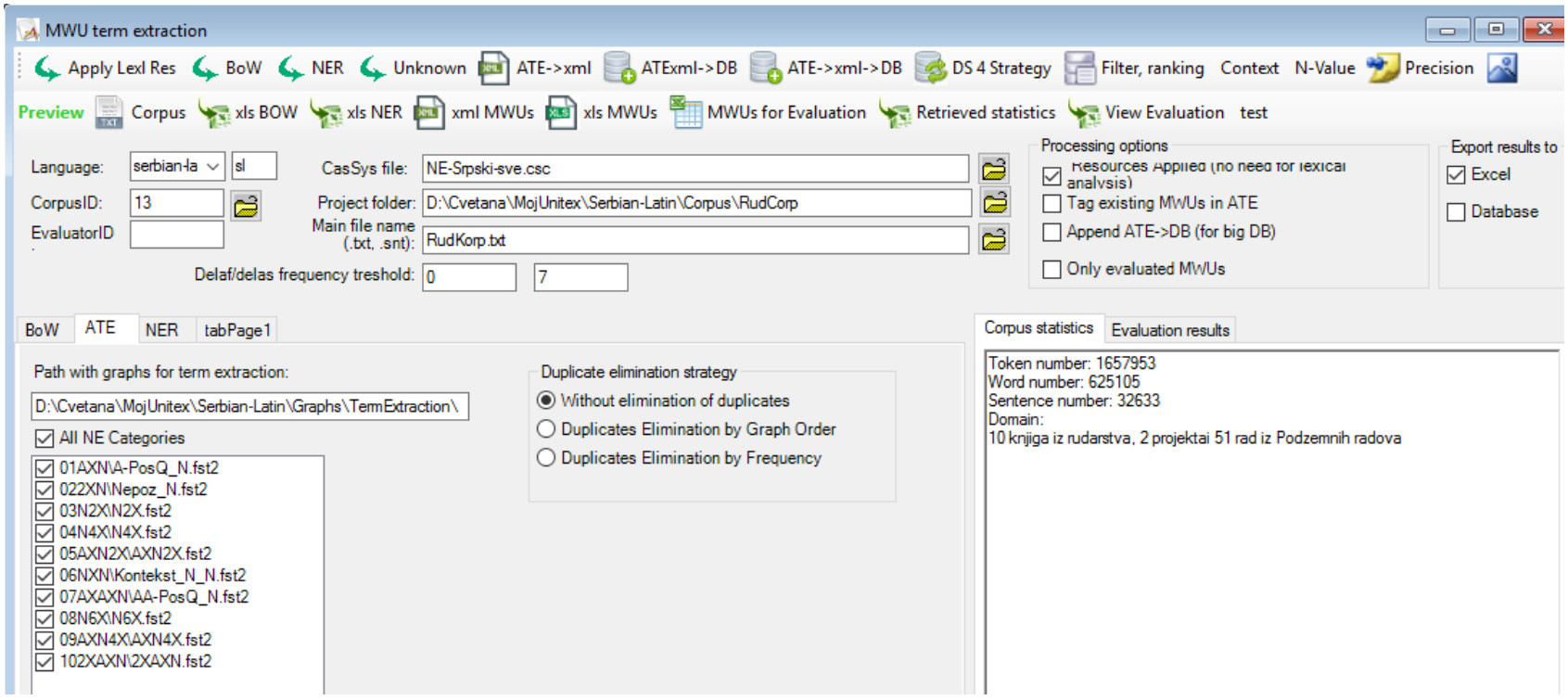
Features of Software solution for MWTE

22 Unitex graphs, grouped in 10 categories integrated into LeXimir with metadata corpus management for specific domains (mathematics, geology, energetics, library science, . . .)

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

Unitex: Deus ex machina

Software solution for MWTE



Results

The results of the term extraction system

- 8 different experiments conducted, 8 parameters:
  - term extraction on the source side (2 cases);

Results

The results of the term extraction system

- 8 different experiments conducted, 8 parameters:
  - term extraction on the source side (2 cases);
  - raw and augmented data (2 cases);

Results

The results of the term extraction system

- 8 different experiments conducted, 8 parameters:
  - term extraction on the source side (2 cases);
  - raw and augmented data (2 cases);
  - term extraction on the target side (2 cases);

Results

Number of pairs obtained after alignment with GIZA++ and initial filtering

Compilation of English terms

using dictionary entries:

- on raw data, $\sim$ 240,000 pairs;
- on augmented data, $\sim$ 215,000 pairs.

using term-extractor:

- on raw data, $\sim$ 497,000 pairs;
- on augmented data, $\sim$ 447,000 pairs.

Results
Number of Serbian MWTs extracted

Compilation of English terms using dictionary entries

- on raw data
  - extracting MWTs from the phrase-table, $\sim 27,000$
  - extracting MWTs from the text: $\sim 50,000$
- on augmented data
  - extracting MWTs from the phrase-table, $\sim 46,000$
  - extracting MWTs from the text: $\sim 51,000$

Results
Number of Serbian MWTs extracted

Compilation of English terms using term-extractor

- on raw data
  - extracting MWTs from the phrase-table, $\sim 35,000$
  - extracting MWTs from the text: $\sim 50,000$
- on augmented data
  - extracting MWTs from the phrase-table, $\sim 45,000$
  - extracting MWTs from the text: $\sim 51,000$

Results
Number of obtained candidate pairs

Compilation of English terms using dictionary entries

- on raw data
  - extracting MWTs from the phrase-table, ~ 800
  - extracting MWTs from the text: $\sim 1,000$
- on augmented data
  - extracting MWTs from the phrase-table, $\sim 1,400$
  - extracting MWTs from the text: $\sim 1,400$

Results
Number of obtained candidate pairs

Compilation of English terms using term-extractor

- on raw data
  - extracting MWTs from the phrase-table, $\sim 2,100$
  - extracting MWTs from the text: $\sim 2,000$
- on augmented data
  - extracting MWTs from the phrase-table, $\sim 3,000$
  - extracting MWTs from the text: $\sim 3,000$

Outline

Upload two sentence-aligned text files.

Files should have same names, but the extension should differ (e.g. medicine.en and medicine.sr). These files are later fed into GIZA++

```
10  Sistem uzajamne katalogizacije uspostavljen je još 1988. godine i pokrivao je celokupnu teritoriju tadašnje Jugoslavije.
11  Autori projekta i programa su bili stručnjaci Računarskog centra Univerziteta u Mariboru (danas IZUM - Institut informacijskih znanosti)
    koji su izabrani na tenderu tadašnjeg Saveznog ministarstva za nauku.
12  Kao projekat koji je finansirala država, program je prihvaćen od strane Zajednice jugoslovenskih nacionalnih biblioteka i u njegovom sastavu
    funkcionisalo je 55 biblioteka iz svih republika.
13  Računarska tehnologija 80-ih godina omogućila je umrežavanje biblioteka i razmenu podataka, zasnovanu na principu jednog centralnog računara
    - servera (smeštenog u Mariboru) koji je preko postojeće telekomunikacione mreže bio povezan sa lokalnim računarima, na kojima su se
    nalazile baze podataka pojedinih biblioteka ili grupa biblioteka.
```

```
10  The system of shared cataloguing was established in 1988, and it was covering the whole territory of Ex-Yugoslavia.
11  Authors of the project and software were professionals from the University Maribor Computer Center (today IZUM-Institute of
    information science) and they were elected on the tender of the former Federal ministry of science.
12  As a project financed by the Federal government, it was accepted by the Yugoslav National Library Association, and it
    consisted of 55 libraries of all republics.
13  Computer equipment in the `80-s enabled shared cataloguing using one central computer - server (located in Maribor) witch
    was connected through existing telecommunication network for data transfer to local computers, where databases of
    participating libraries where located.
```

Upload a List of English terms

First line is a header, each line contains one term

```
eng
full name of person
historical collection
dictionary enhancement
education system
morphological dictionary
newspaper text
co-ordinating library
parent library
special library within state body
total number of citation
language variety
cambridge university
stem class
grammatical category
optimal suffix stemmer
classification method
civil engineering
web frontend
broadest sense
batthyaneum branch
```

Upload a List of Serbian Extracted MWUs

First line is a header, each line contains a MWU and its frequency, separated with | (“pipe” character)

```
Lema|Freq
slučaj sa jezik|2
onlajn sudijski sistem|5
evropski identitet međa građanin|1
aplikacija za jezički resurs|2
različit klasa|3
nov program|3
određen različitost|1
škola grad|4
opasan sadržaj|2
isključivanje objekt|8
broj primer|4
usluga na brod|7
profesionalan savetodavan usluga|1
autorov analiza|3
određen autor|5
zagovornik biblioteka|2
hiljada član|8
dostupan stemera|10
izvor biblioteka|14
```

System Overview
Processing

1st step

Run GIZA++ on aligned sentences

System Overview
Processing

1st step

Run GIZA++ on aligned sentences

2nd step

Filtering: discard out-of-the-domain

System Overview
Processing

1st step

Run GIZA++ on aligned sentences

2nd step

Filtering: discard out-of-the-domain

3rd step

Lemmatization: English chunks with WordNet and Serbian chunks with Unitex

System Overview
Results

1st step

Keep only candidates present in English list

System Overview
Results

1st step

Keep only candidates present in English list

2nd step

Retrieve intersection with Serbian extracted MWUs

System Overview
Results

1st step

Keep only candidates present in English list

2nd step

Retrieve intersection with Serbian extracted MWUs

3rd step

Additional filtering: eliminate bad candidates from the previous step

Results

Experimental results on INFOtheca

http://bilteresults.jerteh.rs/

	A	B	C	D	E
1	EXTRACTED_SR	DICTIONARY_EN	EVAL	GIZA_SR_ORIGINAL	GIZA_SR_LMTZ
2	aktivan činilac doprinos	active contributor	OK	aktivni činioci doprinosa	oni su aktivan činilac doprinos
3	administrativan osoblje	administration	OK	administrativnim osobljem	administrativan osoblje
4	prateći dokumentacija	accompanying documentation	OK	i prateću dokumentaciju	i prateći dokumentacija u
5	model saradnja	model of cooperation	T	modela saradnje	model saradnja
6	algoritam za konflaciju	algorithm for conflation	T	algoritma za konflaciju	algoritam za konflaciju
7	morfološki elektronski rečnik	electronic morphological dictionary	T	morfološkog elektronskog rečnika	elektronski morfološki rečnik
8	akademski biblioteka	academic library	LIS	i akademske biblioteke se	biblioteka u akademski
9	akademski izdavač	academic publisher	LIS	akademski izdavači su	akademski izdavač
10	akademski zajednica	academic community	LIS	akademsku zajednicu	koji naš akademski zajednica
11	alat informacija	information tool	NOK	alatima informacije	alat informacija
12	akcent glagol	accent	NOK	akcenta glagola	akcent glagol
13	alat za jezik	language tool	NOK	jezik već alat	alat za taj jezik
14	mreža reč za drugi jezik	wordnet of other language	X	mreža reči za druge jezike	mreža reč za drugi jezik
15	nacionalan biblioteka republika	national library of republic	X	nacionalnoj biblioteci republike	nacionalan biblioteka republika
16	naučni istraživački rad	scientific work	X	naučno istraživačkog rada	naučni istraživački rad

Results
Annotations

LIS	the extracted pair is correct and the extracted terms belong to the LIS domain
T	the extracted pair is correct and the extracted terms belong to some other domain
OK	the extracted pair contains translational equivalents, but does not represent a term
NOK	the extracted pair does not contain translational equivalents
X	if English term extractor extracted neither a term nor a complete noun phrase

Practical Demo

http://bilte.jerteh.rs/

Watch demo on:

https://youtu.be/mIJjycrY5ZE

Thank you for your attention!