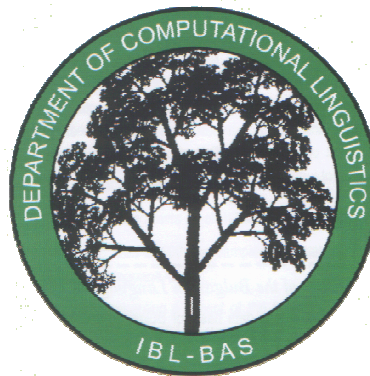


# Using for Analysis of Translational Asymmetries in Verb Argument Structure

Rositsa Dekova & Ivelina Stoyanova



14<sup>th</sup> национален форум за езиковедство  
inalco

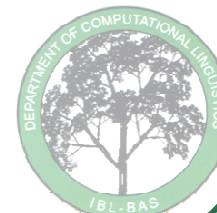


*15<sup>th</sup> NooJ Conference, 14-16 June 2012, Paris, France*

# Outline of the talk

---

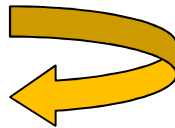
- Introduction and the Bulgarian-English Clause Aligned Corpus (BulEnAC)
- Problem statement
- Typology of translational asymmetries
- Translational asymmetries and clause alignment
- Analysis with NooJ
- Conclusions



# Introduction

---

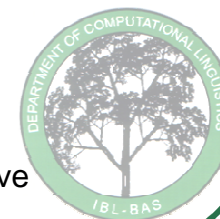
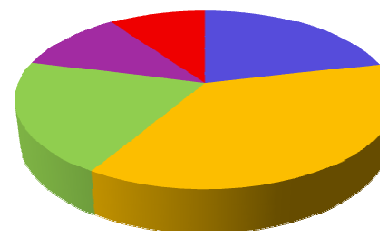
- For the purposes of training applications for machine translation, we need large parallel corpora
- These need to be aligned at different linguistic level
  - sentence level
  - phrase level
  - word level
- We introduce an intermediate level – a **clause level**, more suitable for studying verb argument structure
- Asymmetries pose problems for clause alignment



## Clause Aligned Bulgarian-English Corpus

---

- The clause-aligned corpus (BulEnAC) is part of the Bulgarian-English Parallel Corpus
- BulEnAC consists of 363 402 tokens altogether (174 790 for Bulgarian and 188 612 for English)
- Originally developed as a training corpus for automatic clause alignment.
- Texts are distributed over five thematic domains:
  - Fiction (21.4%),
  - News (37.1%),
  - Administrative (20.5%),
  - Science (11.2%)
  - Informal (subtitles) (9.8%).



# Clause Aligned Bulgarian-English Corpus

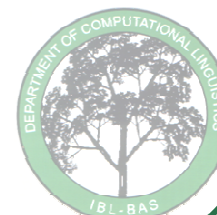
---

■ Both parts (Bulgarian and English) of BulEnAC are annotated with the following linguistic information:

- POS;
- Lemma;
- Sentence boundaries;
- Clause boundaries.

■ BulEnAC is aligned at:

- Sentence level;
- Clause level.



# Problem Statement

---

- Translational asymmetries are a problem for successful automatic alignment and extraction of translational equivalents
- We need to find ways to register translational asymmetries and treat them accordingly
- Problematic cases of translational asymmetries are demonstrated with examples from Bulgarian-English clause aligned corpus(BulEnAC)
- We use NooJ environment to analyse problematic cases



# Problem Statement

---

- Translational asymmetry: *a linguistic unit in the source language text which is not rendered in a regular way in the target language*
  
- There are different levels:
  - Lexical
  - Semantic
  - Syntactic
  - Stylistic
  - Pragmatic
  - Combined – at several levels simultaneously.



## General typology of translational asymmetries

---

- Word → MWE or MWE → Word (lexical)
- Changes in POS (lexical)
- Translation with a word with different sense (semantic)
- Asymmetry in the realisation of arguments and syntactic alternations (syntactic)
- Asymmetry due to extra-linguistic factors (pragmatic)





# Typology of translational asymmetries in verb argument structure

---

## ■ Lexical asymmetries:

- $V \rightarrow A / N / \dots$  change in POS of the head

## ■ Morphosyntactic asymmetries:

- Difference in Tenses

## ■ Asymmetries in syntax-semantics interface:

- $NP_1 V NP_2 \rightarrow NP_2 V NP_1$  different Arg realisation
- Active vs. passive



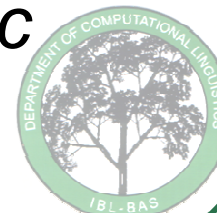
# Lexical Asymmetry – examples

---

■ Word → MWE or MWE → Word

*Other indications may appear on the bottle provided they do not **<MWE>give rise** </MWE> to confusion with the compulsory indications.*

*На бутилката могат да бъдат изписвани и други обозначения, при условие че не **<W>предизвикват** </W> объркване със задължителните обозначения.*



# Examples

---

## ■ Changes in POS

*In view of the above, the Commission considers that the scheme <VP> is [still] <A>**applicable**</A></VP> after Slovenia's accession to the European Union.*

*С оглед на гореизложеното Комисията смята, че схемата <VP>[продължава] да <V>**се използва**</V></VP> и след присъединяването на Словения към Европейския съюз.*



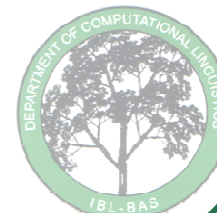
# Morphosyntactic asymmetries

---

## ■ Asymmetry in Tense

Моля ви, мистър, **<V tense=aorist>**  
**настъпихте** </V> ми кучето...

'Please, mister, you **<V tense=present cts.>**'re  
**stepping**</V> on my dog —'



## Asymmetries in syntax-semantics interface

---

### ■ Semantic asymmetries:

- $V_{ID_1} \rightarrow V_{ID_2}$  verbs from different synsets, corresponding to different sense
- Often: hypernym/hyponym;
- It can be used for revising synsets in Wordnet

### ■ Asymmetries in syntactic properties:

- $V_{INTR} \rightarrow V_{TRANS}$
- $V_{TRANS} \rightarrow V_{INTR}$
- Other



# Semantic asymmetry – examples

---

- Translation with a word with different sense, i.e. from different synset

**<V>Опичат</V>** *краката на човека, за да го накарат да пропее, а после се напъхват право в гостната на едно от другарчетата му.*

**<V>Burn</V>** *a guy 's feet to make him sing and then walk right into the parlor of one of his pals.*

... една молеща държава може да **<V>поиска</V>** поверителност на своята молба.

... a requesting State may **<V>seek</V>** the confidentiality of its request.



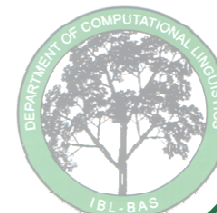
# Syntactic asymmetry – examples

---

- Asymmetry in the realisation of arguments and syntactic alternations

*French President Nicolas Sarkozy <V>mediated</V>  
<NP>a ceasefire agreement that ended the five-day  
conflict.</NP>*

*Френският президент Никола Саркози  
<V>посредничи</V> <PP>за споразумението за  
прекратяване на огъня, което сложи край на  
петдневния конфликт.</PP>*



## When is the clause not enough?

---

- We study the argument structure of verbs within the context of the clause
- Examples so far suggest it is possible and easier to observe verb argument structure and asymmetries are evident within the clauses
- Sometimes asymmetries affect clause structure and we need the context of the whole sentence
- These are problematic as they influence clause alignment





# Case studies

---

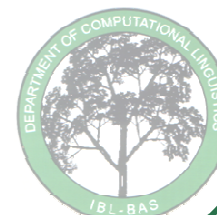
## ■ *Synset:*

- *пораждам, породя, предизвикам, предизвиквам, произведа, произвеждам – transitive*

- *bring about, give rise, produce – intransitive*

## ■ 25% of the cases *give rise* is translated by 'vodya / doveda / dovezhdam do'

## ■ These are not included in the synset although they match the argument structure of 'give rise (to)' which suggests that this is a gap in the wordnet



# Case studies

---

- To demonstrate asymmetries we explored the usage of Bulgarian verb: *настоявам* (*nastoyavam*)
- It is a member of 3 synsets:
  - *insist, take a firm stand* — *настоявам, настоятелен съм, твърд съм, упорствам*
  - *claim, demand* — *изисквам, настоя, настоявам, поисквам*
  - *importune, insist* — *вадя душата на, моля настойчиво, настоявам*
- The English verb *urge* is contained in 2 synsets:
  - *карам, убедя, убеждавам, увещавам, увеща* — *exhort, press, urge, urge on*
  - *карам* — *barrack, cheer, exhort, inspire, pep up, root on, urge, urge on*



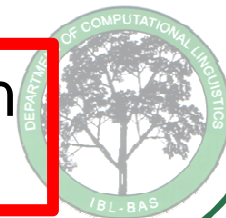
## Results of the study

---

- Synsets containing *nastoyavam* and *urge* do not intersect
- However, they are very often used as translational equivalents

In BulEnAC there are 70 occurrences of *nastoyavam* and 19 of them were translated with *urge* – 27% of the cases

There is clearly a translational equivalence between the two forms which Wordnet does not show.



## Analysis of the results

---

- Different synsets are expected as these have different subcategorisation frames

BG: ***nastoyavam*** – intransitive (EN *insist*)

\_ <SubCl>**s.o.** to do s.th.<SubCl>

<PP>for s.th.<PP>

EN: ***urge*** – transitive (BG *karam*)

\_ **s.o.** <SubCl>to do s.th.<SubCl>

(additional senses of *urge* are not considered here)



# Analysis of the results

---

■ larger context than the clause is needed to identify the asymmetry

<MainCl>Групи родители **настояха**</MainCl>

<SubCl>**румънските власти** да вдигнат  
забраната</SubCl>.

<p>

<MainCl>Parents' groups have **urged**

**Romanian authorities**</MainCl> <SubCl>to lift the  
ban</SubCl>.



## Level of analysis with NooJ

---

- In many cases the clause provides sufficient context to examine argument structure and asymmetries
- Sometimes arguments are distributed differently across clauses and then a larger context is necessary
- Therefore we prefer to operate with the entire sentences



## Analysis with NooJ

---

- NooJ does not offer functionalities for processing parallel texts / multilingual data
- We declare resources for both English and Bulgarian as monolingual (Bulgarian) and process the corpus
- In NooJ the corpus is loaded as a set of bi-sentences separated by  $\langle p \rangle$



# Analysis with NooJ

NooJ-sents-nastoya.not [Modified]

8 / 547 TUs

Characters  
Tokens  
Digrams  
Unknowns  
Ambiguities

Language is "Bulgarian (Bulgaria)(bg)".  
Text Delimiter is: \n (NEWLINE)  
Text contains 547 Text Units (TUs).  
14652 tokens including:  
сана мисл...

Show Text Annotation Structure

<s>  
3. Научният съвет се събира по молба на секретариата, както се изисква от Конференцията на страните. <p>  
===== 3. The Scientific Council shall meet at the request of the Secretariat as required by the Conference of the Parties.  
</s>  
<s>  
===== 6) ===== специалните изисквания за отпадъчните води, ===== които изискват отделно обработване; <p>  
===== (b) special requirements for effluents ===== necessitating separate treatment;  
</s>

|                        |           |           |               |          |         |     |
|------------------------|-----------|-----------|---------------|----------|---------|-----|
| 63                     | 71        | 88        | 91            | 102      | 112     | 116 |
| изисквам,V+R3s+PE+IM+T | от,PREP+Q | на,PREP+Q | страна,N+pd+F | п,N+Nb=s | the,DET | sc: |
| изисквам,V+E2s+PE+IM+T |           |           |               |          |         |     |
| изисквам,V+E3s+PE+IM+T |           |           |               |          |         |     |



## Analysis with NooJ

---

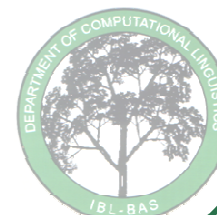
### ■ We apply

- Bulgarian Grammar Dictionary (about 4,000 lemmas and 58,000 word forms).
- Dictionaries and grammars for English (distributed with NooJ)

### ■ Small dictionaries of synonyms in Bulgarian and English – entries from the same synsets marked with identical ID's

настоявам, V+TR+SPEC+SYN=ID132

insist, V+TR+SPEC+SYN=ID132



## Analysis with NooJ

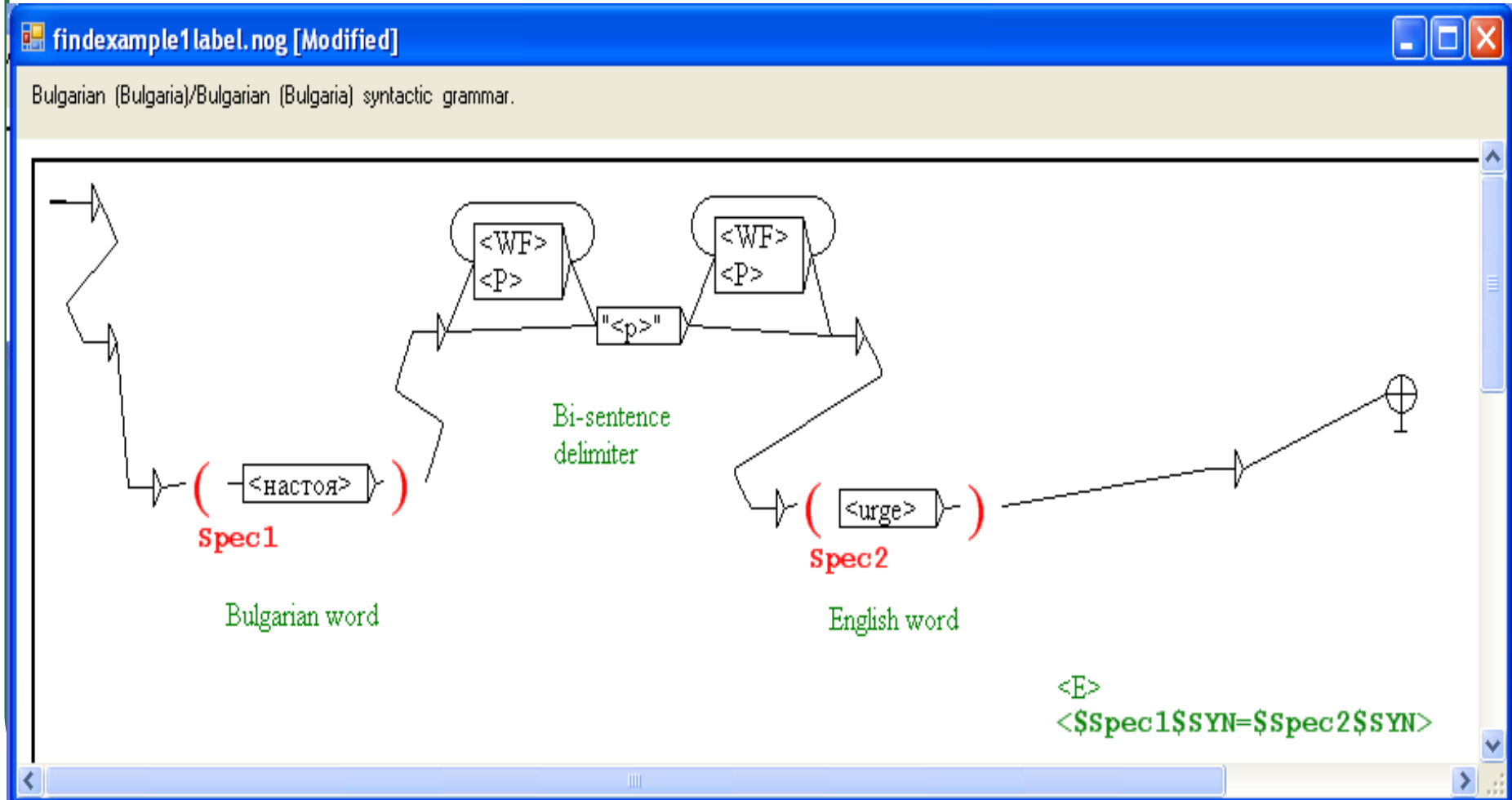
---

- We extract examples either using regular expressions or grammars

$\langle \text{настоя} \rangle (\langle WF \rangle + \langle P \rangle)^* \langle p \rangle (\langle WF \rangle + \langle P \rangle)^* \langle \text{urge} \rangle$



# Analysis with NooJ



# Analysis with NooJ

The screenshot shows the 'Concordance for Text' window in NooJ. The title bar reads 'Concordance for Text NooJ-sents-nastoya.not [Modified]'. The interface includes a 'Reset' button, a 'Display' field set to '5', and radio buttons for 'characters' and 'word forms'. The search parameters are 'before, and 5 after. Display:  Matches  Outputs'. The main area is a table with three columns: 'Before', 'Seq.', and 'After'. The search term 'настояха' is circled in red in the 'Seq.' column, and its English translation 'have urged' is circled in red in the 'After' column. The table contains several rows of text, including mentions of 'Romanian authorities', 'Belgrade', 'NATO', 'Del Ponte', 'Undersecretary', 'EU', 'Croatia', 'BiH', and 'European Commission'. The bottom of the window shows a 'Query' field and a page number '19/19'.

| Before    | Seq.  | After   |
|-----------|---|---|
| родители  | настояха, румънските власти да вдигнат забраната.                                 | Parents' group have urged Romanian authorities to lift th           |
| итгомври  | настоя Белград да сложи край на нарушенията.                                      | ==== In an accompanying letter, ... Belgrade to put an end          |
| ел Понте  | настоя също НАТО да сформира специална военна част.                               | Del Ponte also urged NATO to form a special                         |
| секретар  | настоя ЕС да насрочи дата за започване на преговори.                              | Undersecretary urged the EU ==== to set a                           |
| Те също   | настояха, Белград да изпълни международните си задължения и да подобри сътру...   | Belgrade to fulfil its internatic                                   |
| <s> Той   | настоя да бъдат взети повече мерки за подпомагане завръщането на принудителн...   | for more measures ==== to   |
| арламент  | настоява Хърватска да сътрудничи напълно на трибунала за военни престъплени...    | Croatia to Fully Co-operate   |
| > НАТО    | настоява БиХ да обедини командването на армията                                   | ==== NATO Urges BiH ==== to Consolidate A                           |
| ==== като | настоява да създадат съвместен команден и контролен център за въоръжените си...   | them to set up a  |
| <s> Той   | настоя съюзът да посочи ясно, че е недоволен.                                     | He urged the Union to make ==== cle                                 |
| ия, ====  | настоява страната да напредва с реформите   | ==== European Commission Praises ... Country to Proceed With Re     |
| равила и  | настоя страната да продължи с реформите.  | ==== A delegation representing the Eu... it to proceed with reforms |
| бомбгенс  | настоя в четвъртък властите на БиХ да ускорят заделянето на средства в бюджета... | BiH authorities on Thursday   |
| н отново  | настоява за създаване на общо Министерство на отбраната на БиХ                    | ==== NAT... Creation of Joint Defence M                             |

# Analysis with NooJ

NooJ-clauses-nastoya.not [Modified]

26 / 547 TUs

Characters  
Tokens  
Digrams  
Unknowns  
Ambiguities

Language is "Bulgarian (Bulgaria)(bg)".  
Text Delimiter is: \n (NEWLINE)  
Text contains 547 Text Units (TUs).  
6510 tokens including:  
арно мод форма

Show Text Annotation Structure

<s>  
замолената държава може да поиска от молещата държава <p>===== the requested State may require  
</s>

<s>  
ако това бъде поискано от молещата държава. <p>if such confidentiality is requested by the requesting State.  
</s>

|                    |                       |           |                     |      |
|--------------------|-----------------------|-----------|---------------------|------|
| 4                  | 14                    | 23        | 26                  | 35   |
| този,PRO+sn+DEM+PT | поискам,V+Qsn0+PE+P+T | от,PREP+0 | моля,V+Ysfd+PE+IM+T | дър> |



# General conclusions

---

- These asymmetries are not single examples but a widely present relation between languages
- They need to be treated properly in terms of:
  - lexical description (dictionaries)
  - semantic relations (Wordnet)
  - argument structure (Framenet)



## Conclusions and further research

---

- We believe clause alignment will facilitate phrase and word alignment and will improve Machine translation (e.g. Moses)
- Translational asymmetries are a widely spread phenomena; they are normal across languages and contribute to richness and diversity.
- Asymmetries pose problems before clause alignment and advanced linguistic models are needed to account for them.
- NooJ can be applied for analysis of translational asymmetries.
- Some additional features may be introduced into NooJ for processing of parallel texts.



---

This research was conducted with financial support granted by the “*Human Resources Development*” Operational Programme, co-financed by the European Social Fund of the European Union.







DEPARTMENT OF  
COMPUTATIONAL  
LINGUISTICS



# Thank you for your attention!

*{rosdek, iva}@dcl.bas.bg*

*<http://dcl.bas.bg/>*

