UNITEX for NER

Cvetana Krstev, JeRTeH & University of Belgrade

Denis Maurel, Université de Tours, Lifat

Outline of this presentation

- Why Unitex for NER?
- Why using dictionaries?
- Why using graphs?
- What can graphs in Unitex do?
- Why using cascades?

Why using Unitex for NER?

- Because it enables production of lexicon-based and rule-based systems;
- It enables writing rules that rely on information from dictionaries
  - These rules are in the form of finite-state transducers (even ATN);
  - Sometimes called „local grammars“ or „shallow parser“;
- Advantages:
  - These rules are not „black boxes“;
  - Lexicon-based and rule-based systems are unavoidable for preparation of training sets for machine-learning systems.

What is in dictionaries that can help NER?

- **Jeanne d'Arc,N+PR+Anthroponyme+Individuel+Celebrite**
  - a noun (N)
  - a proper name (PR)
  - an anthroponyme, a proper name of a human being (Anthroponyme)
  - an individual proper name (Individuel)
  - a famous person (Celebrite)
- **Jovanka Orleanka,N+NProp+Hum+Name+Cel+DOM=Hist**
  - a noun (N)
  - a proper name (NProp)
  - a human being (Hum)
  - a name given to a human being (Name)
  - a famous person (Cel)
  - a historical person (DOM=Hist)

What is in dictionaries that can help NER?

- **Bab el=Mandeb,N+PR+Toponyme+Hydronyme**
  - a noun (N)
  - a proper name (PR)
  - a toponym (Toponyme)
  - a hydronym (Hydronyme)
- **Bab-el-Mandeb,N+NProp+Top+Hyd+Strait+Cel**
  - a noun (N)
  - a proper name (NProp)
  - a toponym (Top)
  - a hydronym (Hyd)
  - a strait (Strait)

What is in dictionaries that can help NER?

- **Banque centrale européenne,N+PR+Anthroponyme+Collectif+Groupement+Organisation**
  - a noun (N)
  - a proper name (PR)
  - an anthroponyme, a proper name of a human being (Anthroponyme)
  - a collective name (Collectif)
  - a group (Groupement)
  - an organisation (Organisation)
- **Evropska centralna banka,N+NProp+Org+DOM=Fin+ACR=ECB**
  - a noun (N)
  - a proper name (NProp)
  - an organisation (Org)
  - in a domain of finances (DOM=Fin)
  - its acronym is ECB (ACR=ECB)

What is in dictionaries that can help NER?

proper names that are not related to one individual

- **Eric,N+Prenom:ms**
- **Éric,N+Prenom:ms**
  - a noun (N)
  - a first name (Prenome)
  - masculine (m)
  - singular (s)
- **Marko,N+NProp+Hum+First**
- **Vitas,N+NProp+Hum+Last**
  - a noun (N)
  - a proper name (NProp)
  - a human (Hum)
  - a first name (First)
  - a surname (Last)

What is in dictionaries that can help NER?

common nouns for titles, professions, positions

- **professeur,N+z1+Profession**
- **directeur,N+z1+Profession**
- **roi,N+z1+Profession**
  - a profession (Profession)
- **profesor,N+Hum+Prof**
- **direktor,N+Position**
- **kralj,N+Hum+Ttl**
  - a noun (N)
  - a human (Hum)
  - a profession (Prof)
  - a position (Position)
  - a title (Ttl)

What is in dictionaries that can help NER?
and other common nouns

- **mai,N+z1**
- **dimanche,N+z1**
- **après,PREP+z1**
- **quotidiennement,ADV+z1**

- **maj,N**
- **nedelja,N**
- **posle,PREP**
- **svakodnevno,ADV**
  - a noun (N)
  - an adverb (ADV)
  - a preposition (PREP)

Can a NER system rely on dictionaries only?

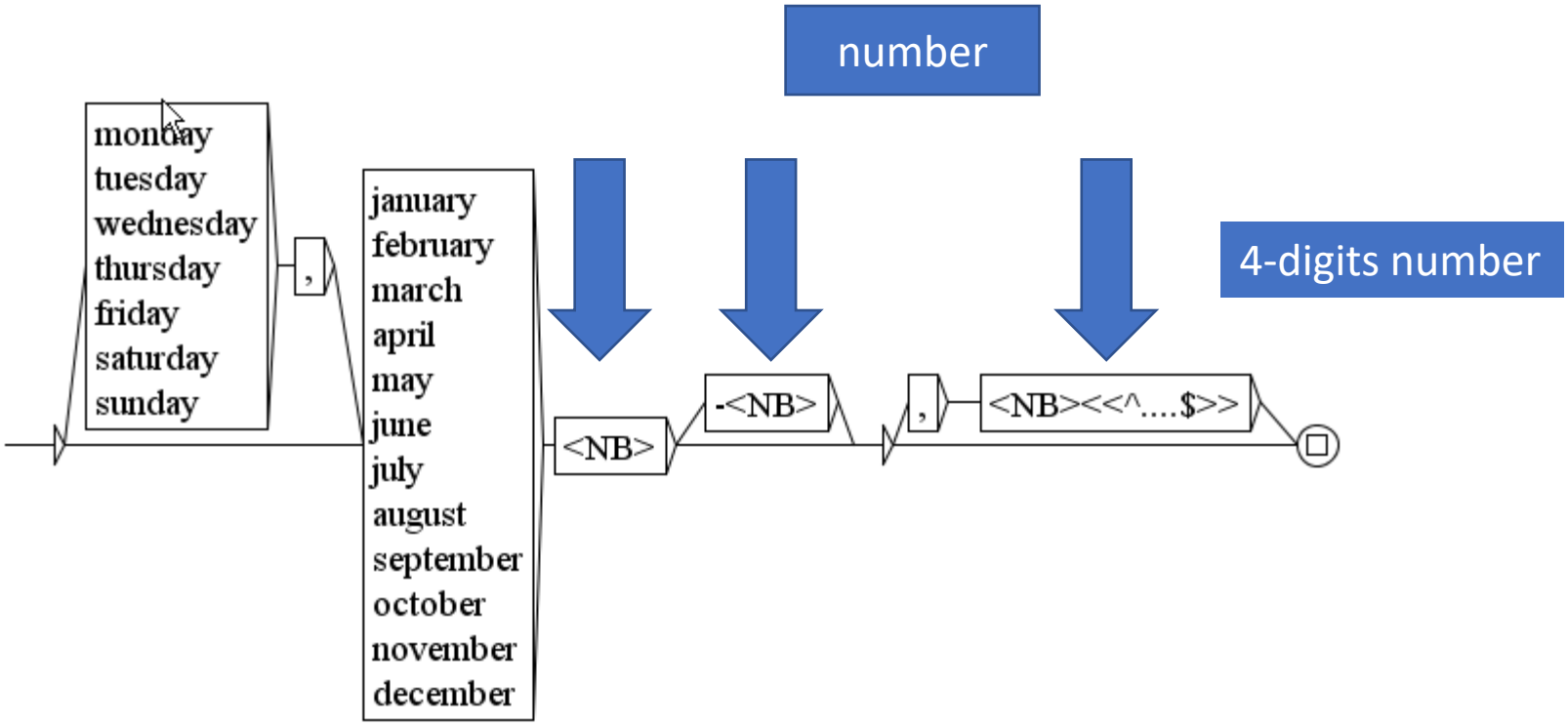
- Would lexical masks like these be enough to recognize NEs in a text with high recall and precision?

French	Serbian	NE type
<N+PR+Toponyme+Hydronyme>	<N+NProp+Hum+Hyd>	hydronym
<N+Groupement+Organisation>	<N+NProp+Org>	organization
<N+Prenom> <N+Prenom>	<N+First> <N+Last>	person's full name

Dictionaries are not enough, why?

- Dictionaries can never cover all named entities (names of people, toponyms, organization names, etc.). (recall drops)
- This is especially true for temporal expressions that can be written in various formats and numerical expressions (measures, money...) that use numerals written with digits, words and ther combination.
- Moreover, word forms are ambiguous that may lead to false recognitions. (precision drops)

A date graph, can that be described by a dictionary?



Examples of ambiguity

- some names are used to designate several things: ***Balkan*** – a mountain and a superregion; ***Congo*** (***Kongo***) – a river and a country

- geographic names: ***Aden*** – a city and a personal name; ***Amazon*** – a river, an ethnonym, an organization

- geographic names: ***Una*** – a river and a feminine name, ***Sava*** – a river and a masculine name and a feminine name; ***Jelica*** – a mountain and a feminine name
- personal names: ***Miloš*** – a first name and a surname;
- common nouns: ***Višnja*** – a feminine first name and a common noun (sour cherry), ***Ali*** – a masculine first name but aslo; a conjunction (but);

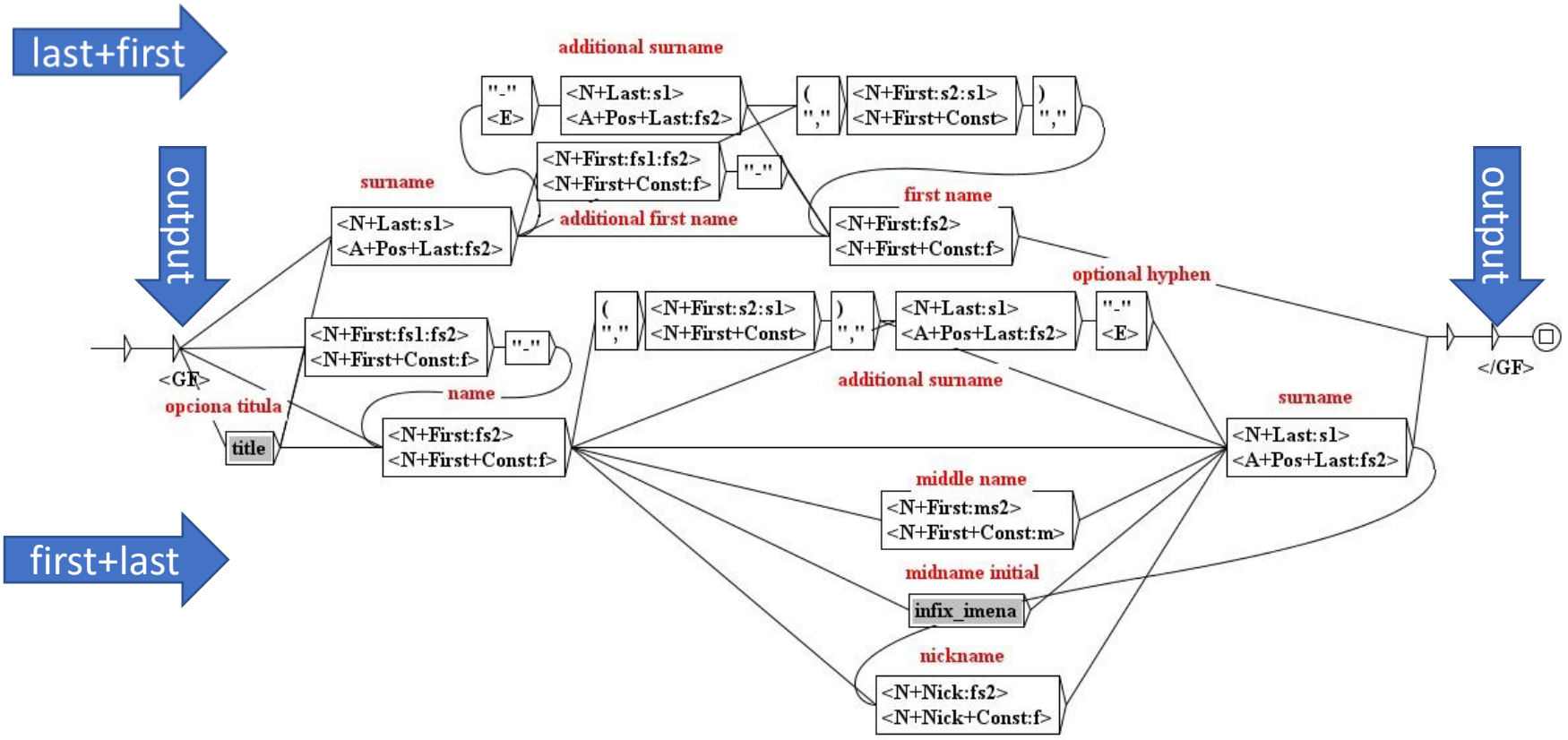
The ambiguity of forms

	Ivan (masculine)	Ivana (feminine)
nominative	Ivan	Ivana
genitive	Ivana	Ivane
dative	Ivanu	Ivani
accusative	Ivana	Ivanu
vocative	Ivane	Ivana
instrumental	Ivanom	Ivanom
locative	Ivanu	Ivani

Graphs can help to solve many of these problems, how?

- **1:**
- They can descibe different forms a certain NE can appear in;
- For example, a full name of a person as it appears in a text can consists of:
  - First name
  - Last name
  - Additional last name
  - Middle name or initial
  - Nick name
  - Title
- Various orders of components; some components are optional
- The recognized information is an output, e.g. the text is modified;

Feminine personal names in the genitive case (Serbian)



Names that would be recognized by some of these graphs (m/f; nominative, genitive...)

- **dr Milan Jovanović Batut** – title, first, last, nick (m/nom);
- **Bojana Petrović-Popović** – first, last,hyphen,last (f/nom);
- **Milan Gale Muškatirević** – first, nick, last (m/nom);
- **Mahtija Ahtisarija** – first, last (m/gen);
- **Mirjana R. Milenković** – first, initial, last (f/nom);
- **prof. dr Dragan Radovanović** – title, title, first, last (m/nom);
- **Aleksandar-Kristijan Golubović** – first, hyphen, first, last (m/nom);
- **Prof. Jovana Hadži-Đokića** – title, first, link, hyphen, last (m/gen);
- **Tešić Zoran** – last, first (m/nom).

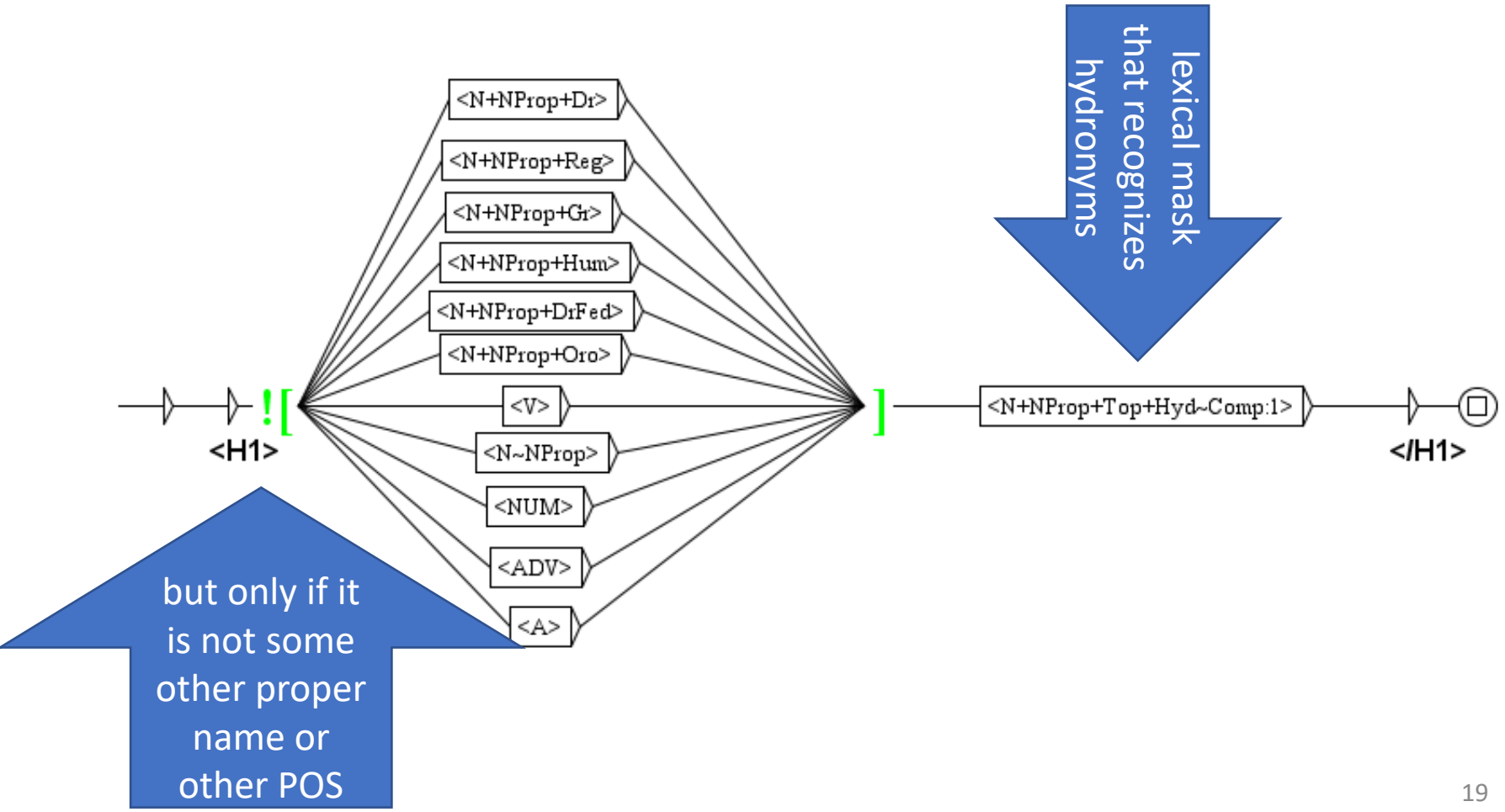
2: How can graphs help with false recognitions?

- Some lexical masks can be restricted if there is an unwanted ambiguity;
- Some concordance lines obtained with the lexical mask **<N+NProp+Hyd>** on a Serbian text:

povećanje vodostaja na	**Nišavi**	, ali i na planinskim
DOSTIGAO JE JUČE	**OKO**	PODNE VRH NAKON ČEGA
ali i plastenike.	**Po**	rečima mnogobrojnih
zbog izlivanja	**Rasine**	, te Zapadne i Južne

- There are false recognitions because
  - ***Oko*** is a form of ***Oka***, a river in Russia and also a preposition ***oko*** and a form of a noun (eye);
  - ***Po*** is a river in Italy, and a preposition ***po***.

Rejecting false recognitions



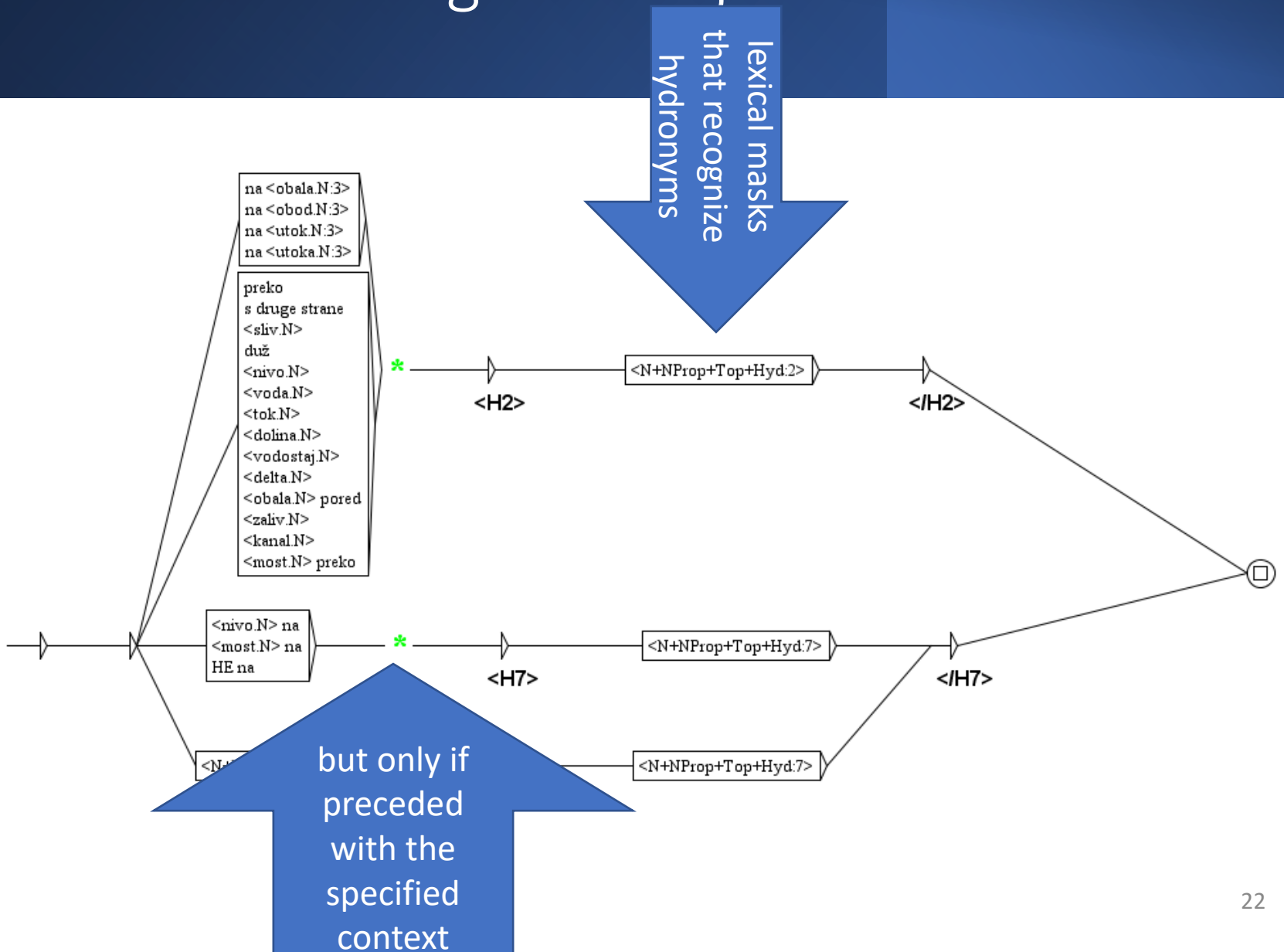
The „right context“

- Green brackets introduce so called „right context“;
- The context can be „positive“ if you want the pattern inside the bracket to match the analysed text at the current position, or „negative“ otherwise;
- Suprisingly, the „right context“ can appear at the beging of the graph (as in our example), in which case it proceeds only if at the current position the pattern inside brackets is matched (positive) or not (negative).

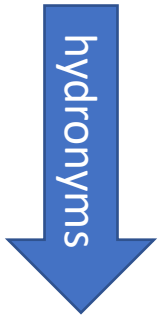
This can be rather restrictive and reject many true positives

- This graph would reject all ocurrences of the river ***Sava*** because it is ambiguous with the first name
- It will reject all ocurrences of the river ***Kolubara*** because ***kolubara*** is an ergonym, e.g. a trade name of a coal.
- It will reject all ocurrences of the river ***Gradac*** because it is also the name of numerous villages in Serbia, Croatia, Bosna and Hercegovina, Montenegro, Slovenia, ... and a diminutive of ***grad*** (city)
- Can this be mended?

3. Recovering false rejections

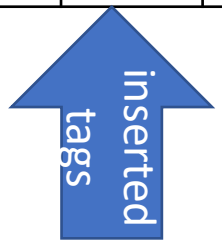


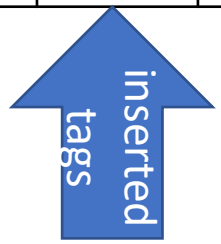
What would be recognized and tagged by this graph?



On the other side of the Drina
a **bridge over** the Danube
on the border **along** the Danube
raft in **Belgrade on** the Sava
on the banks of the Danube

s druge strane	<h2>	**Drine**	</h2>	da se „dignu“
za **most preko**	<h2>	**Dunava**	</h2>	, u prostorijama
na granici **duž**	<h2>	**Dunava**	</h2>	rešiće se
splav u **Beogradu na**	<h7>	**Savi**	</h7>	odakle je krenuo
Na obali	<h2>	**Dunava**	</h2>	pojali su

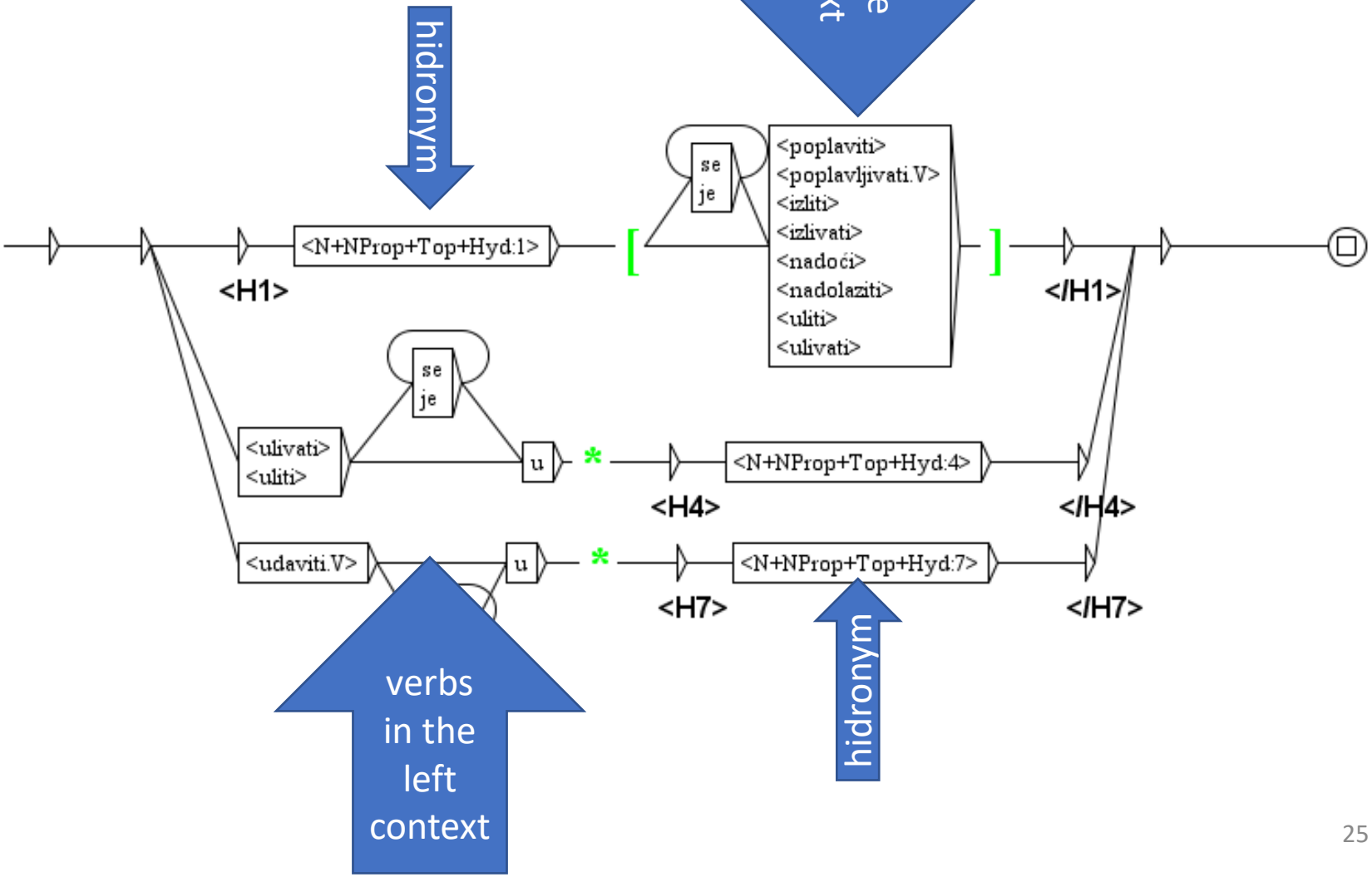




The „left context“

- The green asterix indicates the end of the „left context“ of the expression we want to match;
- Graphs can be written without this special feature, but than the left context is a part of a recognized section (and appears, for example, in concordances);
- Note: output in the „left context“ is ignored.

3. Recovering false re ons



What would be recognized and tagged by this graph?

the river Tamnava **overflowed** in the region
Timok **flooded** 300 houses
she **drowned** in the Sava
She **jumped** into the Sava.

hydronyms

triggers

dok se reka	<h1>	**Tamnava**	</h1>	**izlila** u rejonu
	<h1>	**Timok**	</h1>	**poplavio** 300 kuća
udavila se u	<h7>	**Savi**	</h7>	i da je doneta u
i **skoči u**	<h4>	**Savu**	</h4>	.

triggers

inserted tags

inserted tags

4. Normalization of recognized entities

- Sometimes it is useful to normalize the recognized values, and pass that value to the output;
- This can be done for several purposes:
  - a numerical value in an entity (measures, money, percents, etc.) can be normalized;
  - a recognized date or time expression can be put in a normalized form, according to the standard;
  - for a recognized entity for person, place etc. a lemma can be output.

Normalization of numerals

171,1 milion evra	<money val="**171100000EUR**"/>
18 milijardi i 800 miliona dolara	<money val="**180800000000USD**"/>
160 kilometara na čas	<measure val="**160kmh**"/>
milijardu kubnih metara	<measure val="**1000000000m3**"/>
7,2 posto	<percent val="**7.2%**"/>
tri i po odsto	<percent val="**3.5%**"/>

Special type of graphs, called dictionary graphs, were developed for Serbian that are used in the phase of the application of dictionaries.
They assigne to each numeral its „normalized“ value.
The normalized value is also assigned to all currencies and measurement units in the Serbian dictionaries.

1991. godine avgusta meseca	**<TIMEX3 type="DATE" val="1991-08"/>**
početkom maja 1992. godine	**<TIMEX3 type="DATE" val="1992-05" mod="START"/>**
dva sata i 15 minuta popodne	**<TIMEX3 type="TIME" val="T14:15"/>**
tokom meseca aprila 2011. godine	**<TIMEX3 type="DURATION" val="2011-04"/>**
poslednjih sto godina	**<TIMEX3 type="DURATION" val="P100Y"/>**
triput nedeljno	**<TIMEX3 type="SET" val="P1W" freq="3X"/>**

- Values are assigned to the attributes of TIMEX3 tag according to the standard TimeML

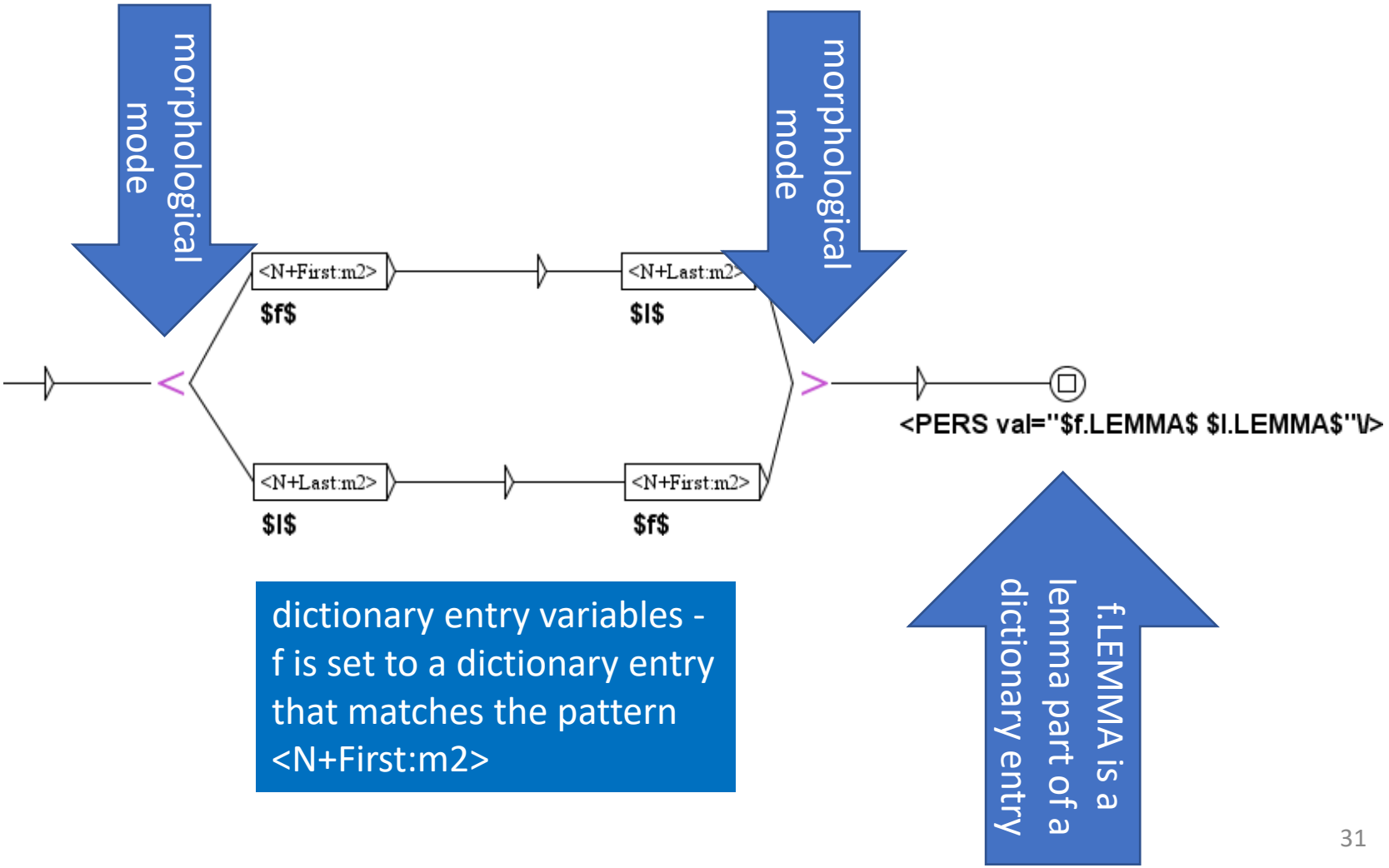
Normalization of temporal expressions

Normalization of proper named entities

Prof. Jovana Hadži-Đokića	<PERS val=„**Prof. Jovan Hadži-Đokić**"/>
Mahtija Ahtisarija	<PERS val=„**Mahti Ahtisari**"/>
udavila se u Savi	udavila se u <LOC val=„**Sava**"/>
Jadranskog mora	<LOC val=„**Jadransko more**"/>
Beogradskih elektrana	<ORG val=„**Beogradske elektrane**"/>
MMF	<ORG val=„**Međunarodni monetarni fond**"/>

- Transducers that perform this normalization rely on the content of dictionaries
  - For instance, the entry for „Međunarodni monetarni fond“ (International Monetary Fund) has to have a marker saying that MMF is its acronym.
- They also use a special „morphological mode“ and dictionary entry variables.

Dictionary entry variables and the morphological mode



Cascades of graphs for NER

What are cascades of transducers?

- Transducers are graphs that produce output (as we saw before)
  - the output can be merged with the text, or;
  - it can replace the recognized sequence.
- A cascade of transducers applies a sequence of transducers one after another;
- Each of these trandusers produces output that can be used by transducers that follow;
- In a cascade, all types of transducers acn be used that can be used;
  - only the use of the left context is of no use;

What can be the output of a cascade?

- It can be whatever one finds suitable for solving one's problem;
- For producing a NER system it is useful if transducers produce output that is in the form of a lexical tag?
- A lexical tag is one type of a token in Unitex
  - For instance, **{aujourd'hui,.ADV}** is a lexical tag in a French text that is introduced during preprocessing;
  - In this way, there are no more three separate tokens: **aujourd** - **'** - **hui**, two of them being nonexistant words in French.

How it works?

- Look at the sequence: **Bank of England**
- A transducer reognizes England as an administrative location and replaces it with a lexical tag:
- **Bank of {England,.entity+loc+adm}**
- A subsequent transducer looks for names of organization and one of patterns it looks for is: **<bank.N> of <entity+loc>**
- It recognizes that patterns and replaces the whole sequence with the new lexical tag:
- **{Bank of {England,.entity+loc+adm},.entity+org+ent}**

Embedded named entities are recognized

---les
<org>
usines
<orgName>
du
<persName>
<roleName type="nobility">marquis</roleName>
<nameLink>de la</nameLink>
<surname>Lande</surname>
</persName>
</orgName>
</org>
factories of marquis de la Lande

Embedded named entities are recognized

<pers>
<persName_full>Aleksandra Miloševića</persName_full>,
<role>učitelja</role>
<org_gen>OŠ „
<persName_full>Velizar Stanković Korčagin</persName_full>"
iz
<top_gr>Velikog Šiljegovca</top_gr>
</org_gen>
</pers>

Aleksandar Milošević, teacher of the EC „Velizar Stanković Korčagin“ from Veliki Šiljegovac (in genitive)

Why are embbeded entities important?

- Because they connect recognized entities;
- What do we know about **Aleksandar Milošević**?
  - He is a teacher;
- Where does he teach?
  - At the **elementary school „Velizar Stanković Korčagin“**
- Where is that elementary school?
  - In **Veliki Šiljegovac**.

What is specific for the use of transducers in a Unitex cascade?

- For each transducer in a cascade one has to specify whether it is used in a merge or a replace mode;
- For instance:
- **Merge mode**: enclosing recognized sequence in XML tags:
  - **…+381 60 1234567…** ⇒ **<tel>+381 60 1234567</tel>**
- **Replace mode**: replace recognized sequence with a new text:
  - **…+381 60 1234567…** ⇒ **<tel-number/>**
  - (anonimization)

Repetition of the use of a transducer

- A transducer in a cascade can be used repeteadly until reaching a fixed point, that is, the point in which it cannot modify the text anymore.
- One has to be careful with the use of such transducers in order not to enter an endless loop (like in programming).
- Example:
- Text uses a tag **** for emphasizing a sequence, and an attribute **type** to say how, with values: **i** for italic, **b** for bold, and **s** for small caps. These tags can be embbeded.
- We want to replace the tag **** with tags **<i>**, **<b>** or **<s>** according to the value of the attribute **type**.

Example:

- Text:
- ...aaa bbb ccc ddd eee fff ggg hhh iii jjj...
- Transducer:
- Recognizes tags and and text between them that consists of [a-j]+ and tags , , , , <s>, </s> (but not and)
- and replaces tags and with <$x> and </$x>.

How does the transducer work in the iteration mode?

- Text:
- ...aaa bbb ccc ddd eee fff ggg hhh iii jjj...
- First run recognizes eee fff
- Resulting text: ...aaa bbb ccc ddd <s>eee fff</s> ggg hhh iii jjj...
- Second run recognizes: ccc ddd <s>eee fff</s> ggg hhh
- Resulting text: ...aaa bbb ccc ddd <s>eee fff</s> ggg hhh iii jjj...
- Third run recognizes: aaa bbb ccc ddd <s>eee fff</s> ggg hhh iii jjj
- Resulting text: ...aaa bbb ccc ddd <s>eee fff</s> ggg hhh iii jjj...
- Forth run: fixed point, transducer cannot recognize anything.

Rules applied to transducers in a Unitex cascade

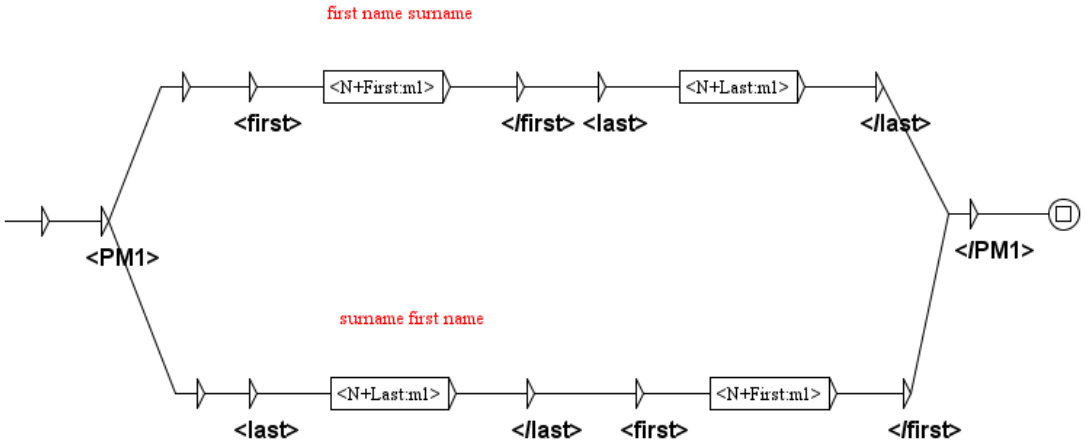
- Priority is always given to the ***longest match***.
- Regularly (outside a cascade) all matches by a local grammar are indexed, which means that they are listed also in concordances. Since a transducer in a cascade modifies text, it has to choose between several possibilities, and it chooses the ***leftmost match***.
- ***Weights*** can be given to paths that match the same sequence:
  - when producing concordances, only line with the highest weight is listed;
  - in a cascade, since a transducer has to produce an output, one path among those that match the same sequence is randomly chosen, unless the wights are assigned to these paths.

Example of the use of weights

- Priority is always given to the ***longest match***.
- Regularly (outside a cascade) all matches by a local grammar are index, which means that they are listed also in concordances. Since a transducer in a cascade modifies text, it has to choose between several possibilities, and it chooses the ***leftmost match***.
- ***Weights*** can be given to paths that match the same sequence:
  - when producing concordances, only line with the highest latest weight is listed;
  - in a cascade, since a transducer has to produce an output, one path among those that match the same sequence is randomly chosen, unless the wights are assigned to these paths.

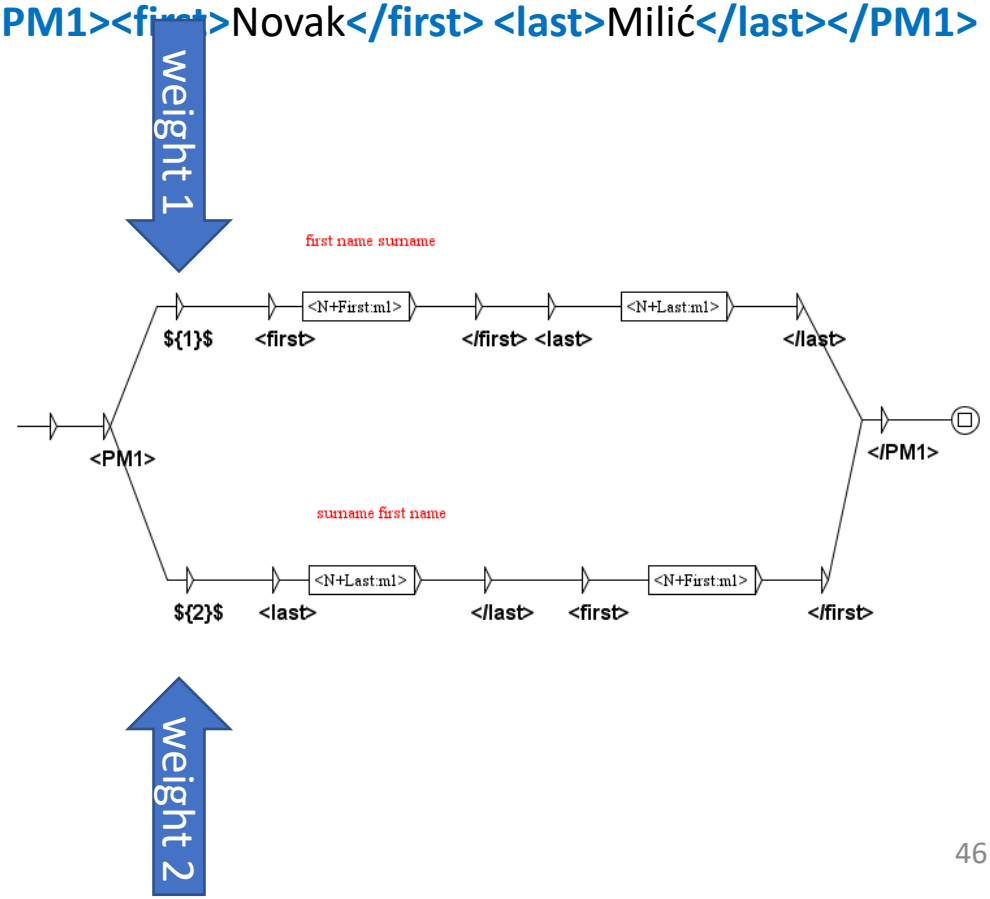
A transducer without weights

- Recognizes masculine names in the nominative case writen in one of two forms:
- first name followed by a surname (a path above)
- a surname followed by a first name (a path below)
- In the case both pats are matched, e.g. **Novak Milić**, one path would be randomly chosen.



A transducer with weights

- Weight **1** is given to the: first name followed by a surname (a path above)
- Because it is the usual (more frequently used) order of a first name and a surname
- *Novak Milić*, would be tagged as **<PM1><first>**Novak**</first> <last>**Milić**</last></PM1>**



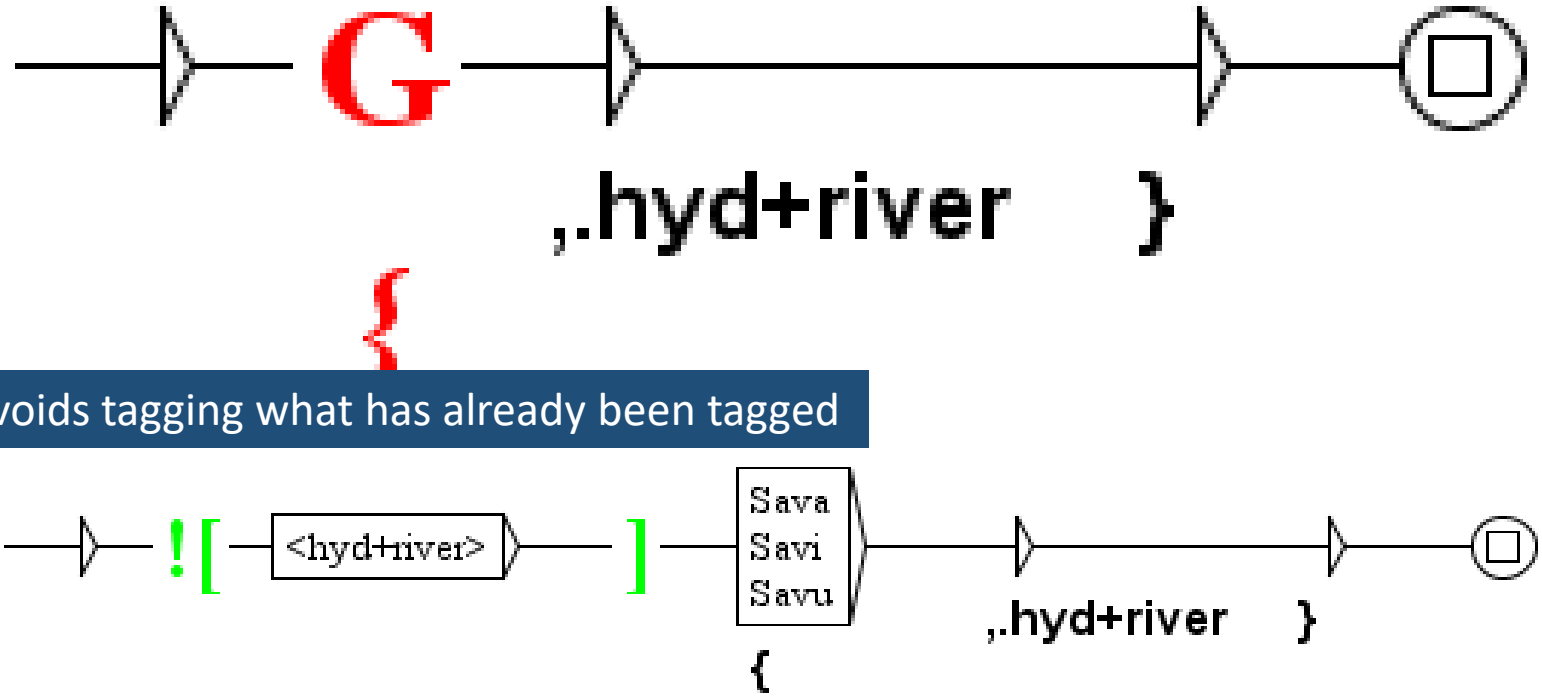
Tagging generalization graphs

- A special kind of graphs to be used in cascades, and useful for NER;
- Sometimes we recognize something in a text due to triggers; e.g. in **...reka Sava...** ***reka*** is a trigger that enables tagging ***Sava*** (which is ambiguous) as a hydronym.
- But ***Sava*** can appear in the same text without any trigger in her context.
- Then generalization graphs can be used that have empty boxes that the program itself fills with forms automatically extracted from a list of tokens of a text that were previously tagged in a given way.

The example of a generalization graph

- At a certain point of the application of a cascade a list of tokens of a text contains the following:
- **{Sava,.NE+hyd+river:ms1q}**
- **{Savi,.NE+hyd+river:ms3q:ms8q}**
- **{Savu,.NE+hyd+river:ms4q}**
- In a text there are still forms ***Sava***, ***Savi***, ***Savu***... that were not tagged

A generalization graph



A generalization graph after processing

Why else are cascades interesting for NER?

- With cascades recognition (analyse) is separated from tagging (synthese).
- One cascade is used for the recognition and the other for the tagging.
- That means that different taggings can be used for different purposes, not changing anything in the recognition process:
  - With or without embbeding
  - Skipping tags for some recognized entities, etc.

An example of the analysis and synthesis – a text FRA02001 from the French sub-collection

D-reading NER tag set

- - Le gouvernement français, messieurs, avait été pressenti par l'**<ORG>**Aéro-Club**</ORG>** au sujet de cette course au Pôle, le **<ORG>**Conseil des ministres**</ORG>** m'a chargé de vous dire qu'il ajoute cent mille francs aux trois cent mille dont on vient de vous parler !
- Le **<PERS>**marquis de la Lande**</PERS>** agita le bras pour demander la parole puis il annonça :
- - Je donne deux cent mille francs ! Ne voulant pas demeurer en reste avec le richissime marquis, ami de tous les sports, les gros instructeurs d'automobiles, les **<ROLE>**industriels**</ROLE>** qui se trouvaient là s'empressèrent de s'engager pour de fortes sommes...
- Si bien que le **<ROLE>**président**</ROLE>**, lorsque le tumulte fut un peu calmé, put annoncer :
- - Messieurs, le prix du voyage au Pôle sera de un million six cent mille francs !

An example of the analysis and synthesis – a text FRA02001 from the French sub-collection

TEI NER annotation

- <s>- Le gouvernement français, messieurs, avait été pressenti par l'**<orgName>**Aéro-Club**</orgName>** au sujet de cette course au Pôle, le **<orgName>**Conseil des ministres**</orgName>** m'a chargé de vous dire qu'il ajoute **<measure type="currency" quantity="cent mille" unit="francs">**cent mille francs**</measure>** aux trois cent mille dont on vient de vous parler !</s></s>
- <s>Le **<persName><roleName type="nobility">**marquis**</roleName> <nameLink>**de la**</nameLink> <surname>**Lande**</surname></persName>** agita le bras pour demander la parole puis il annonça :</s></s>
- <s>- Je donne **<measure type="currency" quantity="deux cent mille" unit="francs">**deux cent mille francs**</measure>** !</s> <s>Ne voulant pas demeurer en reste avec le richissime marquis, ami de tous les sports, les gros instructeurs d'automobiles, les **<roleName type="office">**industriels**</roleName>** qui se trouvaient là s'empressèrent de s'engager pour de fortes sommes...</s></s>
- <s>Si bien que le **<roleName type="office">**président**</roleName>**, lorsque le tumulte fut un peu calmé, put annoncer :</s></s>
- <s>- Messieurs, le prix du voyage au Pôle sera de **<measure type="currency" quantity="un million six cent mille" unit="francs">**un million six cent mille francs**</measure>** !</s></s>

How can cascades be used?

- They can be used in the Unitex environment;
- Preferable mode for developers – it is easy to try and correct;
- Unitex commands used in the Unitex environment can be translated into a script, which subsequently can be used in the command line of the operating sequence.
- The script can be applied to a single file or to a collection of files in a folder.
- Preferable mode for end users.
- That is how French NER system (**CasEN**) was used to annotated the whole French collection.

Is this all that can be said about Unitex?

- Certainly not.
- There is much more that can be said about graphs, transducers, cascades etc.
- For instance, we have not mentioned various sort of variables and how they can help in solving probles.
- Fortunately, there is a comprehensive manual that can be downloaded from the Unitex official site
- https://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf

You can also read some articles:

- Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. "A system for named entity recognition based on local grammars." *Journal of Logic and Computation* 24, no. 2 (2014): 473-489.
- Jaćimović, Jelena. "Recognition and normalization of temporal expressions in Serbian medical narratives." *Infotheca-Journal for Digital Humanities* 19, no. 2 (2019): 26-60.
- Jaćimović, Jelena, Cvetana Krstev, and Drago Jelovac. "A rule-based system for automatic de-identification of medical narrative texts." *Informatica* 39, no. 1 (2015).
- Friburger, Nathalie, and Denis Maurel. "Finite-state transducer cascades to extract named entities in texts." *Theoretical Computer Science* 313, no. 1 (2004): 93-104.
- Maurel, Denis, Enza Morale, Nicolas Thouvenin, Patrice Ringot, and Angel Turri. "ISTEX: A database of twenty million scientific papers with a mining tool which uses named entities." *Information* 10, no. 5 (2019): 178.