Incorporating Valency Lexicon into BURGER

Petya Osenova and Kiril Simov

DELPH-IN, Sofia, 2012

Plan of the Talk

- The BulTreeBank-driven Valency Lexicon: Overview
- Towards incorporation within BURGER
- Conclusions and future work

Our Aim

- Constructing a valency lexicon, which:
  - covers the verbs in the syntactically analyzed corpus of Bulgarian – *BulTreeBank* (www.bultreebank.org)
  - adopts surface syntactic structure
  - consists of ontological constraints
- Incorporation of the result in BURGER

A Variety of Valency Lexicon Creation Projects

- (Hinrichs and Telljohann 2009) - German
- (Zabokrtsky and Lopatkova 2007) – Czech
- (Bielický and Smrž 2008) – Arabic
- (Agic et al. 2010) – Croatian
- (Amussen and Ørsnes 2005) – Danish
- (McGillivray and Passarotti 2009) - Latin

Lexicon Coverage

- the whole set of **3283** lemmas in BulTreeBank
- The number of distinct valence frames for these lemmas is **6469**
- the average is almost **3 valence frames** per lemma

Bulgarian Ontology-based Lexicon

- The valence lexicon is a part the Bulgarian Ontology-based Lexicon (BOL) – (Simov and Osenova, 2010).
- The current version of BOL is based on DOLCE ontology extended with concepts from OntoWordNet - a version 1.6 of WordNet aligned to DOLCE
- Intersection of EuroWordNet Base Concepts and Core WordNet (**1504 synsets**)
- Extended with lexical units extracted from the Bulgarian National Reference Corpus (**www.webclark.org**).

OntoValence Lexicon Extraction and Manipulation

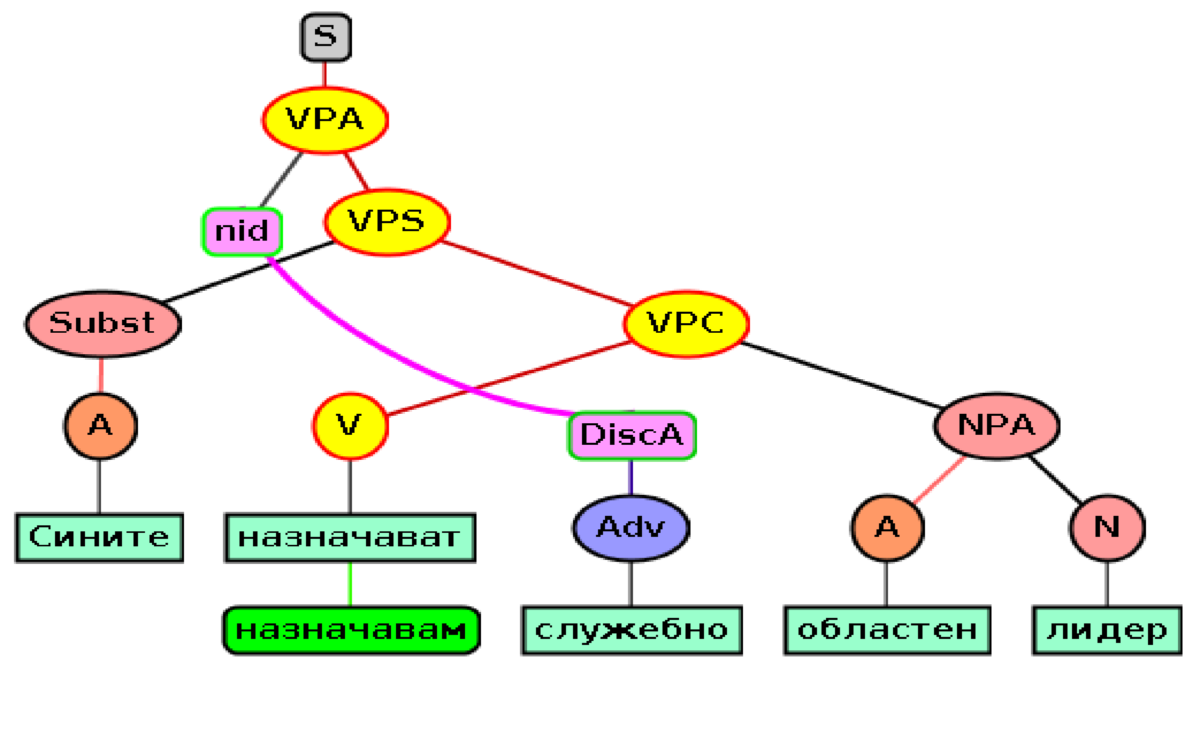
- All the verbs have been extracted together with the sentences they have been used in
- Then they have been lemmatized and sorted by the lemma marker
- A default valence frame has been inserted, which presents a predicate with a **SUBJ**, **DIROBJ** and **INDOBJ**

Why Such an Approach?

- The pre-annotated frames in BulTreeBank might differ syntactically from our present postulations of constructing valence frames due to an error or different view;
- The pure copying of the annotated frame, which might be considered a trivial step, has been abandoned, since our aim is to add also ontological constraints.

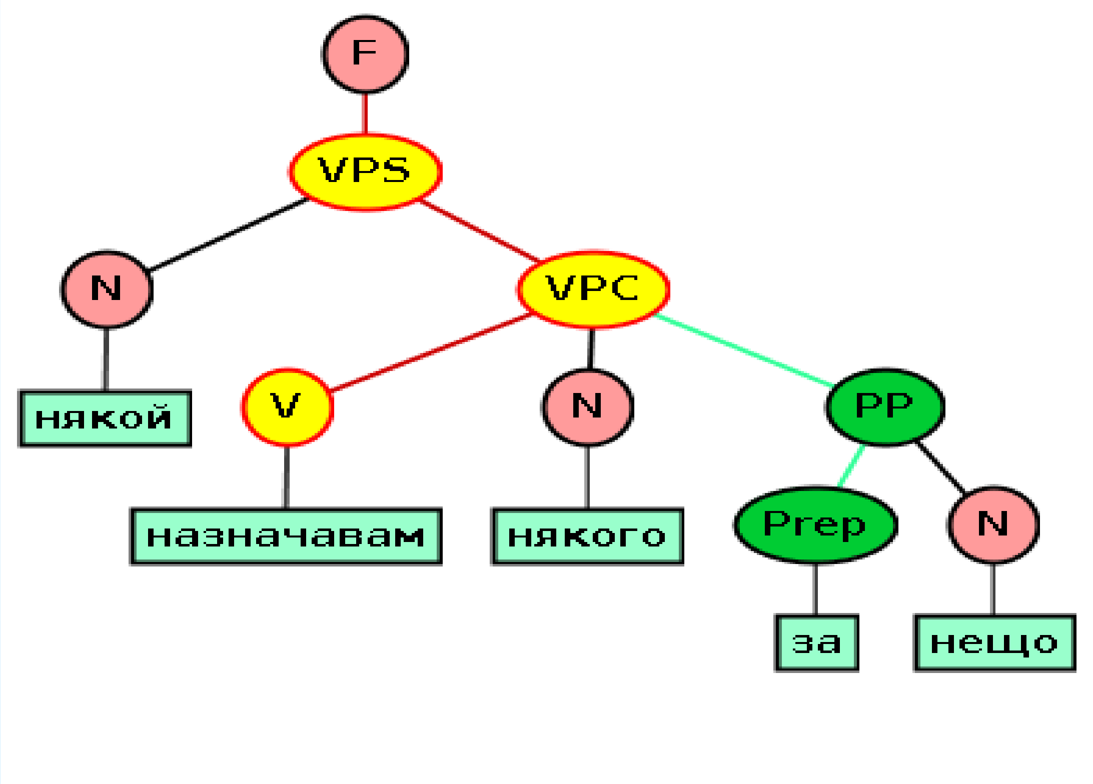
Original representation of a sentence tree

Gloss: *Blue-the appoint officially area leader*.
Translation: *The blue team ex officio appoints an area leader*.



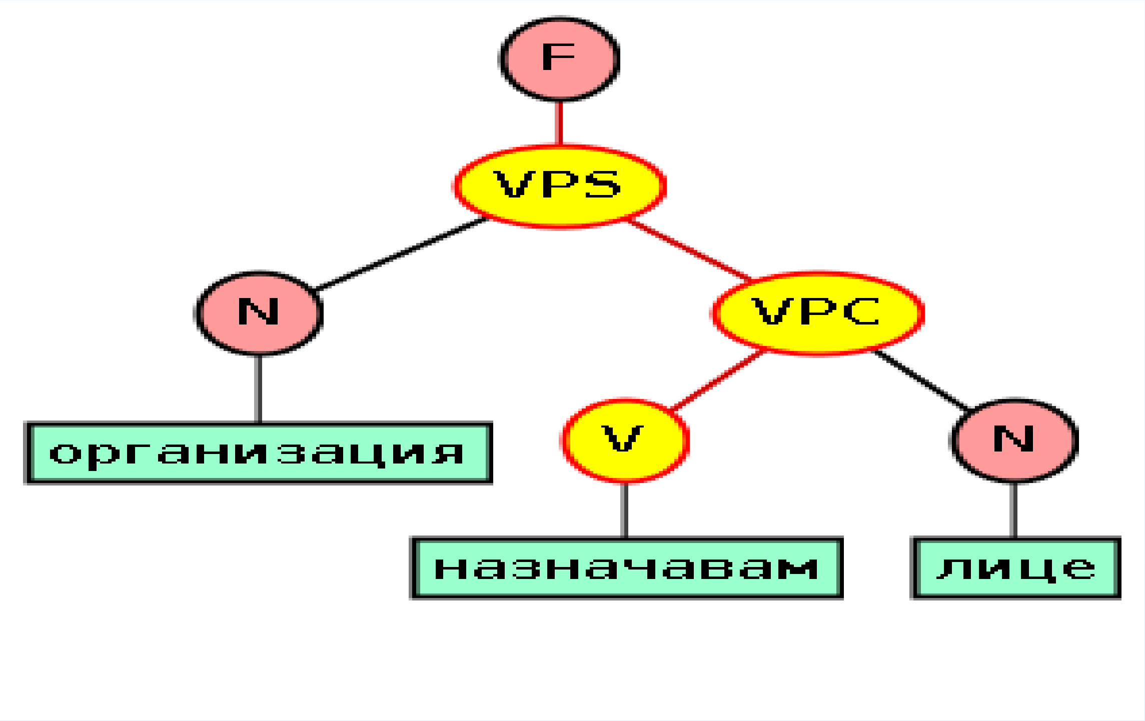
Default inserted tree

[SOMEBODY appoints SOMEONE for SOMETHING]



Resulting Frame

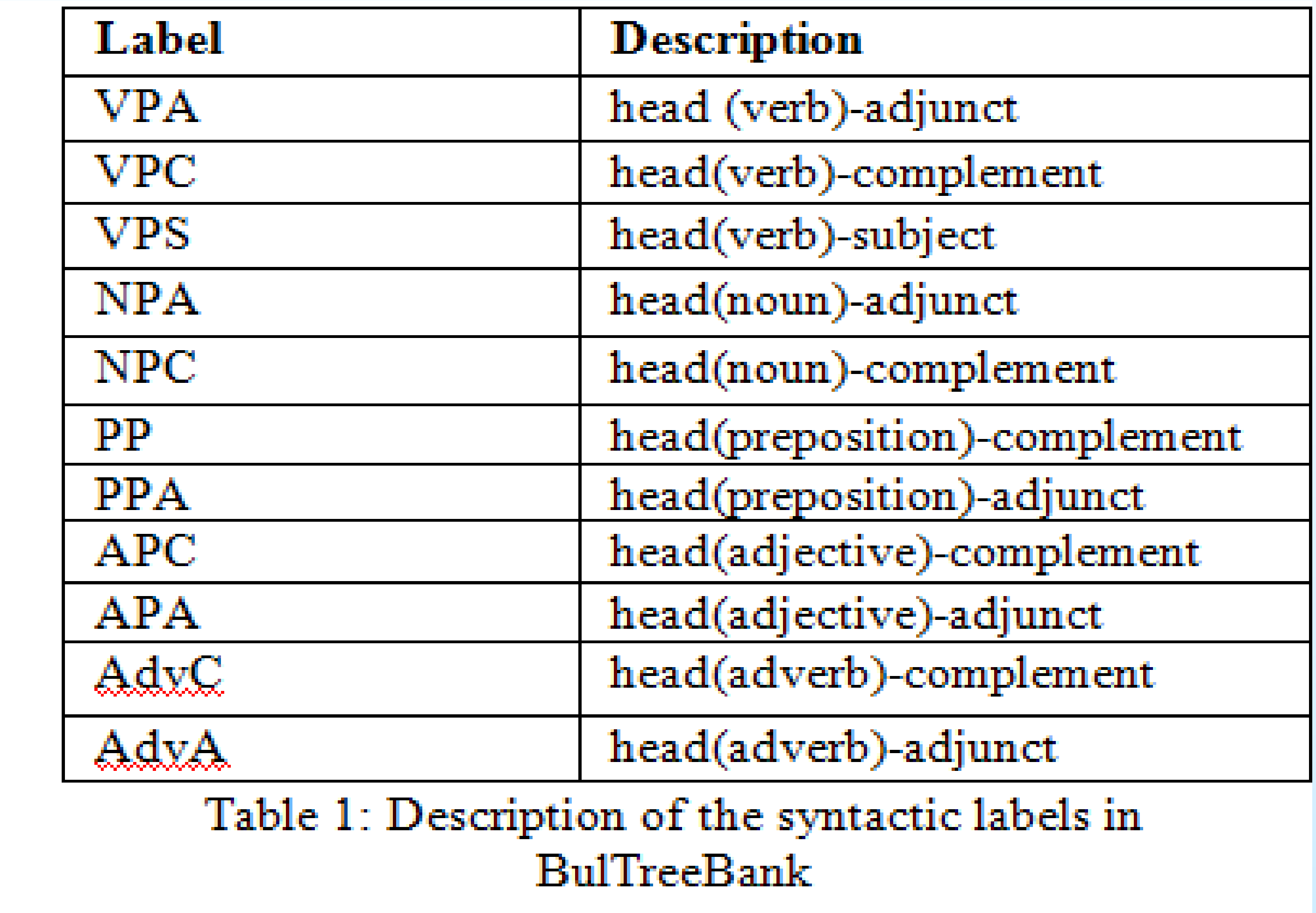
ORGANIZATION [appoint - lemma] PERSON



Some statistics

- The extracted annotated frames from BulTreeBank are **18081**
- Additional example material has been extracted also from the Bulgarian National Reference Corpus (when examples **< 5**)
- In BulTreeBank:
  - **920 verb lemmas** have occurred in only **once**;
  - **313 lemmas** have occurred **2 times**;
  - **200 lemmas – 3 times**;
  - **115 lemmas** – **4 times**;
  - **94 lemmas** – **5 times**

OntoValence Lexicon Architecture and Principles



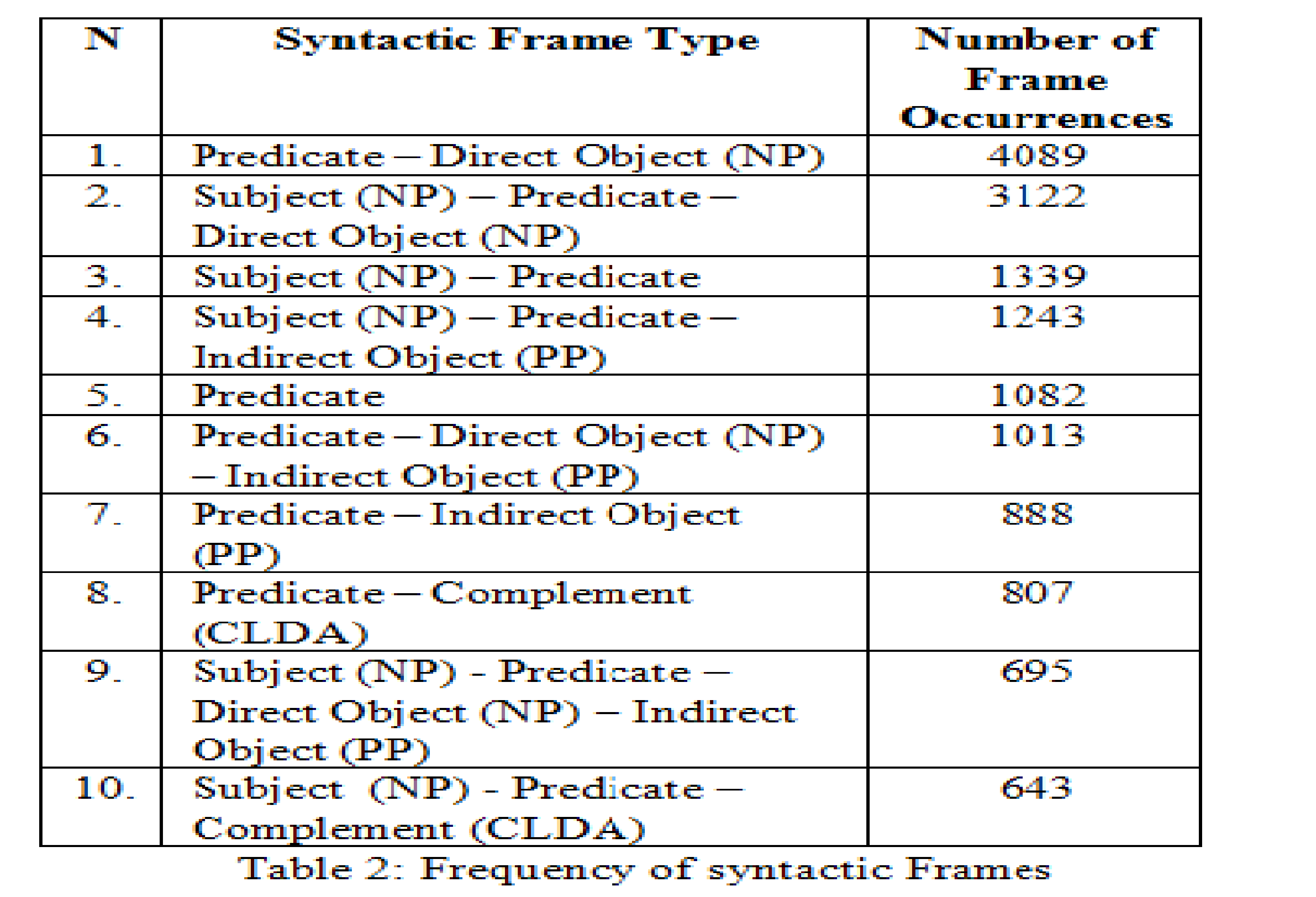
Label	Description
VPA	head (verb)-adjunct
VPC	head(verb)-complement
VPS	head(verb)-subject
NPA	head(noun)-adjunct
NPC	head(noun)-complement
PP	head(preposition)-complement
PPA	head(preposition)-adjunct
APC	head(adjective)-complement
APA	head(adjective)-adjunct
AdvC	head(adverb)-complement
AdvA	head(adverb)-adjunct

Table 1: Description of the syntactic labels in BulTreeBank

Specifics

- The valence frame is kept to the surface syntax
- Thus, the *pro-drops* of any kinds are also presented within the frames
- The frame considers the *clausal complements* as well
- We encode the verb usage in *active voice*
- The verbs in *perfective and imperfective aspect* are considered separate lemmas
- The frame includes only the *inner participants* (semantically obligatory for the event or situation, presented by the predicate, but might be unexpressed on the surface level)

Some Observations

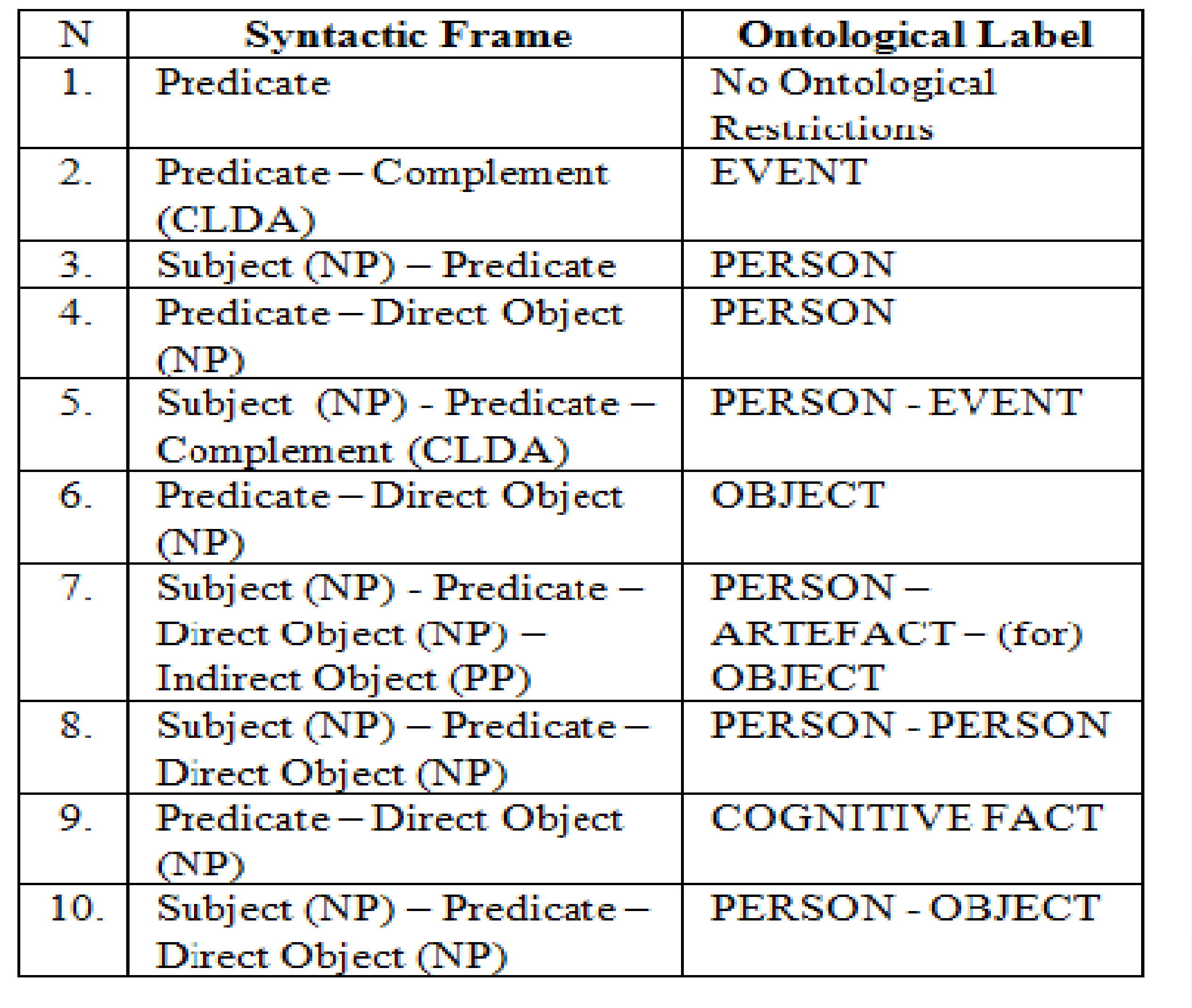


N	Syntactic Frame Type	Number of Frame Occurrences
1.	Predicate – Direct Object (NP)	4089
2.	Subject (NP) – Predicate – Direct Object (NP)	3122
3.	Subject (NP) – Predicate	1339
4.	Subject (NP) – Predicate – Indirect Object (PP)	1243
5.	Predicate	1082
6.	Predicate – Direct Object (NP) – Indirect Object (PP)	1013
7.	Predicate – Indirect Object (PP)	888
8.	Predicate – Complement (CLDA)	807
9.	Subject (NP) - Predicate – Direct Object (NP) – Indirect Object (PP)	695
10.	Subject (NP) - Predicate – Complement (CLDA)	643

Table 2: Frequency of syntactic Frames

Ontological Types:

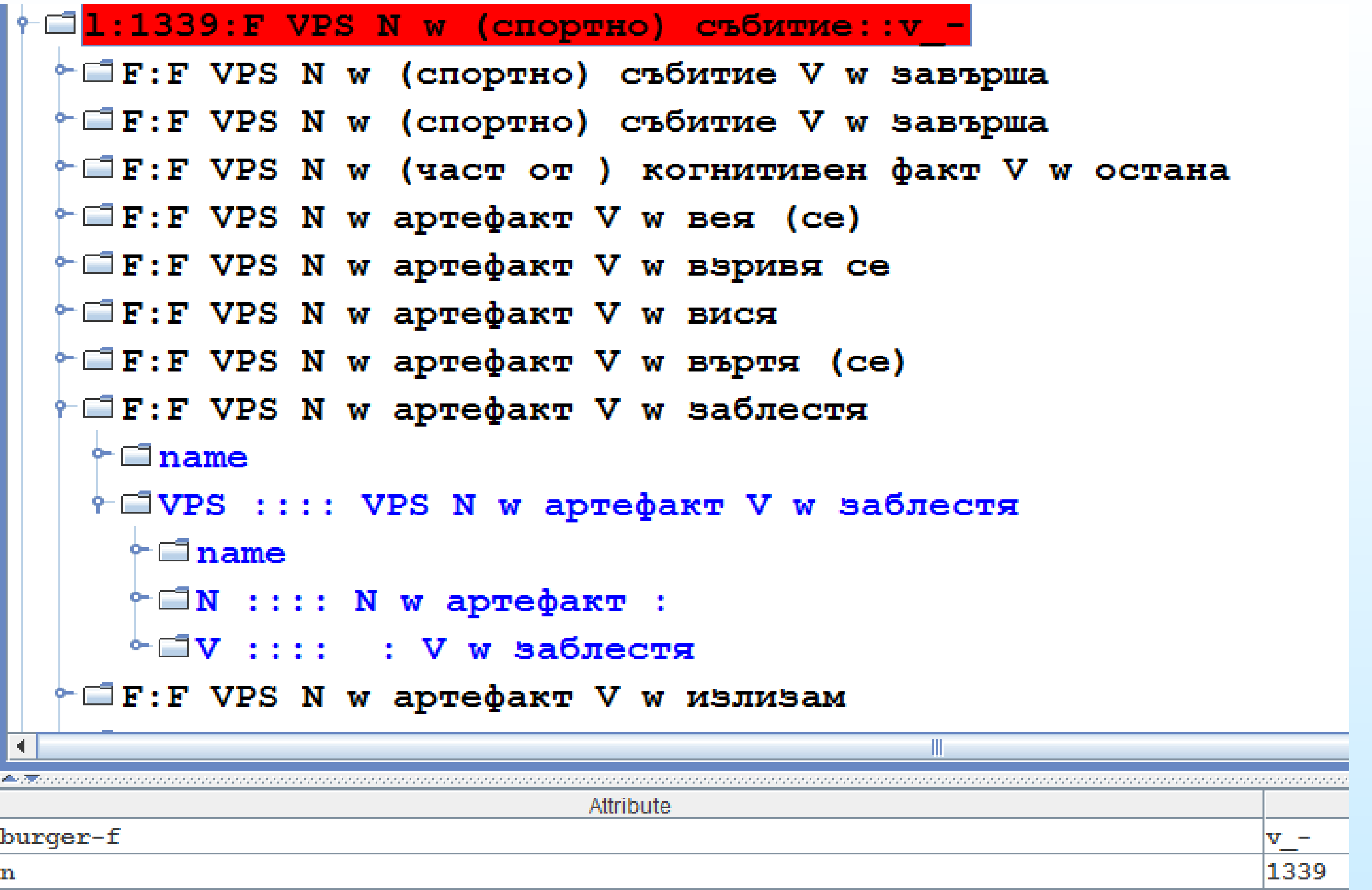
EVENT > PERSON > OBJECT > ARTEFACT > COGNITIVE FACT



N	Syntactic Frame	Ontological Label
1.	Predicate	No Ontological Restrictions
2.	Predicate – Complement (CLDA)	EVENT
3.	Subject (NP) – Predicate	PERSON
4.	Predicate – Direct Object (NP)	PERSON
5.	Subject (NP) - Predicate – Complement (CLDA)	PERSON - EVENT
6.	Predicate – Direct Object (NP)	OBJECT
7.	Subject (NP) - Predicate – Direct Object (NP) – Indirect Object (PP)	PERSON – ARTEFACT – (for) OBJECT
8.	Subject (NP) – Predicate – Direct Object (NP)	PERSON - PERSON
9.	Predicate – Direct Object (NP)	COGNITIVE FACT
10.	Subject (NP) – Predicate – Direct Object (NP)	PERSON - OBJECT

Incorporation of the Lexicon into BURGER

- The verbs have been sorted by frames
- The frames have been automatically transformed into partial syntactic types (v_-; v_pp; v_np…..)
- The information about the value of the aspect has been derived from the morphological dictionary for each verb lemma
- Tuning of the types
- Detecting of missing types



```
1:1339:F VPS N w (спортно) събитие::v_-
   F:F VPS N w (спортно) събитие V w завърша
   F:F VPS N w (спортно) събитие V w завърша
   F:F VPS N w (част от ) когнитивен факт V w остана
   F:F VPS N w артефакт V w вея (се)
   F:F VPS N w артефакт V w взривя се
   F:F VPS N w артефакт V w вися
   F:F VPS N w артефакт V w въртя (се)
   F:F VPS N w артефакт V w заблестя
      name
      VPS :::: VPS N w артефакт V w заблестя
         name
         N :::: N w артефакт :
         V ::::  : V w заблестя
   F:F VPS N w артефакт V w излизам
```

Attribute	
burger-f	v_-
n	1339

Processing Steps

- **Step 1:** Automatic assignment of the types to all the verbs that share a certain frame
- **Step 2:** Extending the types in BURGER
- **Step 3:** Automatic generation of the respective morphological paradigm

Statistics

- All valency types – 268 (including optional subject, impersonal verbs)
- All complement valency types – 41 (22 new)

Lemmas / Complement Frame

v_np	1307
v_-	874
v_pp	661
v_np-pp	546
v_che	158
v_da	143
v_advp	82
v_ques	50
v_pp-pp	48

Subject Realizations / Complement Frame

v_np	52
v_pp	40
v_np-pp	29
v_da	18
v_-	14
v_che	12
v_advp	11
v_pp-pp	9
v_cl	8
v_pp-da	6

Conclusions

- The OntoValency lexicon has been processed in full of its coverage – both on syntactic and ontological layers;
- More efforts are needed for testing the correct level of abstraction for the ontological labeling;
- The verb frames have been mapped to the BURGER types (they are the most frequent types)
- The missing types have been identified (mostly not so frequent)

Future work

- Near future
  - Adding of the new types to BURGER
  - Generation of verb paradigms for the mapped verbs
  - Lexicon Extension
- Farther future
  - Incorporation of more verbs by derivation types
  - Extension of types for other POS lemmas, generation of paradigms and lexicon expansion