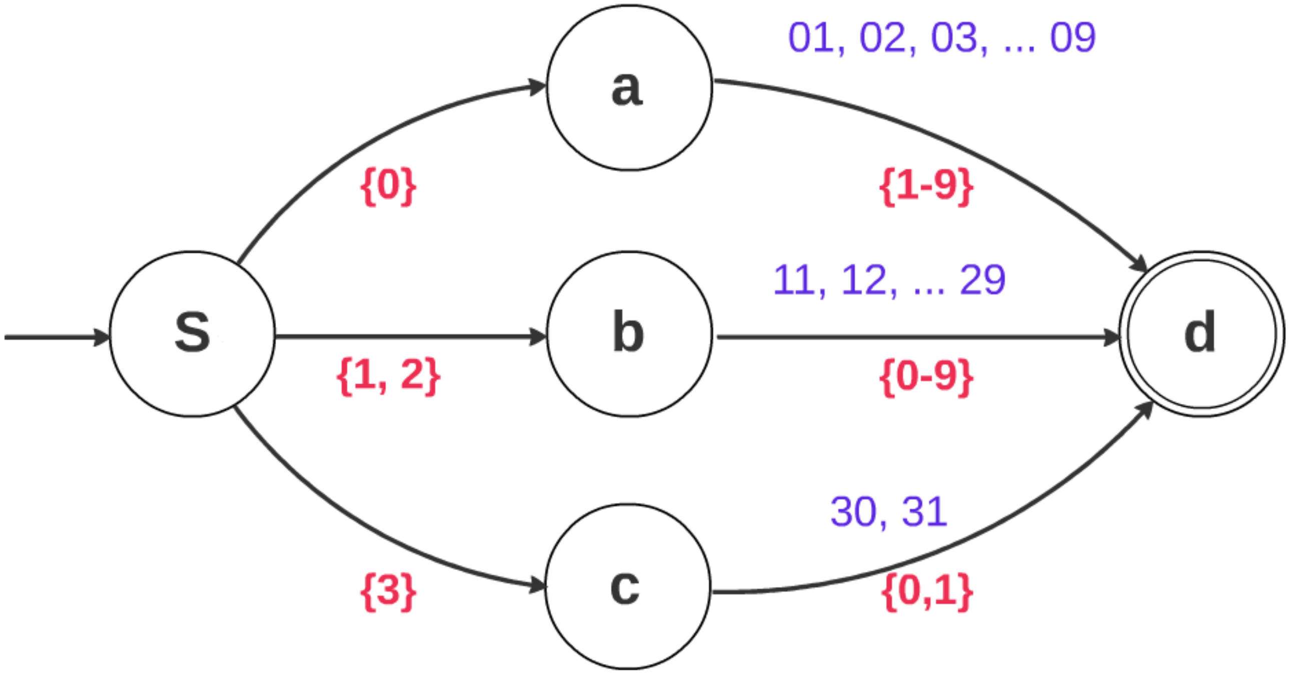
Serbian Unitex Day

Serbian NER

Dictionaries, Graphs and Cascades



Jelena Jaćimović

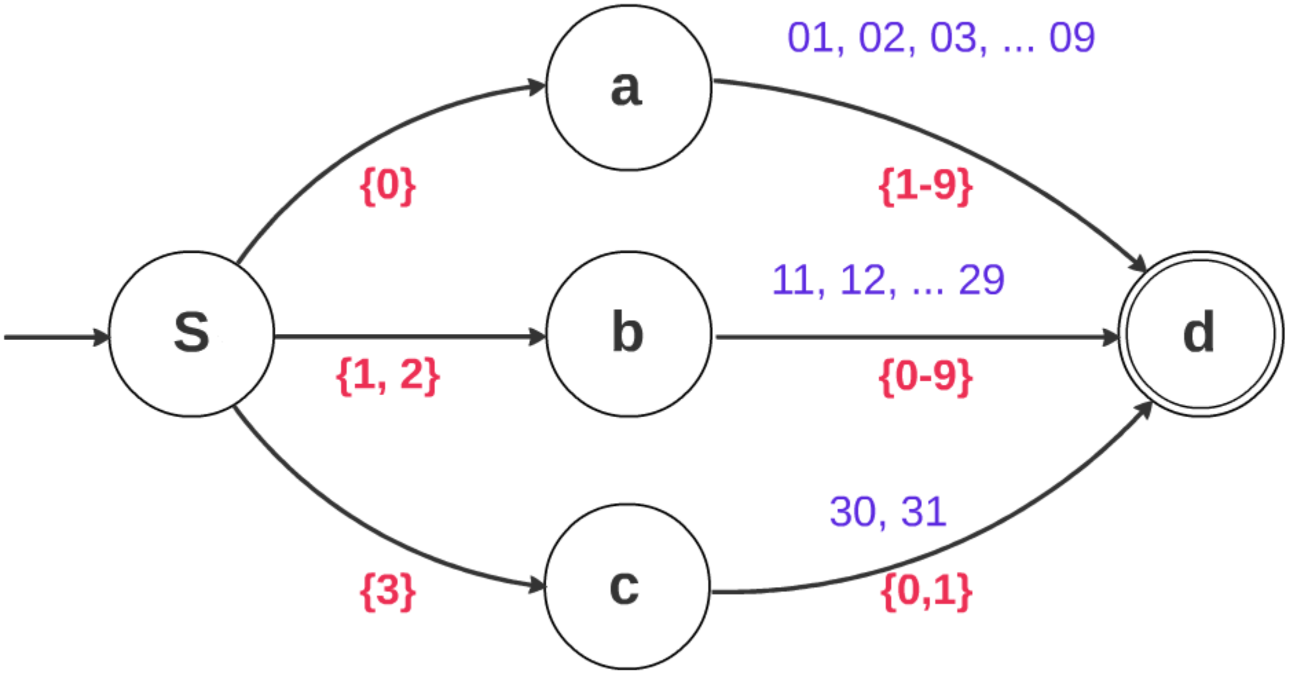
University of Belgrade, Faculty of Mining and Geology

March 11, 2019

Serbian Unitex Day

Serbian NER

Dictionaries, Graphs and Cascades



Jelena Jaćimović

University of Belgrade, Faculty of Mining and Geology

March 11, 2019

NER for Serbian

Work on NER for Serbian started in 2006 with recognition of personal names

In 2010 the system was recognizing: full personal names, temporal expressions, amounts, subset of geopolitical names

Evaluation results were very good:

- **Precision 0.98**
- **Recall 0.94**
- **F-measure 0.96**

In 2012 NER system for Serbian was transformed from the collection of transducers to the cascade of transducers

New entities were added

What we are recognizing now?

- numerical expressions
- temporal expressions
- name expressions

numerical expressions

- currencies
  50 USD, 200 €
- measurement expressions
  10 kg, pet metara 'five meters'
- percentage
  35%, 20 procenata '20 percent'
- amounts
  jednom 'once'

temporal expressions

- date
  25.12.2018, marta 2007.
- time
  jutros u pet, oko 8 h
- duration
  nekoliko meseci, 2 nedelje
- sets of recurring times
  šest puta nedeljno, svakog dana

name expressions

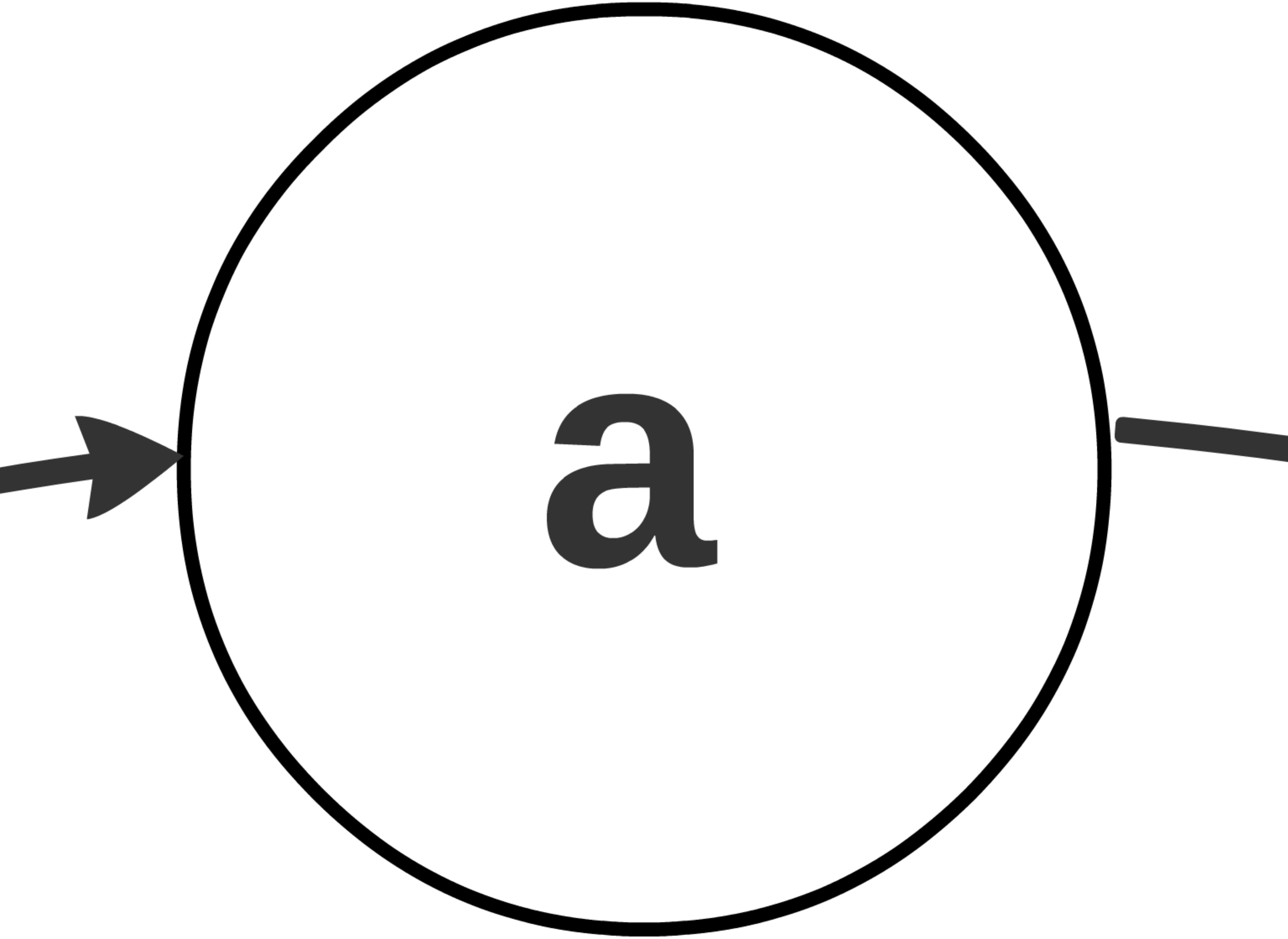
- geopolitical names
  - settlements (*Beograd* ‘Belgrade’)
  - countries (*Nemačka* ‘Germany’)
  - hydronyms (*Dunav* ‘Danube’, *Jadransko more* ‘Adriatic Sea’)
  - oronyms (*Alpi* ‘Alpes’)
  - regions (*Vojvodina*)
  - municipalities (*Savski venac*)
- personal names
  - full names
  - first names
  - surnames,
  - nicknames
  - names of distinguished persons
  - (functions and roles of recognized personal names)
- organizations (general and political)
- urban names

What we are using for NER?

e-dictionaries

graphs

cascades


a

Dictionaries

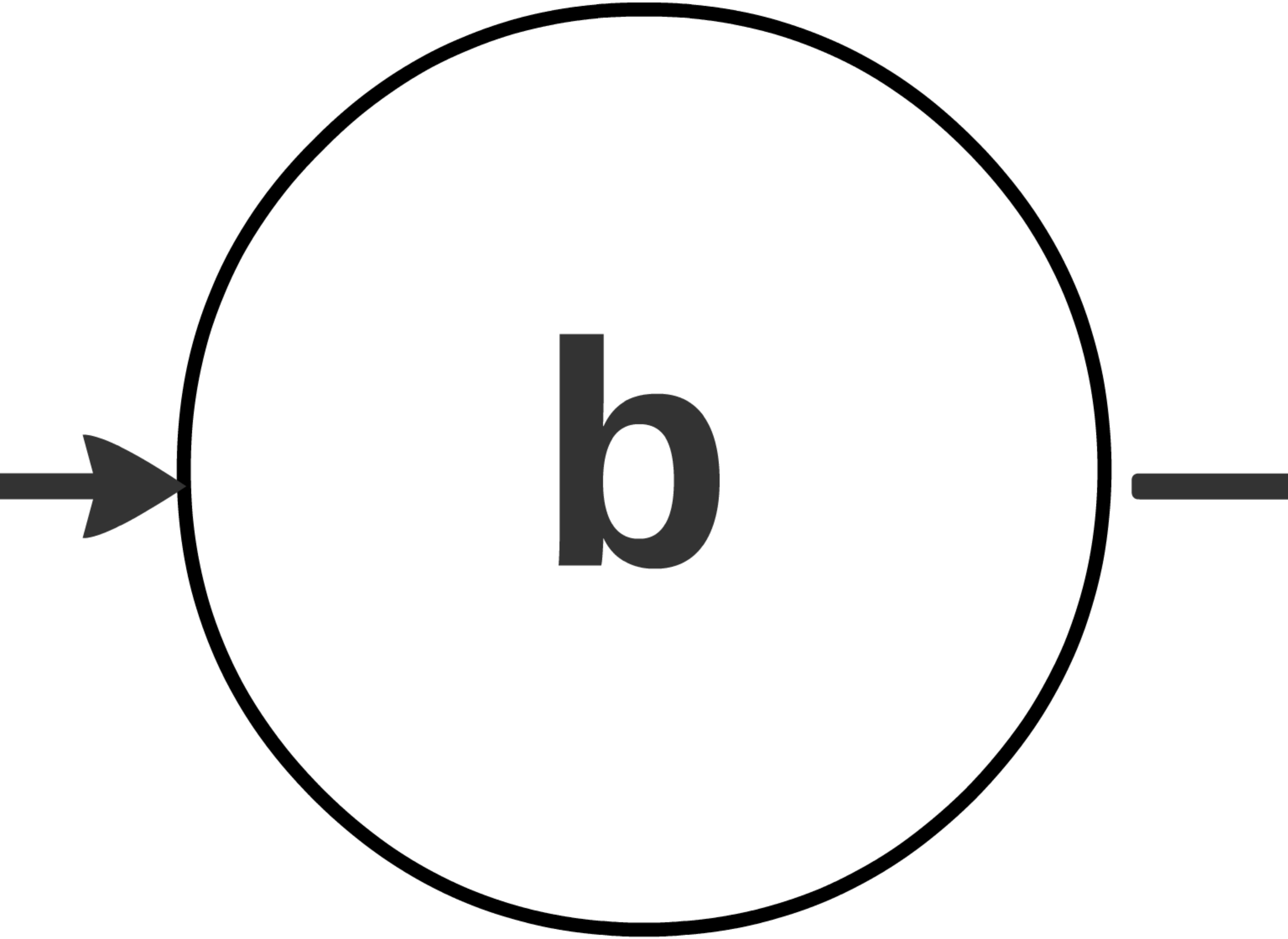
Comprehensive morphological e-dictionaries of Serbian in DELA/DELAF format:

- simple words
- compounds
- general lexica
- geographic names
- personal names

word form ➡ possible values of grammatical categories (case, number, gender, etc.) and lemma together with various additional markers that specify derivational, syntactic, semantic or usage features of lemma

Alena, Alen.N+NProp+Hum+First+EN+Val=Allen:ms2v

Beograđanina, Beograđanin. N+Hum+NProp+Inh+CC2=RS:ms2v


b

Graphs

Dictionary graphs / FSTs

recognition and morphosyntactic tagging of open classes of simple words and compounds generally not found in dictionaries

simple words → Roman numerals, interjections with repetition of one or more graphemes (e.g. *jaoooo* 'ouuuuch') and acronyms which are not generally known and regularly used

numerals → written with digits, words and their combinations

2,52 milijarde ‘2.52 billions’

compound nouns, adjectives and adverbs → derived from numerals and written with digits

adjective *18-dnevni* ‘18 days long’

21-godišnji, 21-godišnji.A+PosQ+C:adms1g:aems4q:aemp1g

Graphs

Transducers

- capture various structures of named entities, and
- disambiguate

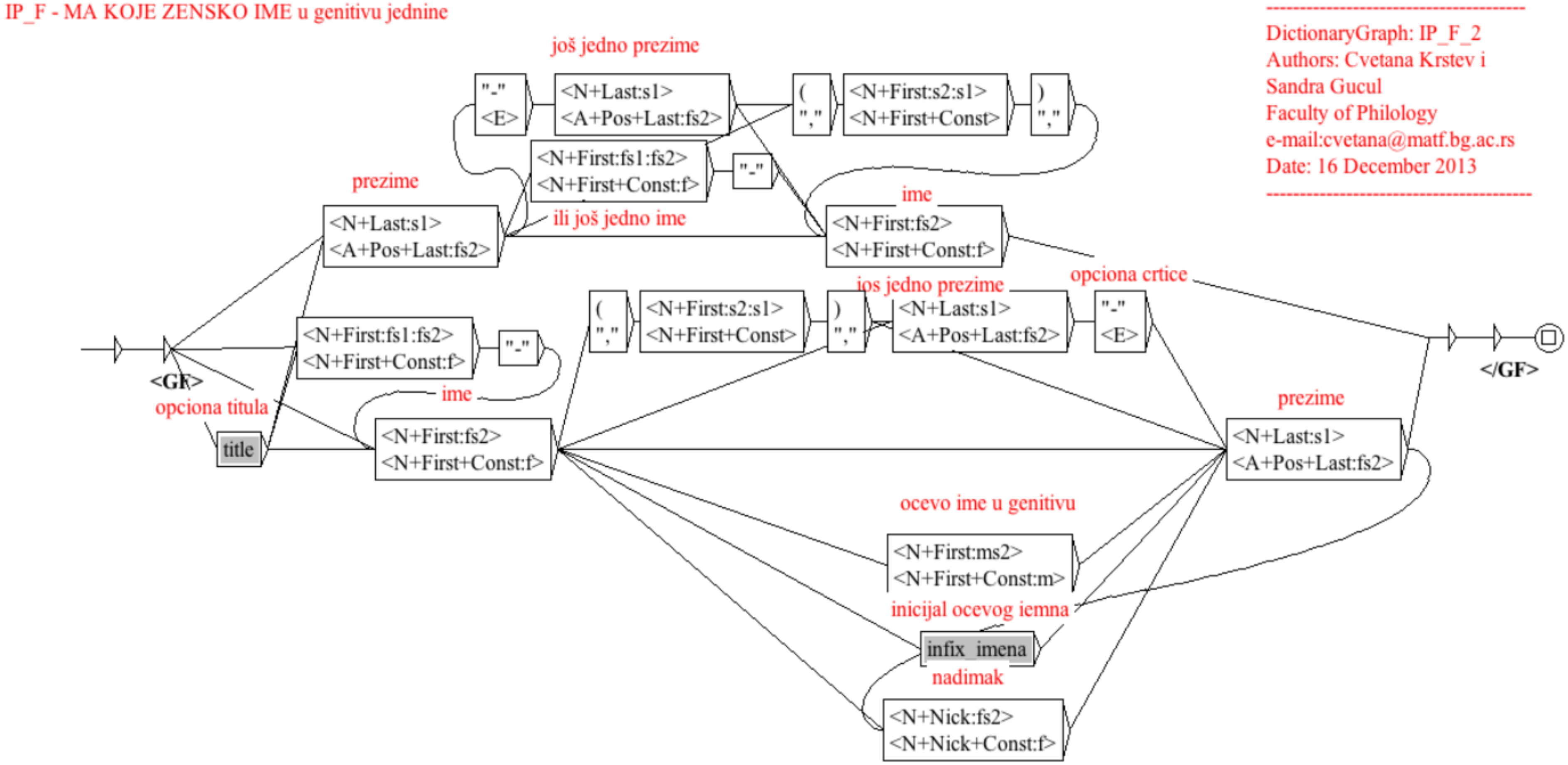
They rely on:

- general and specific e-dictionaries and
- semantic markers assigned to dictionary entries

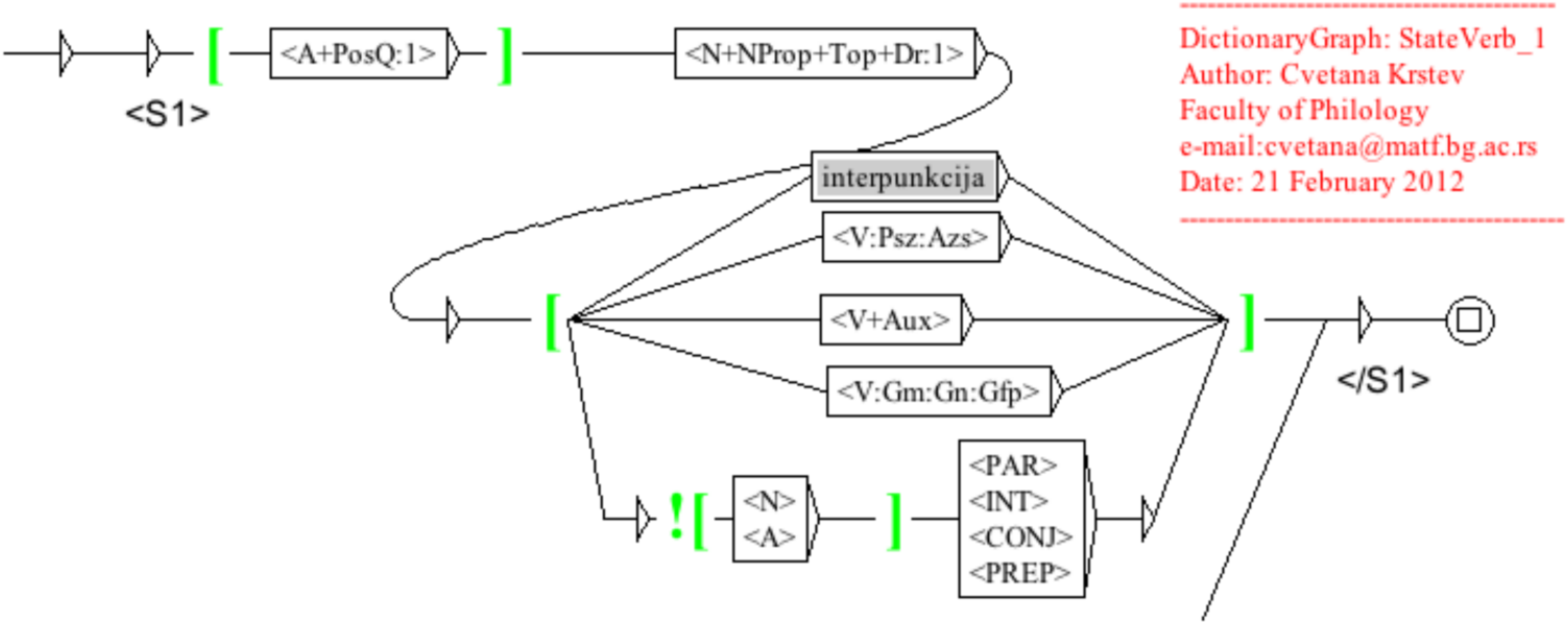
+Cur (currencies),
+Mes (measurement units),
+First (first names),
+Last (surnames),
+Nick (nicknames),

+Dr (states),
+Gr (settlements),
+Hyd (water bodies),
+Oro (elevations),
+Reg (regions), etc.

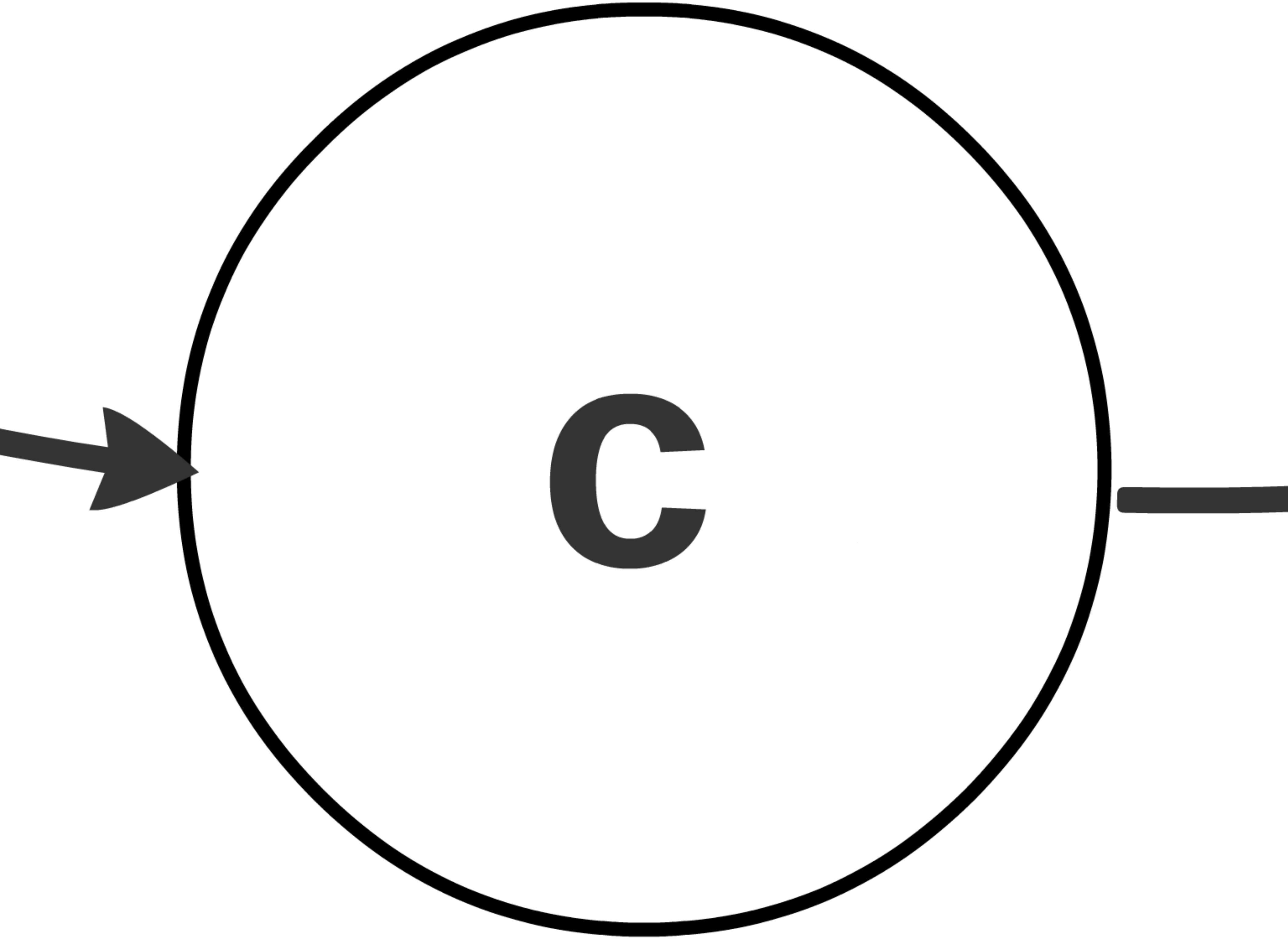
Various structures of full feminine names (the gentive case)



This graph recognizes as a name of a country form that coincide with the feminine gender relational adjectives if followed by a punctuation sign or an auxiliary verb



Hrvatska ('Croatia/Croatian'), *Francuska* ('France/French')

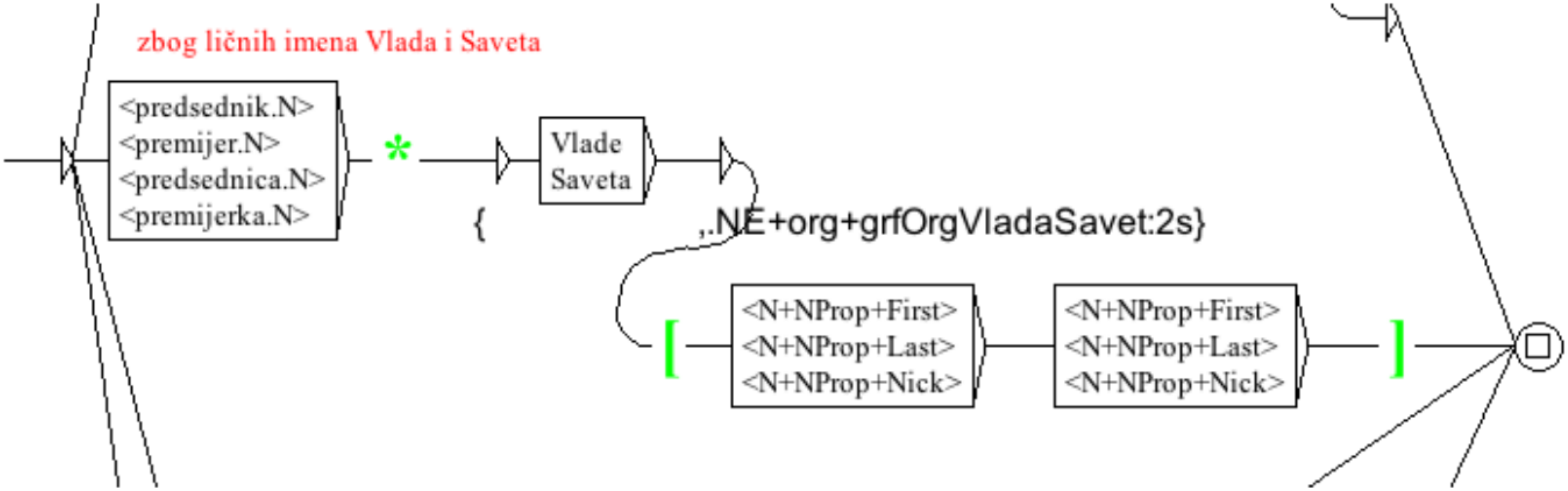

C

Cascade of transducers

- addressesID (phone number, url, email, twitter; personal citizens data - passport, ID card, index number, etc.) 1
- Foreign organization names (2)
- orgVlada Savet 1
- Amounts – basic and approx (2)
- Temporal expressions - date abs and rel (3), date period 1, set 1, duration 2, time 3, time period 1, duration basic and relative 1+1, date rel 1, duration period and rel date 1, time complex 1 (16)
- Amounts – approximation, basic and range (3)
- Full names – masculine (6), feminine (6) +2
- Amounts – approximation 1
- Geopolitical names – MWUs (6+1), unambiguous (1), simple (6+1), lists (6+1)
- Organizations (6+1)
- Distinguished persons (1)
- Lists of names with roles (6+1), full names with roles – masculine (6), feminine (6) +1, unknown names with roles - masculine (6), feminine (6) +1
- Surnames (7+1)
- Surnames list (6+1)
- Personal name 1
- Streets 1
- Masculine first name, surname and occupation 1

Vlade 'of the Government' / e.g. personal masculine first name Vlada in genitive case

Saveta 'of the Council' / e.g. personal feminine first name Saveta in nominative case



zbog ličnih imena Vlada i Saveta

<predsednik.N>
<premijer.N>
<predsednica.N>
<premijerka.N>

*

{

Vlade
Saveta

,.NE+org+grfOrgVladaSavet:2s}

[

<N+NProp+First>
<N+NProp+Last>
<N+NProp+Nick>

<N+NProp+First>
<N+NProp+Last>
<N+NProp+Nick>

]

predsednik Vlade Mirko Cvetković 'Prime Minister Mirko Cvetković'

predsednik {Vlade.NE+org:2s} Mirko Cvetković

Thank you!