Derivation and Graphs

Duško Vitas

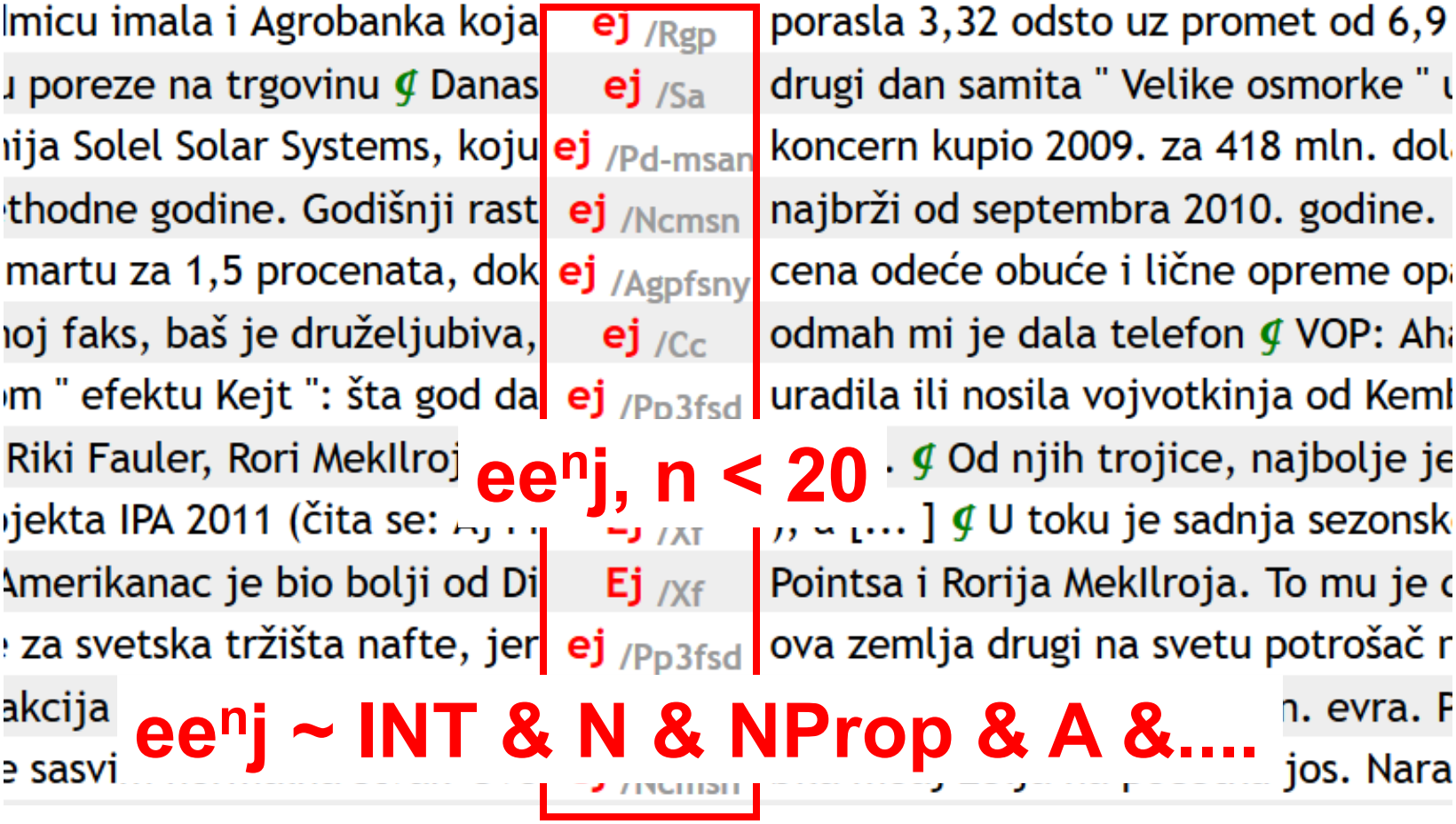
Problems

- The problem of unknown words in corpora processing
- The problem of the description of derivational and the consequences to e-dictionary organization (and dictionaries in general)

A question?

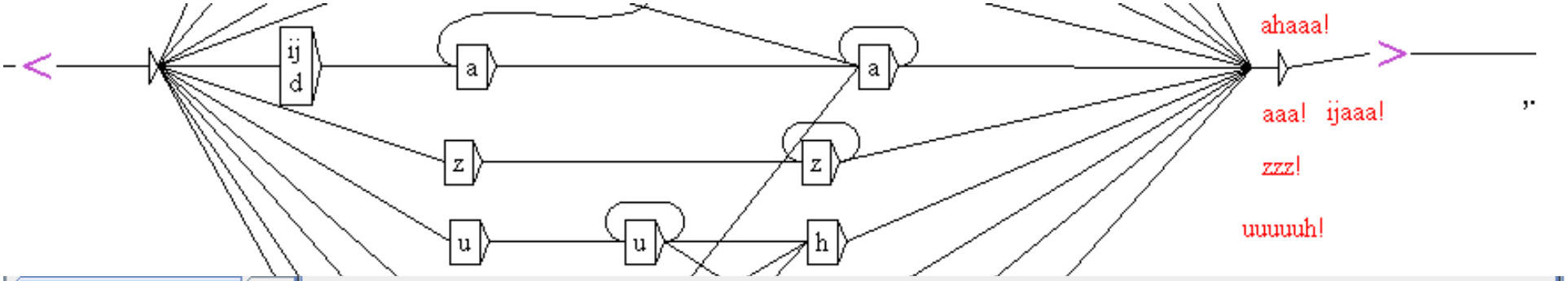
- Should each word be tagged, even if a tag might be wrong, or should words be left unprocessed if correct tags cannot be assigned to them?
- The relation between statistical methods and rule-based methods.

An example - srwac - ej!



With dictionary graphs

Input: ***ej, eej, eeej, eeeej, eeeeej, eeeeeeeeeeeeeeeeeej!***



Locate configuration | Adv

Locate pattern in the form of

Regular expression:

<INT>

Graph:

Activate debug mode

Index | Gran

6 matches

ej, eej, eeej, eeeej, eeeeej, eeeeeeeeeeeeeeeeeej!
ej, eej, eeej, eeeej, eeeeej, eeeeeeeeeeeeeeeeeej!
ej, eej, eeej, eeeej, eeeeej, eeeeeeeeeeeeeeeeeej!
ej, eej, eeej, eeeej, eeeeej, eeeeeeeeeeeeeeeeeej!
ej, eej, eeej, eeeej, eeeeej, eeeeeeeeeeeeeeeeeej!
ej, eej, eeej, eeeej, eeeeej, eeeeeeeeeeeeeeeeeej!

Unknown words

... or unrecognized words by a system of e-dictionaries:

- No matter how comprehensive e-dictionary might be, new unknown words occur with each new text!
- Even the most ordinary newspaper article is usually not covered by e-dictionaries!

An example

The corpus of the weekly magazine ***Vreme*** 2010-2018:

29.930.758 (**422.508** diff) tokens

13.062.037 (**422.230**) simple forms

...but in **ERR**:

129.124 (**46.315** diff) simple forms or
0.431% (of the text)

~ 11% of diff. simple forms

Some examples from err

sveta na portalu www . **wikileaks** . org (u međuvremenu

doncite ... ni e vo dobro **seќavanje**", glasi makedonski

je da Srbija slučaj stavi **ad akta**, pošto je rezultat u

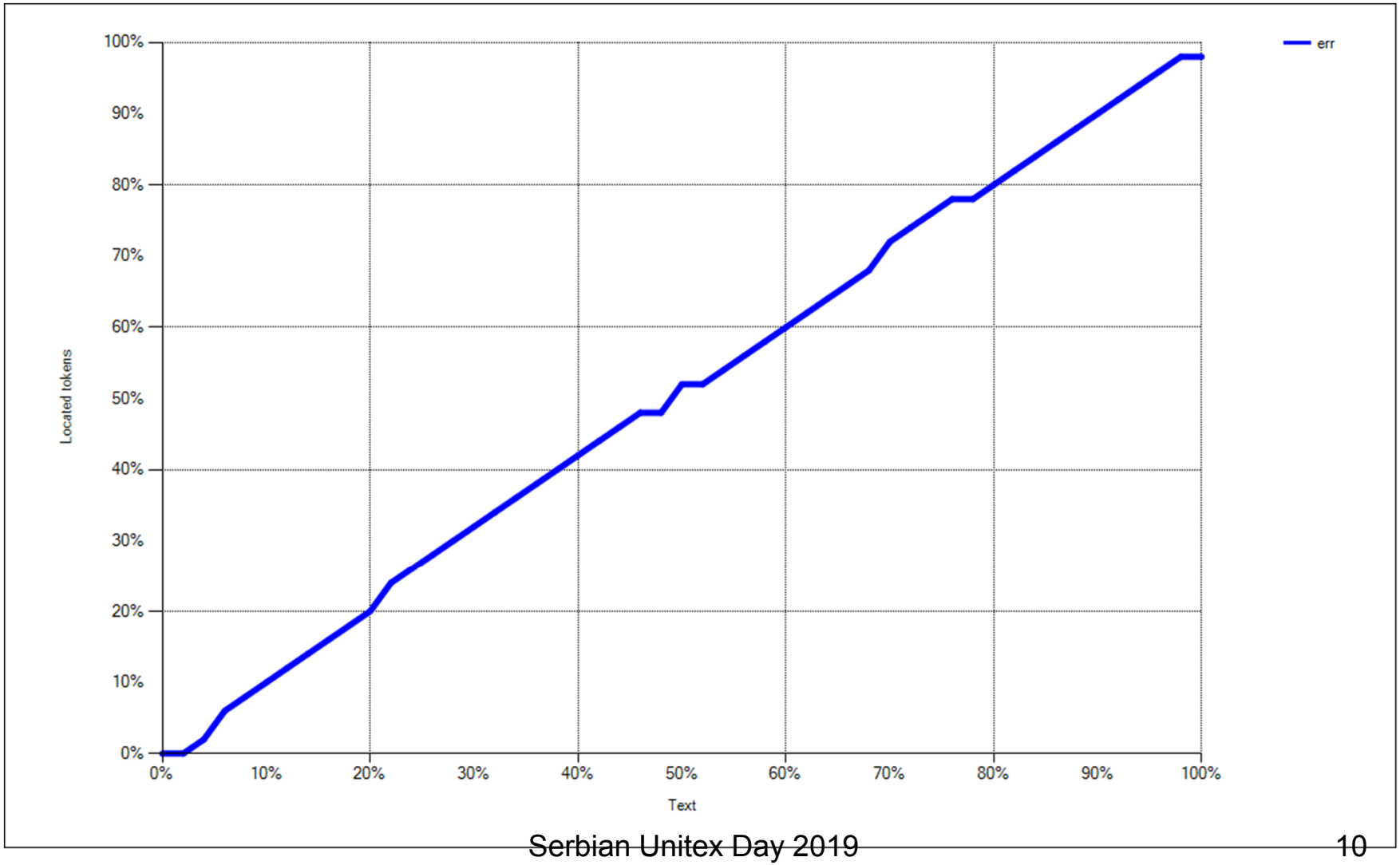
obila (što je, reći će naš **amerikanoFILski** autor, omaž m

određenih mera protiv **antidejtonskih** političara, na

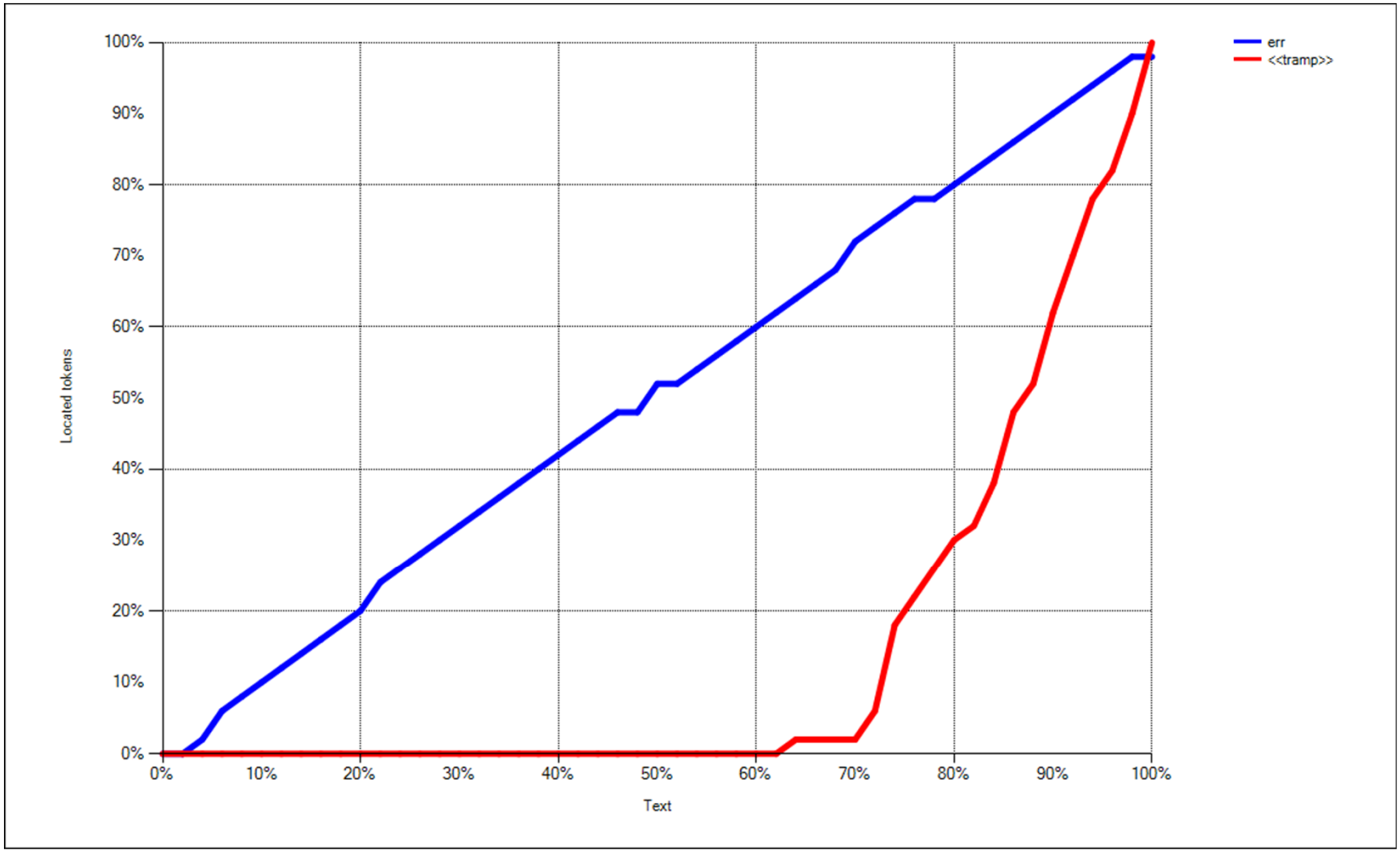
nifestacije, još od doba **antifederalista**, koji su se

avidnu volju i umeće da **atrofirajućim** odnosima

Distribution of err vs. new word



Distribution of err vs. new word



An approximation

- If words tagged by graphs as **Unk** and **ABB** are included, and this set also includes some potentially incorrect words, then numbers are:

  908.326 (**877.006** diff) simple forms or **3.034**%

- (correct solutions are included).

Examples of Unk and ABB

N+NProp+Unk:

J. Evaristo, Monti, **Arico** Suarez – Peucelle,	**OK**
istrage (takozvana **Chilcot** Inquiry).{S} Javn	**OK**
sno SIS-a (Secret **Intelligence** Service).{S}	**OK?**
iranjem u lepoti" i **Ronaldinjovim** majstorijam	**A**
anizmu čoveka.{S} **Biočip** je zapravo mala pl	**Not NP**

ABB:

priznao u vestima **B92** da jeste reč o njegov:	**B is ABB**
javnosti.{S} I MI5 i **MI6** izgubile su zbog zlou	**MI is ABB**
a objašnjavaju da **rad** tajnih službi, uz sve	**OK**
a u goste 1994. u **SAD** i 1998. u Francusku,	**OK**

<<merkel>> (1300 occ.)

In DLF:

Merkel,.N+NProp+Hum...

Merkelova,Merkelov.A+Pos+NProp+Hum+Last+... **???**

Merkelici,.N+NProp+**Unk**

Merkelina,.N+NProp+**Unk ?**

In ERR:

merkelice

("***merkelice***" ovog sveta)

merkelovci

(***merkelovci*** svih partija)

merkelovih

(od svih tih makrona , ***merkelovih*** , junkera i rahoja)

merkelovski

Twitter in 10 years

Derivation

Derivational dictionary

Worth, S., A. Kozak, D. Johnson: ***Russian Derivational Dictionary***. RAND Corporation, 1970.

МАРКС

suffixes

МАРКС ЙЗМ
МАРКС ЙСТ
МАРКС ЙСТ К А
КВАЗИ МАРКС ИСТ
НЕ МАРКС ЙСТ
ПО - МАРКС ЙСТ СК И
МАРКС ЙСТ СК ИЙ
АНТИ МАРКС ЙСТ СК ИЙ
ПСЕВД О МАРКС ЙСТ СК ИЙ
ДО МАРКС ЙСТ СК ИЙ

Derivational dictionary

Worth, S., A. Kozak, D. Johnson: ***Russian Derivational Dictionary***. RAND Corporation, 1970.

МАРКС

prefixes

	МАРКС ЙЗМ
	МАРКС ЙСТ
	МАРКС ЙСТ К А
КВАЗИ	МАРКС ЙСТ
НЕ	МАРКС ЙСТ
ПО -	МАРКС ЙСТ СК И
	МАРКС ЙСТ СК ИЙ
АНТИ	МАРКС ЙСТ СК ИЙ
ПСЕВД О	МАРКС ЙСТ СК ИЙ
ДО	МАРКС ЙСТ СК ИЙ

Capital letters

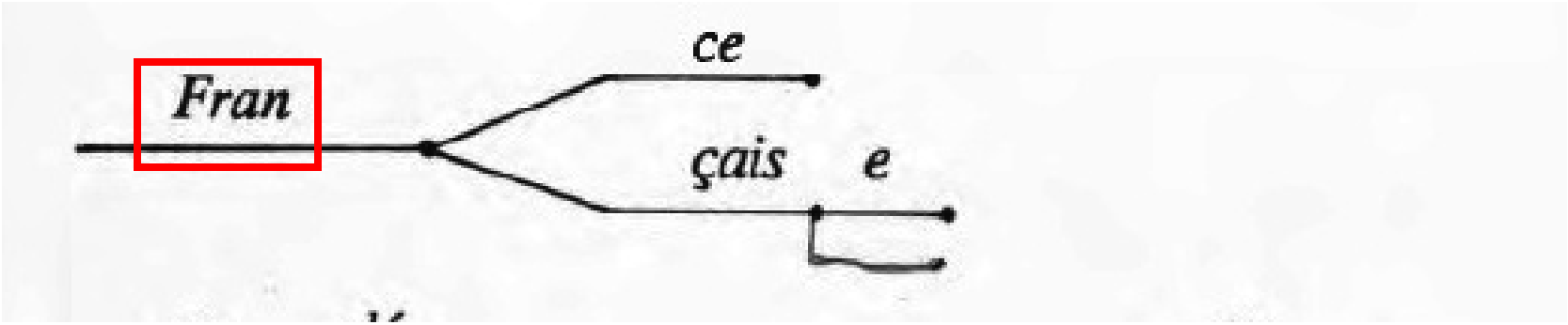
МАРКС

МАРКС ЙЗМ
МАРКС ЙСТ

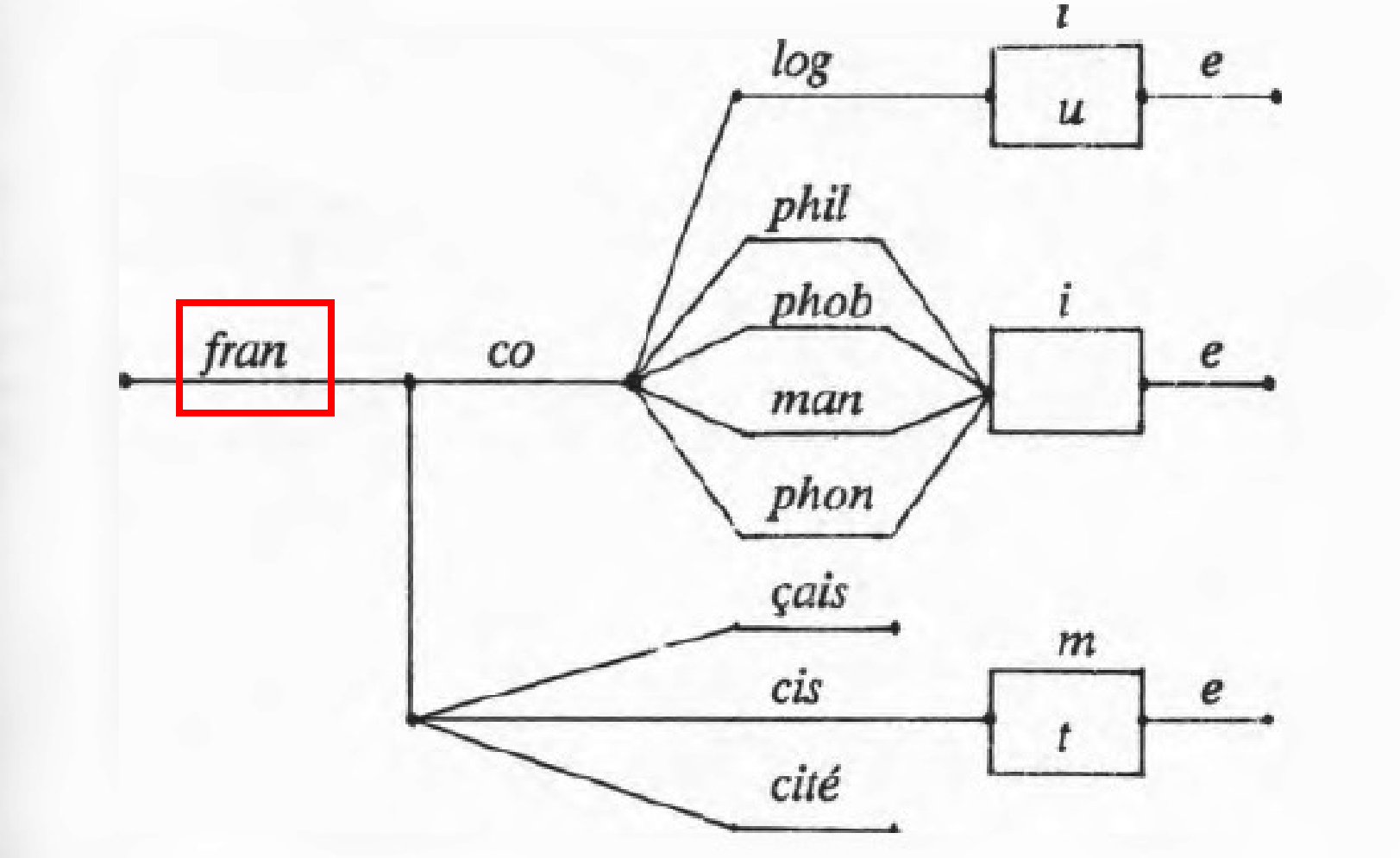
Marks, but **m**arksizam

The representation done by Gross

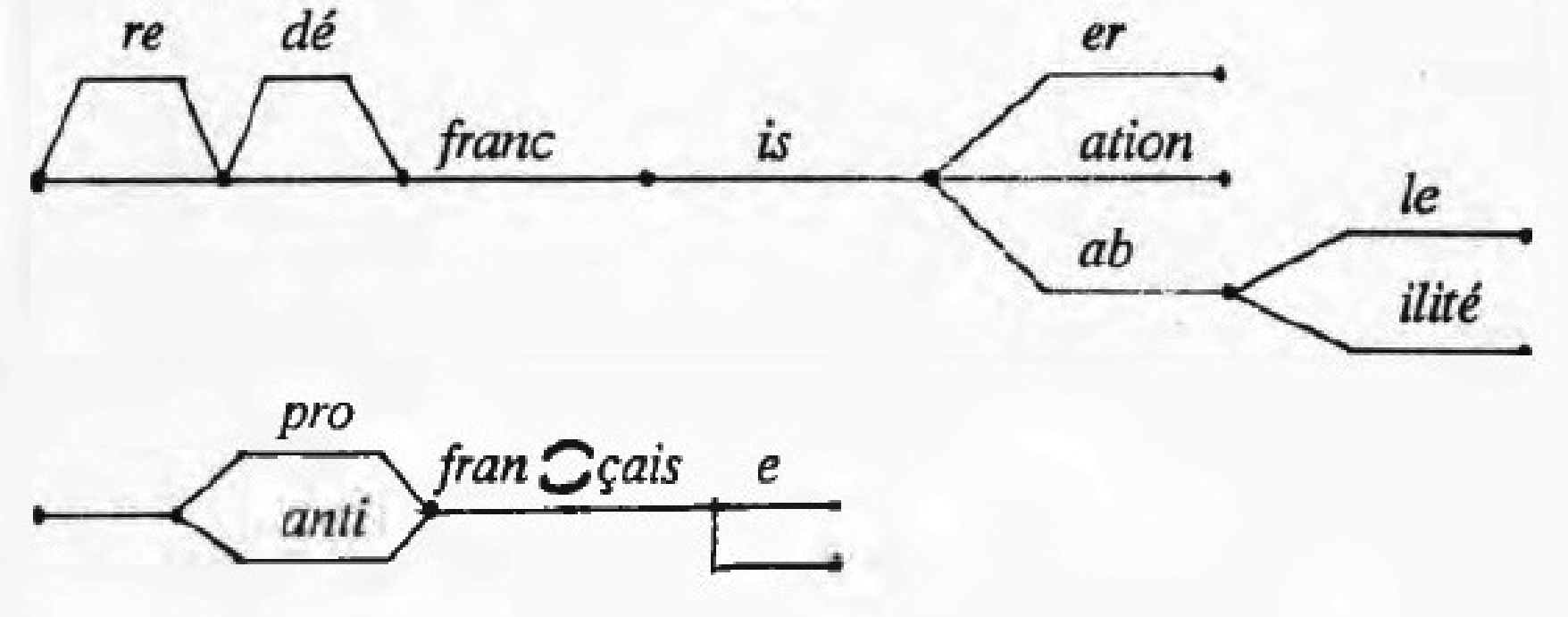
Gross, Maurice: *The Use of Finite Automata in the Lexical Representation of Natural Language*. LNCS 377, pp. 34-50, 1989



The factorization to the right - suffixes



The factorization to the left - prefixes



The case of Serbian

In dictionary RMS6:

***glumičica*, dim. of *glumica*?**
***glumičin*, adj. poss. of *glumica*?**
glumčićev, adj. poss. of glumčić?
...

gluma	n.
glumac	n.
glumački, -a, -o	adj. related to actors
glumački	adv. in a way of actors
glumiti	v. to act
glumica	n. actress (gender motion)
glumičin	adj. possessive of *glumica (actresse)*
glumčev	adj. possessive of *glumac (actor)*
glumčina	n. augmentative of *glumac* (actor)
glumčić	n. diminutive of *glumac* (actor)

Redundancy (and incompleteness)

In dictionary RMS6:

gluma	n. 1. acting art
glumac	n. actor
glumački, -a, -o	adj. related to actors
glumački	adv. in a way of actors
glumiti	v. to act
glumica	n. actress (gender motion)
glumičin	adj. possessive of *glumica (actresse)*
glumčev	adj. possessive of *glumac (actor)*
glumčina	n. augmentative of *glumac* (actor)
glumčić	n. diminutive of *glumac* (actor)

Prolex 3.0

Beograd	N
beogradski	A+Rel
Beograđanin	N+Hab:m
beograđanski	A+Rel
Beograđaninov	A+Hab+Poss
Beograđanka	N+Hab:f
Beograđankin	A+Hab+Poss

Regular derivation

For the animate (human) nouns:

- gender motion: *glumac --> glumica*
- possessive adjectives: *glumac --> glumčev*
- diminutives: *glumac --> glumčić*
- augmentatives: *glumac --> glumčina*

For adjectives:

- negation: *pristojan --> nepristojan*
- abstract nouns: *mudar -> mudrost*

For verbs:

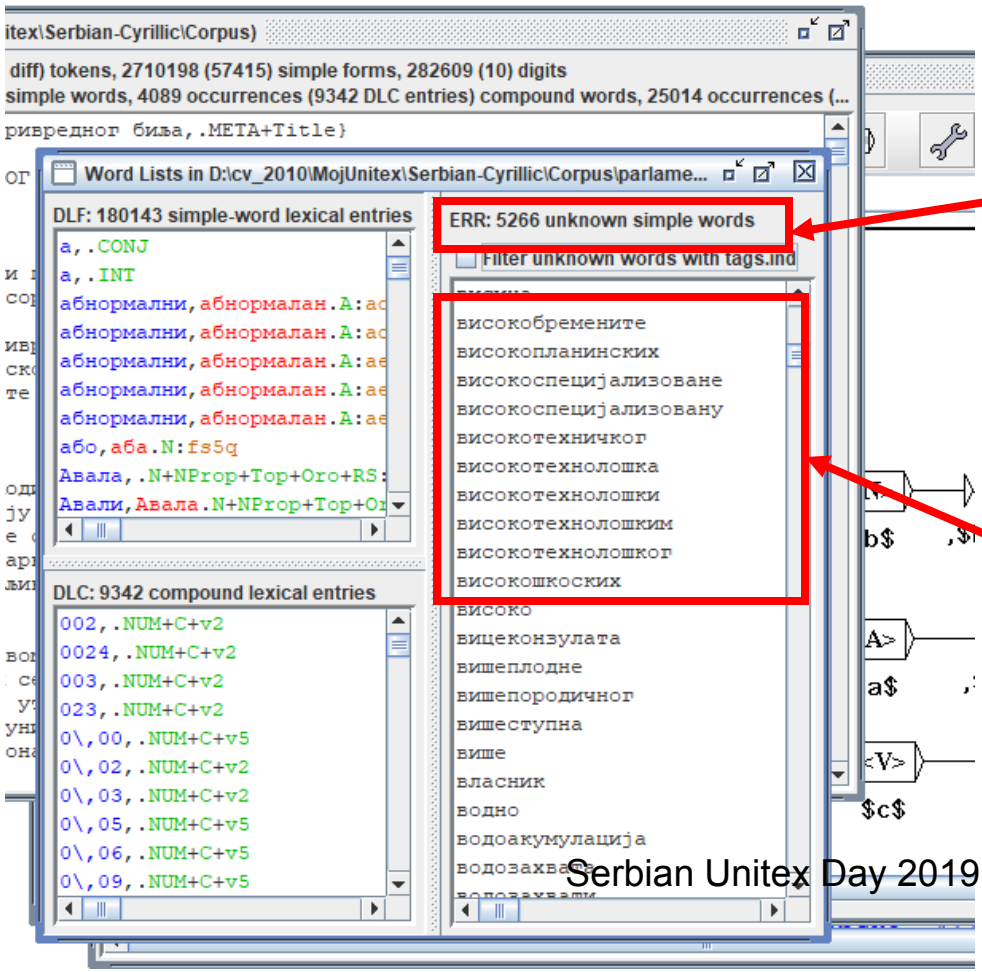
- verbal noun: *pevati --> pevanje*
- verb forms acting as adjectives. *atrofirati --> atrofirajući*

Derivation in Unitex

1. Morphological graphs
2. Super - lemma

Morphological graphs

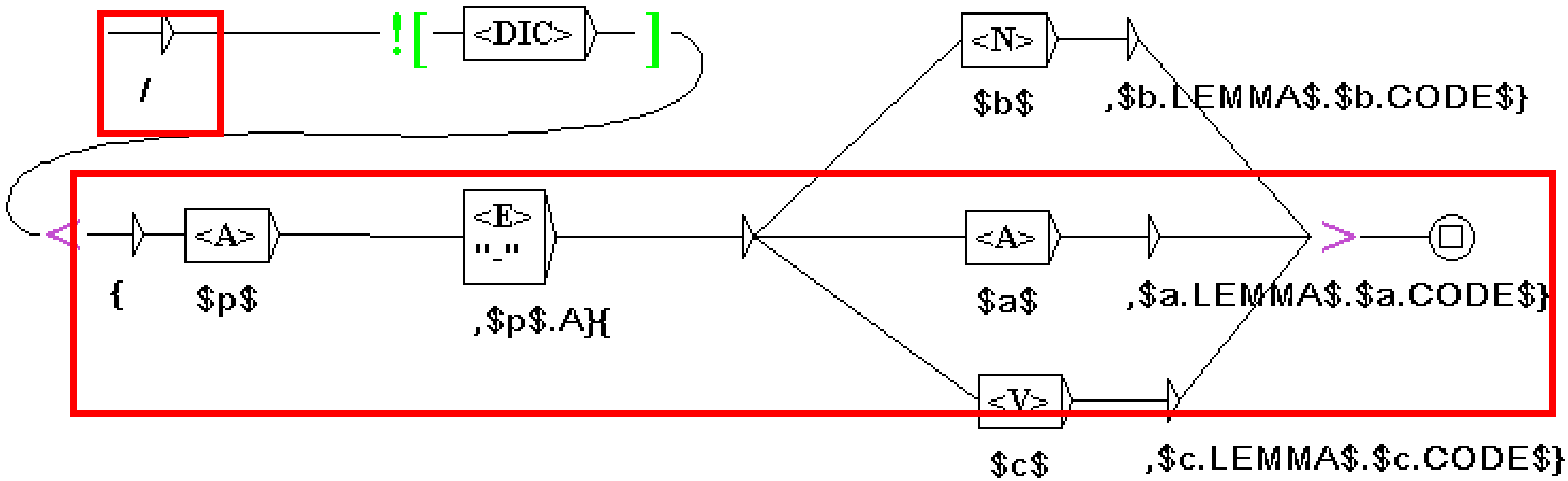
The corpus of Serbian laws (2013) **~ 3MW**



in ERR ~ 5266

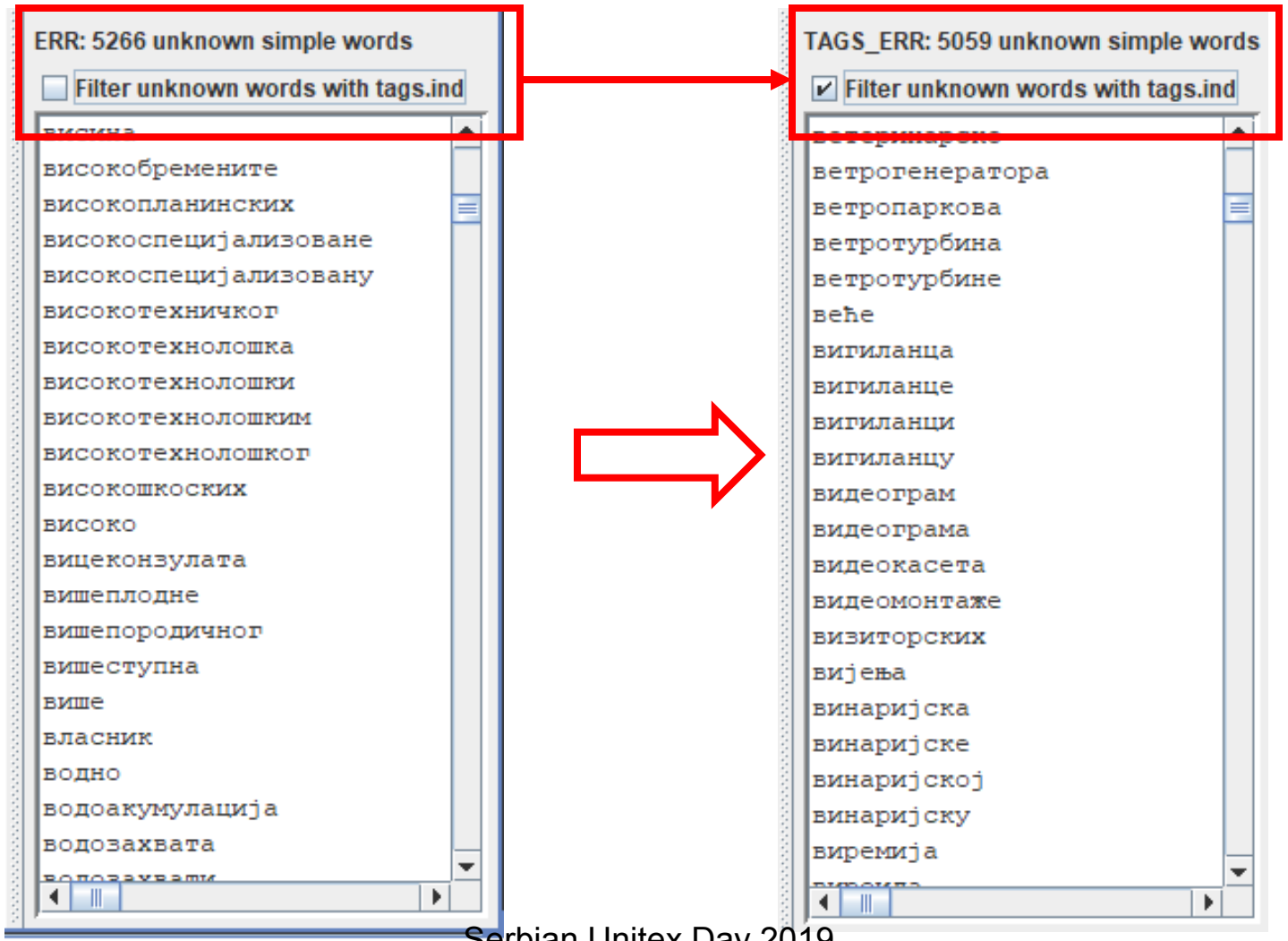
visokobremenite
visokoplaninskih
visokospecijalizovane
visokotehničkog
visokotehnološkog
...

~ A<E>A



}Члан 26. Забрањено је превозити: {S}1) /{високо,.A}{бремените,бременит.A:aefp5g} женке у перио
}Члан 26. Забрањено је превозити: {S}1) /{високо,.A}{бремените,бременит.A:aefs2g} женке у перио
}Члан 26. Забрањено је превозити: {S}1) /{високо,.A}{бремените,бременит.A:aefw2g} женке у перио
}Члан 26. Забрањено је превозити: {S}1) /{високо,.A}{бремените,бременит.A:aefw4g} женке у перио
}Члан 26. Забрањено је превозити: {S}1) /{високо,.A}{бремените,бременит.A:aemp4g} женке у перио
а Члан 17. Заштита екосистема (шумских, /{високо,.A}{планинских,планински.A+PosQ:aefp2g}, воден
а Члан 17. Заштита екосистема (шумских, /{високо,.A}{планинских,планински.A+PosQ:aemp2g}, воден
а Члан 17. Заштита екосистема (шумских, /{високо,.A}{планинских,планински.A+PosQ:aenp2g}, воден
датке који се односе на: {S}1) обављање /{високо,.A}{специјализоване,специјализован.A+PP+DER=OvatiIr
датке који се односе на: {S}1) обављање /{високо,.A}{специјализоване,специјализован.A+PP+DER=OvatiIr
датке који се односе на: {S}1) обављање /{високо,.A}{специјализоване,специјализован.A+PP+DER=OvatiIr
датке који се односе на: {S}1) обављање /{високо,.A}{специјализоване,специјализован.A+PP+DER=OvatiIr
датке који се односе на: {S}1) обављање /{високо,.A}{специјализоване,специјализован.A+PP+DER=OvatiIr
датке који се односе на: {S}1) обављање /{високо,.A}{специјализоване,специјализован.A+PP+DER=OvatiIr
датке који се односе на: {S}1) обављање /{високо,.A}{специјализоване,специјализован.A+PP+DER=OvatiIr
датке који се односе на: {S}1) обављање /{високо,.A}{специјализоване,специјализовати.V+Imperf+Perf+T
датке који се односе на: {S}1) обављање /{високо,.A}{специјализоване,специјализовати.V+Imperf+Perf+T
ика је здравствена установа која обавља /{високо,.A}{специјализовану,специјализован.A+PP+DER=OvatiIr

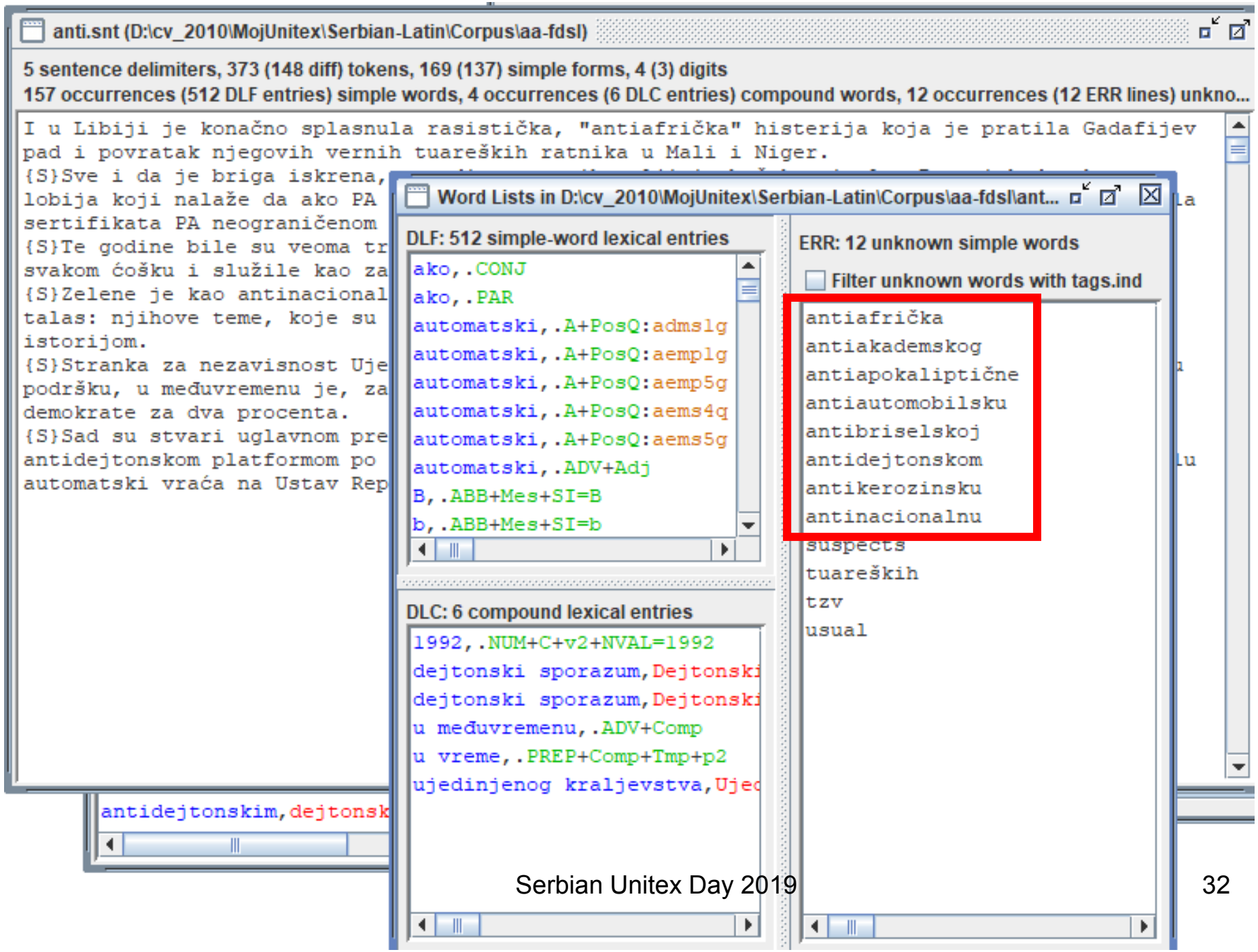
tags.ind

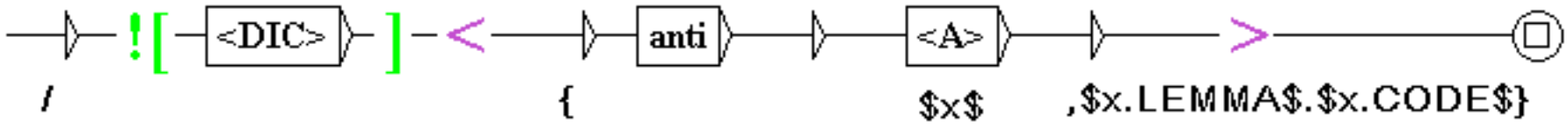

ERR: 5266 unknown simple words
Filter unknown words with tags.ind
високобремените
високопланинских
високоспецијализоване
високоспецијализовану
високотехничког
високотехнолошка
високотехнолошки
високотехнолошким
високотехнолошког
високошкоских
високо
вицеконзулата
вишеплодне
вишепородичног
вишеступна
више
власник
водно
водоакумулација
водозахвата
TAGS_ERR: 5059 unknown simple words
Filter unknown words with tags.ind
ветрогенератора
ветропаркова
ветротурбина
ветротурбине
веће
вигиланца
вигиланце
вигиланци
вигиланцу
видеограм
видеограма
видеокасета
видеомонтаже
визиторских
вијења
винаријска
винаријске
винаријској
винаријску
виремија

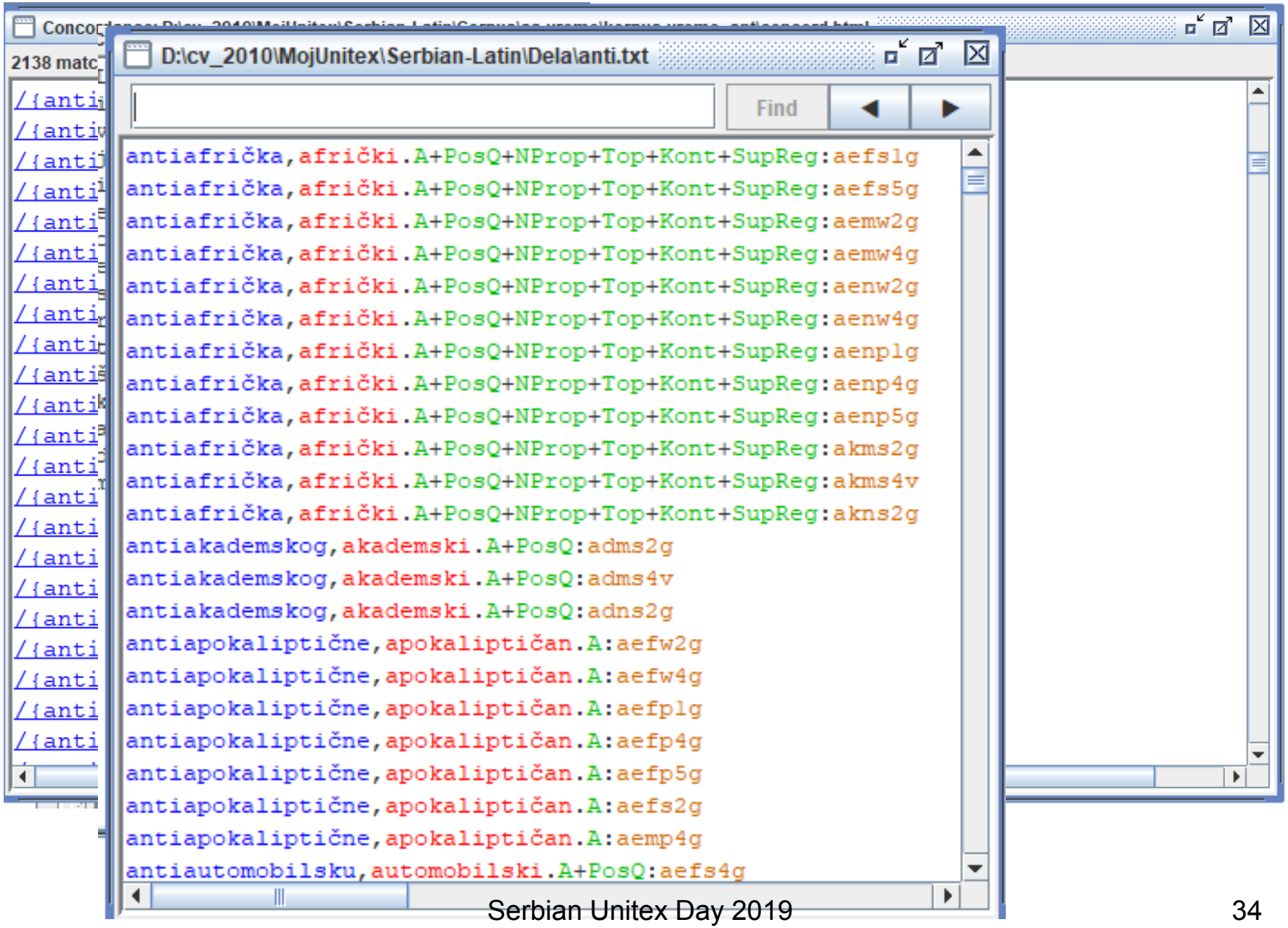
Another way

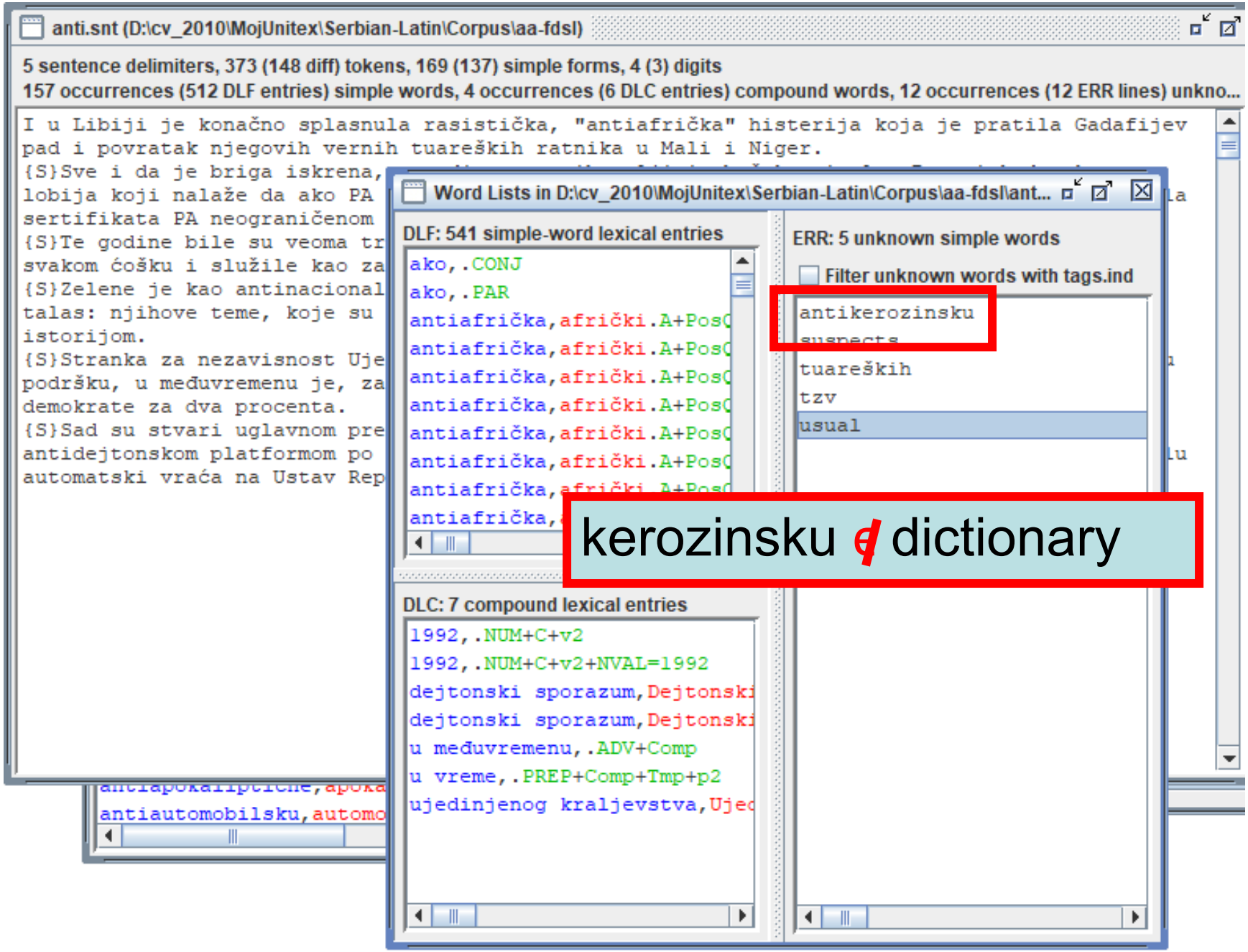
- Examples of unknown words from the ***Vreme*** corpus having the prefix ***anti-***

čno splasnula rasistička, "antiafrička" histerija koja je pratila G
vatljivi, baš kao i plan B antiakademskog lobija koji nalaže da ako
dine bile su veoma tražene antiapokaliptične pilule i maske koje su
ene je kao antinacionalnu, antiautomobilsku i antikerozinsku partij
nu je, zahvaljujući svojoj antibriselskoj politici, nadmašila Liber
provinijencija: od onih sa antidejtonskom platformom po kojima se,
onalnu, antiautomobilsku i antikerozinsku partiju zahvatio ledeni t
j Zemlje. {S}Zelene je kao antinacionalnu, antiautomobilsku i antik


anti.snt (D:\cv_2010\MojUnitex\Serbian-Latin\Corpus\aa-fdsl)
5 sentence delimiters, 373 (148 diff) tokens, 169 (137) simple forms, 4 (3) digits
157 occurrences (512 DLF entries) simple words, 4 occurrences (6 DLC entries) compound words, 12 occurrences (12 ERR lines) unkno...
I u Libiji je konačno splasnula rasistička, "antiafrička" histerija koja je pratila Gadafijev pad i povratak njegovih vernih tuareških ratnika u Mali i Niger.
{S}Sve i da je briga iskrena,
lobija koji nalaže da ako PA
sertifikata PA neograničenom
{S}Te godine bile su veoma tr
svakom ćošku i služile kao za
{S}Zelene je kao antinacional
talas: njihove teme, koje su
istorijom.
{S}Stranka za nezavisnost Uje
podršku, u međuvremenu je, za
demokrate za dva procenta.
{S}Sad su stvari uglavnom pre
antidejtonskom platformom po
automatski vraća na Ustav Rep
antidejtonskim,dejtonsk
Word Lists in D:\cv_2010\MojUnitex\Serbian-Latin\Corpus\aa-fdsl\ant...
DLF: 512 simple-word lexical entries
ako,.CONJ
ako,.PAR
automatski,.A+PosQ:adms1g
automatski,.A+PosQ:aemp1g
automatski,.A+PosQ:aemp5g
automatski,.A+PosQ:aems4q
automatski,.A+PosQ:aems5g
automatski,.ADV+Adj
B,.ABB+Mes+SI=B
b,.ABB+Mes+SI=b
DLC: 6 compound lexical entries
1992,.NUM+C+v2+NVAL=1992
dejtonski sporazum,Dejtonski
dejtonski sporazum,Dejtonski
u međuvremenu,.ADV+Comp
u vreme,.PREP+Comp+Tmp+p2
ujedinjenog kraljevstva,Ujed
ERR: 12 unknown simple words
Filter unknown words with tags.ind
antiafrička
antiakademskog
antiapokaliptične
antiautomobilsku
antibriselskoj
antidejtonskom
antikerozinsku
antinacionalnu
suspects
tuareških
tzv
usual


!
[
<DIC>
]
<
anti
<A>
>
/
{
x
,$x.LEMMA$.$x.CODE$}


D:\cv_2010\MojUnitex\Serbian-Latin\Dela\anti.txt
Find
antiafrička,afrički.A+PosQ+NProp+Top+Kont+SupReg:aefs1g
antiafrička,afrički.A+PosQ+NProp+Top+Kont+SupReg:aefs5g
antiafrička,afrički.A+PosQ+NProp+Top+Kont+SupReg:aemw2g
antiafrička,afrički.A+PosQ+NProp+Top+Kont+SupReg:aemw4g
antiafrička,afrički.A+PosQ+NProp+Top+Kont+SupReg:aenw2g
antiafrička,afrički.A+PosQ+NProp+Top+Kont+SupReg:aenw4g
antiafrička,afrički.A+PosQ+NProp+Top+Kont+SupReg:aenp1g
antiafrička,afrički.A+PosQ+NProp+Top+Kont+SupReg:aenp4g
antiafrička,afrički.A+PosQ+NProp+Top+Kont+SupReg:aenp5g
antiafrička,afrički.A+PosQ+NProp+Top+Kont+SupReg:akms2g
antiafrička,afrički.A+PosQ+NProp+Top+Kont+SupReg:akms4v
antiafrička,afrički.A+PosQ+NProp+Top+Kont+SupReg:akns2g
antiakademskog,akademski.A+PosQ:adms2g
antiakademskog,akademski.A+PosQ:adms4v
antiakademskog,akademski.A+PosQ:adns2g
antiapokaliptične,apokaliptičan.A:aefw2g
antiapokaliptične,apokaliptičan.A:aefw4g
antiapokaliptične,apokaliptičan.A:aefp1g
antiapokaliptične,apokaliptičan.A:aefp4g
antiapokaliptične,apokaliptičan.A:aefp5g
antiapokaliptične,apokaliptičan.A:aefs2g
antiapokaliptične,apokaliptičan.A:aemp4g
antiautomobilsku,automobilski.A+PosQ:aefs4g


anti.snt (D:\cv_2010\MojUnitex\Serbian-Latin\Corpus\aa-fdsl)
5 sentence delimiters, 373 (148 diff) tokens, 169 (137) simple forms, 4 (3) digits
157 occurrences (512 DLF entries) simple words, 4 occurrences (6 DLC entries) compound words, 12 occurrences (12 ERR lines) unkno...
I u Libiji je konačno splasnula rasistička, "antiafrička" histerija koja je pratila Gadafijev pad i povratak njegovih vernih tuareških ratnika u Mali i Niger.
{S}Sve i da je briga iskrena, lobija koji nalaže da ako PA sertifikata PA neograničenom
{S}Te godine bile su veoma tr svakom ćošku i služile kao za
{S}Zelene je kao antinacional talas: njihove teme, koje su istorijom.
{S}Stranka za nezavisnost Uje podršku, u međuvremenu je, za demokrate za dva procenta.
{S}Sad su stvari uglavnom pre antidejtonskom platformom po automatski vraća na Ustav Rep
antiautomobilsku,automo
Word Lists in D:\cv_2010\MojUnitex\Serbian-Latin\Corpus\aa-fdsl\ant...
DLF: 541 simple-word lexical entries
ako,.CONJ
ako,.PAR
antiafrička,afrički.A+Pos
antiafrička,afrički.A+Pos
antiafrička,afrički.A+Pos
antiafrička,afrički.A+Pos
antiafrička,afrički.A+Pos
antiafrička,afrički.A+Pos
antiafrička,
DLC: 7 compound lexical entries
1992,.NUM+C+v2
1992,.NUM+C+v2+NVAL=1992
dejtonski sporazum,Dejtonski
dejtonski sporazum,Dejtonski
u međuvremenu,.ADV+Comp
u vreme,.PREP+Comp+Tmp+p2
ujedinjenog kraljevstva,Ujed
ERR: 5 unknown simple words
Filter unknown words with tags.ind
antikerozinsku
tuareških
tzv
usual
kerozinsku dictionary

„The super-lemma“

- ... as a mean to gather forms obtained by the regular derivation.

- We should remember that, as an example, the regular derivation produces from the lemma ***programer*** forms:

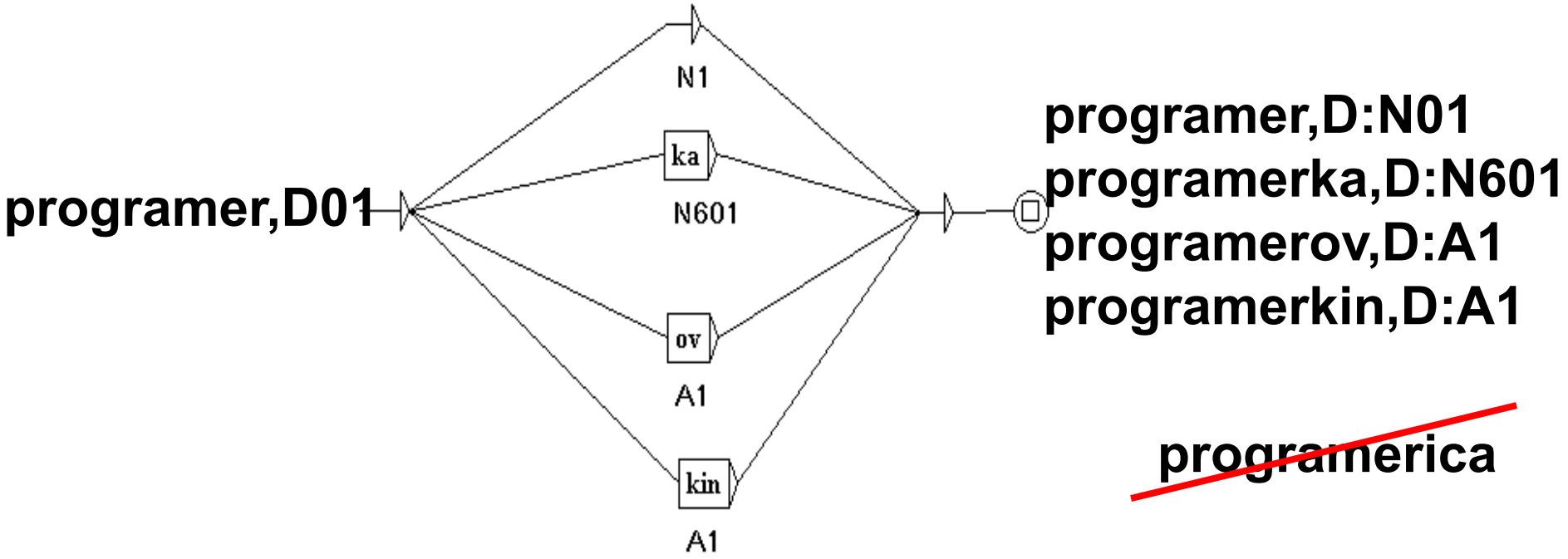
  gender motion: ***programer --> programerka***

  possessive adjectives: ***programer--> programerov***

  diminutives: ***programer --> programerčić***

  augmentatives: ***programer --> programerčina...***

„The super-lemma“ - graph D01



...N/A+...+super-lemma=“programer”

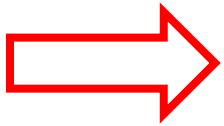
The super-lemma *programer*

programer,D:N01

programerka,D:N601

programerov,D:A1

programerkin,D:A1



DELAS:

programer,N01

programerka,N601+sI=“programer”+FG

programerov,A1+Poss+sI=“programer”

programerkin,A1+Poss+sI=“programer”

DELAF

The classification problems

The processing of unknown words can be divided to three general cases:

1. A really unknown word (neologism, occasionalism, a foreign word,...) – a specific processing
2. A word from DELAS to which a prefix is added (a morphological graph)
3. A word derived by suffixation from a word from DELAS (a generalized inflectional graph)

Possible consequences

- In the example of the noun ***glumac*** and its derivatives, instead of having ten or more separate entries that refer to ***glumac***, only one entry appears with a „code“ that points to possible derivatives.
- Problems? Re-defining of inflectional graphs?
- Applications? Maintenance of dictionaries, but also avoiding redundant lexicographic descriptions.

Thank you!