Identification of Croatian Light Verb Constructions with NooJ

Krešimir Šojat, Kristina Kocijan and Božo Bekavac

University of Zagreb, Croatia

Outline

- Multiword expressions
  - Types and criteria
  - Verbal multiword expressions
  - Light verb constructions
- Application in NooJ
  - NooJ grammars
    - 5 types of syntactic grammars
- Results and future work

Multiword expressions (MWEs)

- MWEs
  - refer to various types of constructions consisting of two or more words
  - they act as a single unit at some level of analysis (syntactic, semantic)
    - MWEs: "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al, 2002)
- Identification and annotation of MWEs in Croatian corpora and treebanks → so far little attention / work
    - although these constructions pose a challenge for various NLP tasks (tagging, parsing)

MWEs

- Semantics
  - the meaning of MWEs can vary from more or less compositional to completely idiosyncratic
    - black coffee / to make a decision / to kick the bucket

- MWEs usually include

  noun compounds, multiword named entities, different types of complex verb phrases, idioms and others

Baldwin and Kim (2010) / Rosen et al (2016)

- Nominal MWEs
  - Multiword named entities
  - NN compounds
    - other nominal MWEs
- Verbal MWEs
  - Phrasal verbs
  - Light verb constructions
  - VP idioms
    - other verbal MWEs
- Prepositional MWEs / Adjectival MWEs / MWEs of other categories
- Proverbs

MWEs - fixed VS. flexible

- fixed / frozen expressions
  - paradigmatic selection of elements and their syntagmatic order is never altered
  - lexicalized phrases
  - idiosyncratic syntax and semantics
    - idioms – kick the bucket
    - fixed collocations - by and large, in short, cash and carry

MWEs – (relatively) flexible

- can be modified to a certain degree
- syntactically and semantically compositional, but co-occur with markedly high frequency
  - semi-fixed expression – compound nouns (attorney general, mineral water)
  - syntactically-flexible expression – light-verb constructions
    - morphosyntactic properties of elements –
      - take / takes / is taking / took a walk
      - donijeti (*pf*) odluku (*sg*) – to make a decision
      - donijeti (*pf*) odluke (*pl*) – to make decisions
      - donositi (*impf*) odluku (*sg*) – to make a desicion
      - donositi (*impf*) odluke (*pl*) - to make a desicion
    - selection / insertion of elements
      - donijeti važnu odluku – to make an important decision / to reach a decison etc.

Verbal MWEs in Croatian

- In this research we focus on **light verb constructions** in Croatian
  - verbal MWEs → subgroups
    - phrasal verbs
    - **light verb constructions**
    - VP idioms and
    - other verbal MWEs
  - a) there is no previous research done on the identification and annotation of light verb constructions in Croatian corpora / treebanks
  - b) there is no resource / database which would enable an extensive / further research of MWEs in Croatian

Light Verb Constructions (LVCs) - features

- *donijeti odluku*
    - to make a decision / to reach a decision
  - made up of a verbal and a nominal component
  - the nominal component consists of a NP or a PP
  - nouns in NPs are usually derived from verbal stems
    - usually in accusative case
  - light verbs have entirely or partially lost their lexical meaning and the meaning of the whole construction is actually expressed by NPs or PPs
- LVCs can frequently be substituted with a single "heavy " verb
  - e.g. *donijeti odluku – odlučiti*
    - the meanings of the LVC and their paraphrases, i.e. semantically full verbs, often do not exactly correspond

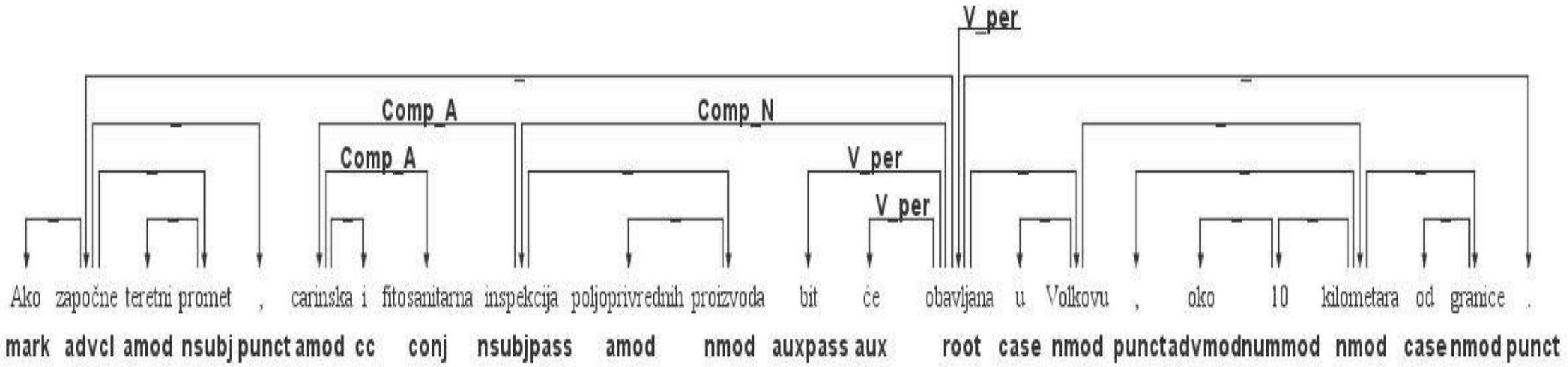
LVCs in Croatian

- LVCs in Croatian are syntactically flexible
  - inversion of elements – NPs and PPs can appear before light verbs
  - modifying elements can be inserted – adjectives, pronouns ets.
  - light verbs can be inflected
  - light verbs can be passivized
  - light verbs are marked as perfectives or imperfectives
  - In some LVCs nouns can be used both in singular and plural and/or in different cases

MWEs - Annotation in corpus

- Universal Dependency (UD) Treebank for Croatian
  - The UD treebank version used in this experiment consisted of 3557 sentences
  - additional / independet level of annotation
  - manuallly annotated
  - the total number of annotated LVCs in the treebank is 466

LVC - treebank



LVC database

Light verb	AUX	AUX	V_per	AUX	AUX	REF	AUX	PREP	A	A	N	N - lema	PREP	N	Example
dobiti			dobio	je							nagradu	nagrada			dobio V_per je V_per nagradu Comp_N
dobiti			dobio	je						posebnu	nagradu	nagrada			dobio V_per je V_per posebnu Comp_A nagradu Comp_N
dobiti			dobio	je							nagradu	nagrada	za		Nagradu Comp_N je V_per dobio V_per za Comp_Prep
dobiti			dobio								naknadu	naknada			dobio V_per naknadu Comp_N
dobiti			dobili							punu	neovisnost	neovisnost	od		dobili V_per punu Comp_A neovisnost Comp_N od Comp_Prep
dobiti			dobio	je							odobrenje	odobrenje			dobio V_per je V_per odobrenje Comp_N
dobiti			dobiti								odobrenje	odobrenje			treba V_comp_fin dobiti V_per odobrenje Comp_N
dobiti			dobiti								posao	posao			dobiti V_per posao Comp_N
dobiti			dobiti							snažnu	potporu	potpora			dobiti V_per snažnu Comp_A potporu Comp_N

5 types of LVCs

- Type 1
  - verb + PP — dobiti na težiti / to gain on weight
- Type 2
  - verb + NP + preposition — baciti sumnju na / to cast doubt on
- Type 3
  - verb + PP + preposition — biti u skladu s / to be in accoradnce with
- Type 4
  - verb + NP + PP — izdati nalog za uhićenje / to issue a warrant for arrest
- Type 5
  - verb + NP — doživjeti rast / to achieve growth

LVC Dictionary

LVCs - types

```
#########
# LVCType1 with PP
#########
dobiti,V+FXC+LVCType1+PREP=na+PREPX=težina+FLX=ČUTI

#########
# TYPE2 with NP  + preposition
#########
baciti,V+FXC+LVCType2+SUFX=sumnja+PREP=na+FLX=BACITI

#########
# TYPE3 with PP + preposition
#########
poslužiti,V+FXC+LVCType3+AFIX=kao+SUFX=temelj+PREP=za+FLX=UGOJITI

#########
# TYPE4 with NP + PP (PREP+SUFXB)
#########
izdati,V+FXC+LVCType4+SUFX=nalog+PREP=za+SUFXB=uhićenje+FLX=IZDATI

#########
# LVCType5 with NP
#########
bilježiti,V+FXC+LVCType5+SUFX=rezultat+FLX=BILJEŽITI
```

Number of dic entries vs number of LVCs

STATISTICS

- Type

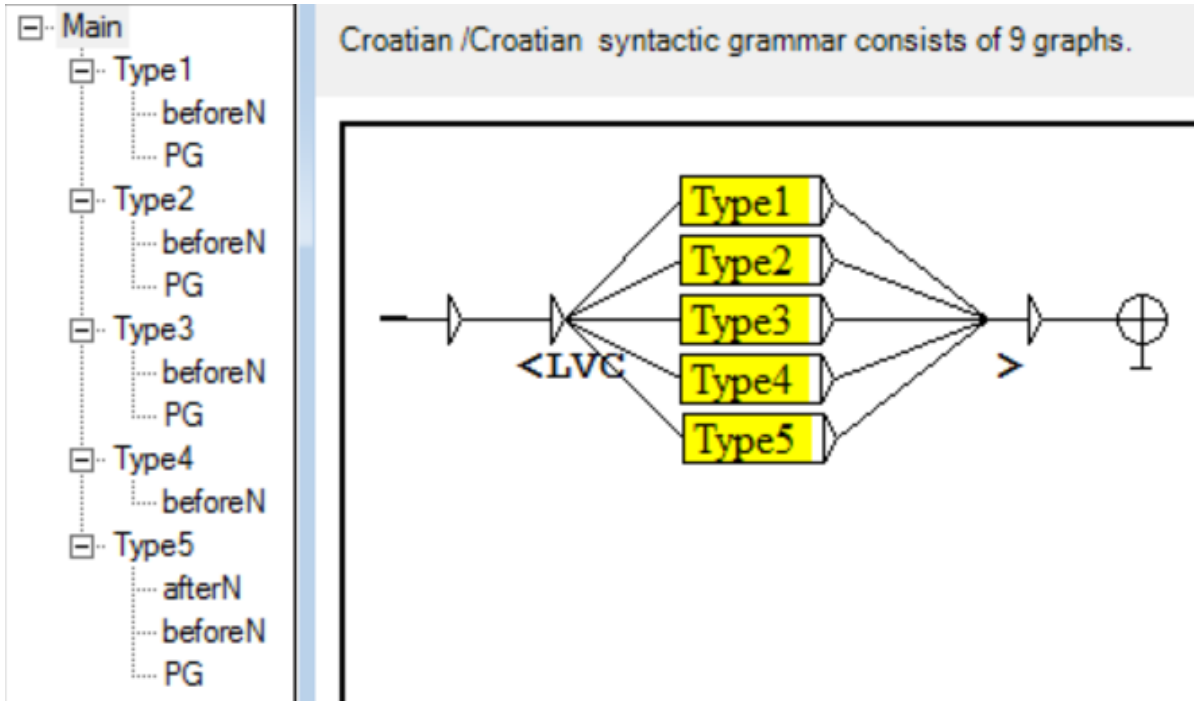
1. 53 entries – 91 LVC
2. 45 entries – 51 LVC
3. 5 entries – 5 LVC
4. 4 entries – 4 LVC
5. 191 entries – 315 LVC

WHY IS THERE A DIFFERENCE IN NUMBERS?

- Verb → head of construction
- Same verb –> different expressions
  - voditi,V+FXC+LVCType5+FLX=SUDITI +SUFX=borba+SUFX=razgovor+SUFX=rasprava+SUFX=pregovori
  - to lead:
    - SUFX=fight
    - SUFX=conversation
    - SUFX=discussion
    - SUFX=negotiations
  - Entry ‘imati’ <to have> has 32 possible expressions
    - the most frequent / productive

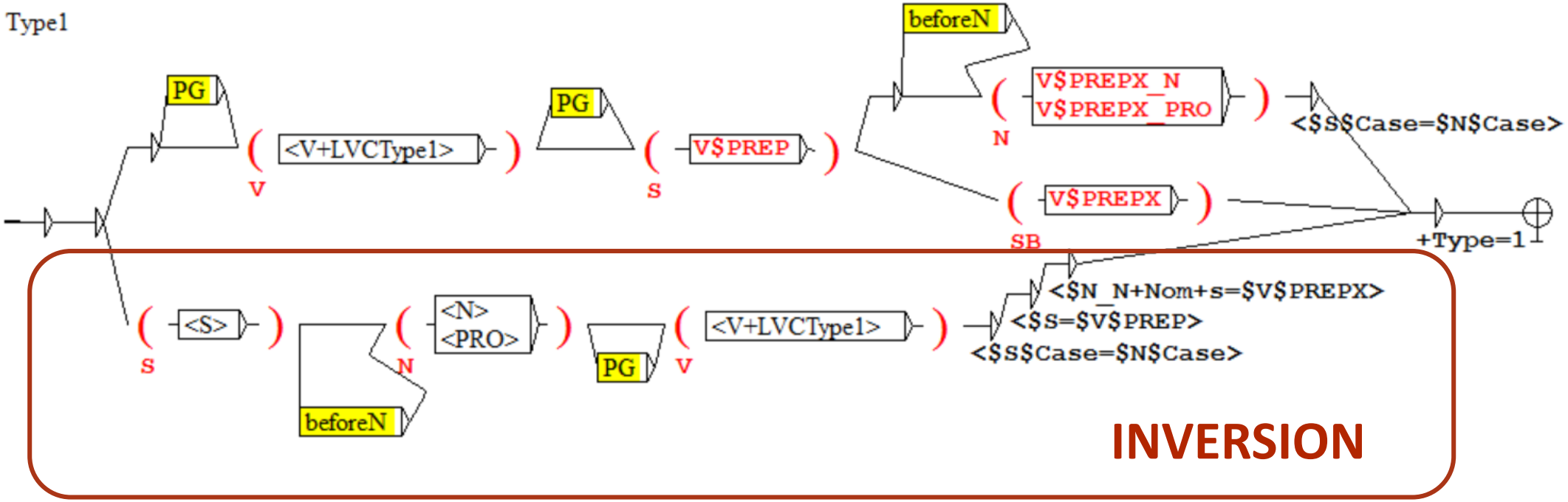
LVC Syntactic Grammar

Main LVC grammar

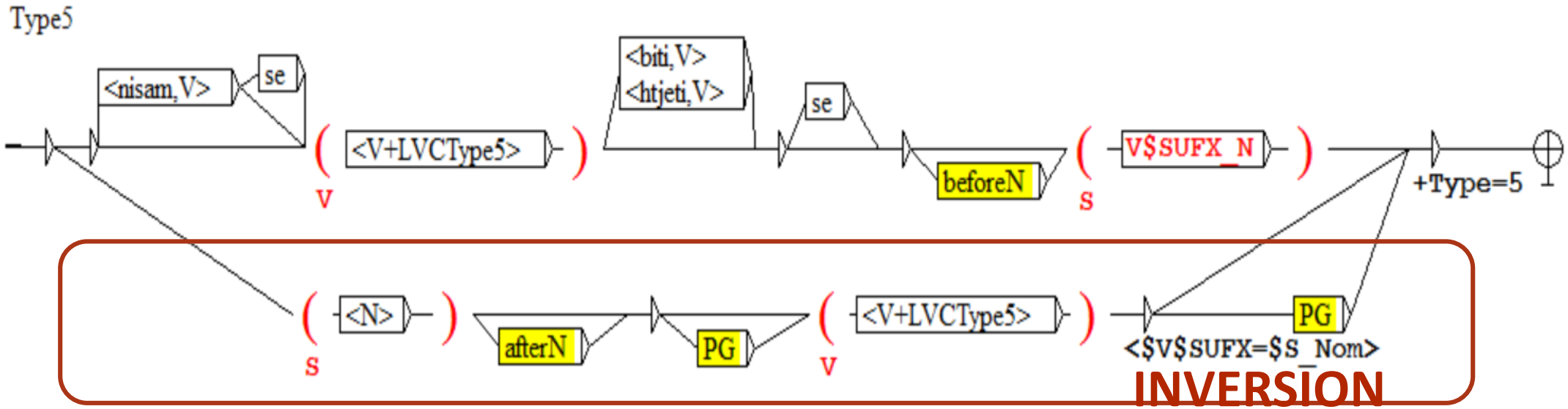


- 5 first order subgraphs correspond to LVC Types
- 3 second order subgraphs
  - beforeN (adjectives, pronouns, numbers)
  - afterN (1 or more NPs in genitive)
  - PG (auxiliary verbs *to be* and *to have* for complex VP, negation, reflexive pronoun '*se*')

Type 1

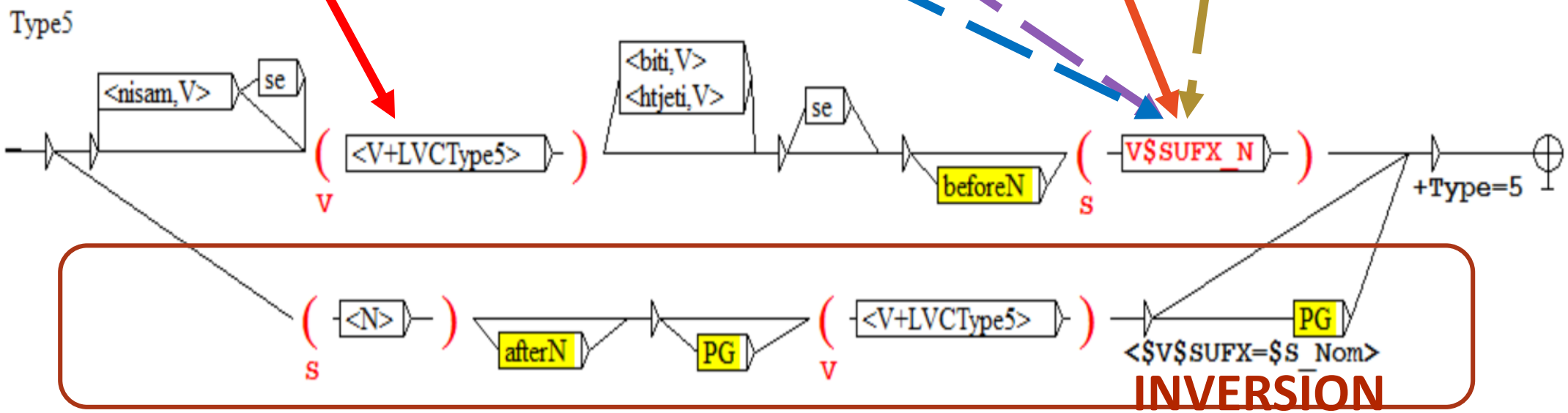

Type1
PG
<V+LVCType1>
V
PG
V$PREP
S
beforeN
V$PREPX_N
V$PREPX_PRO
N
<SCase=NCase>
V$PREPX
SB
+Type=1
<S>
S
<N>
<PRO>
N
beforeN
PG
<V+LVCType1>
V
<$N_N+Nom+s=$V$PREPX>
<$S=$V$PREP>
<SCase=NCase>
INVERSION

Type 5


Type5
<nisam,V>
se
<V+LVCType5>
V
<biti,V>
<htjeti,V>
se
beforeN
V$SUFX_N
S
+Type=5
<N>
S
afterN
PG
<V+LVCType5>
V
PG
<VSUFX=$S_Nom>
INVERSION

- voditi,V+FXC+LVCType5+FLX=SUDITI +SUFX=borba+SUFX=razgovor +SUFX=rasprava+SUFX=pregovor

Type 5



LVC Grammar Results

Measures and problems

MEASURES PER TYPE AND OVERALL PERFORMANCE

TYPE	Precision	Recall	F-measure
1 (100)	0.99	1	0.995
2 (54)	1	1	1
3 (4)	1	1	1
4 (3)	1	1	1
5 (369)	0.94	1	0.97
OVERALL	0.96	1	0.98

PROBLEMS

- Američke Države i Rusija predstavili *su suprotstavljena stajališta*.
  - Found: *su suprotstavljena stajališta*
  - ***Should find**: predstavili su stajališta*
- Mogla bi postati prva zemlja EU s nuklearnom elektranom…
  - Found: *postati prva zemlja*
  - ***No LVC***

Conclusion and future work

- LVCs in Croatian – divided into 5 types
- syntactic grammars – identify all 5 types
  - identify LVCs when modifying elements are inserted (adjectives, pronouns…)
  - identify LVCs when NPs and PPs occur before light verbs
- Future research – from NPs and PPs towards light verbs
  - to determine all light verbs that can form a LVC with a particular NP or PP
    - e.g. POTPORA – dati / pružiti / imati / izraziti POTPORU
  - to determine whether other elements can be inserted into LVCs
    - e.g. adverbs / particles / relative clauses

THANK YOU FOR
YOUR ATTENTION!