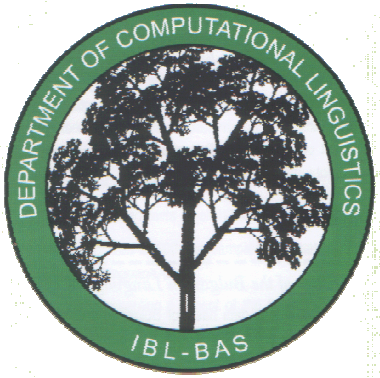
Using NooJ for Analysis of Translational Asymmetries in Verb Argument Structure

Rositsa Dekova & Ivelina Stoyanova



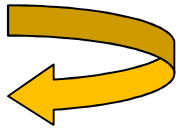
15th NooJ Conference, *14-16 June 2012, Paris, France*

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

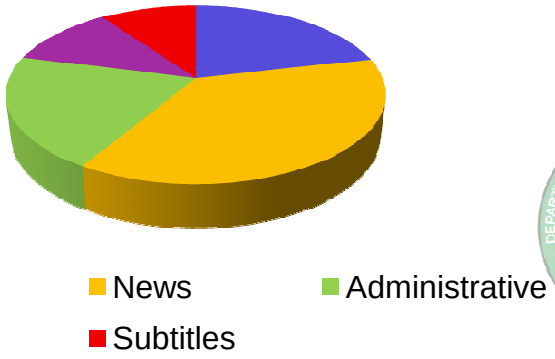


- We introduce an intermediate level – a **clause level**, more suitable for studying verb argument structure
- Asymmetries pose problems for clause alignment

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%).

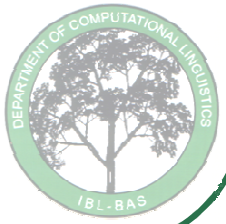


Clause Aligned Bulgarian-English Corpus

- 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.

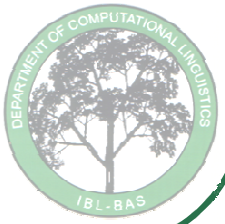
Problem Statement

- 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.

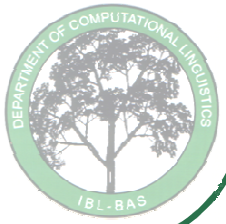


General typology of translational asymmetries

- 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 → A / N / … change in POS of the head
- Morphosyntactic asymmetries:
  - Difference in Tenses
- Asymmetries in syntax-semantics interface:
  - NP_1 V NP_2 → NP_2 V NP_1 different Arg realisation
  - Active vs. passive

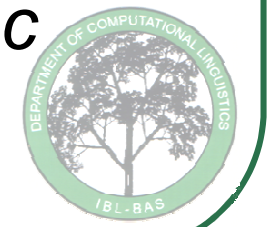


Lexical Asymmetry – examples

- 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>*

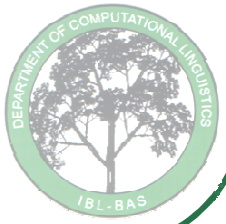
*Френският президент Никола Саркози <V>**посредничи**</V> <PP>за споразумението за прекратяване на огъня, което сложи край на петдневния конфликт.</PP>*

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 matche 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, увещая — 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>.

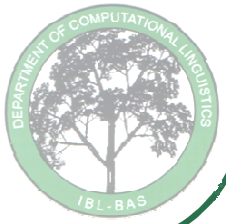
<p>

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

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

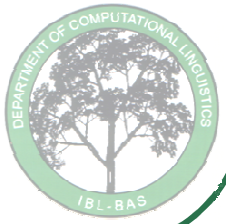
Level of analysis with NooJ

- 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

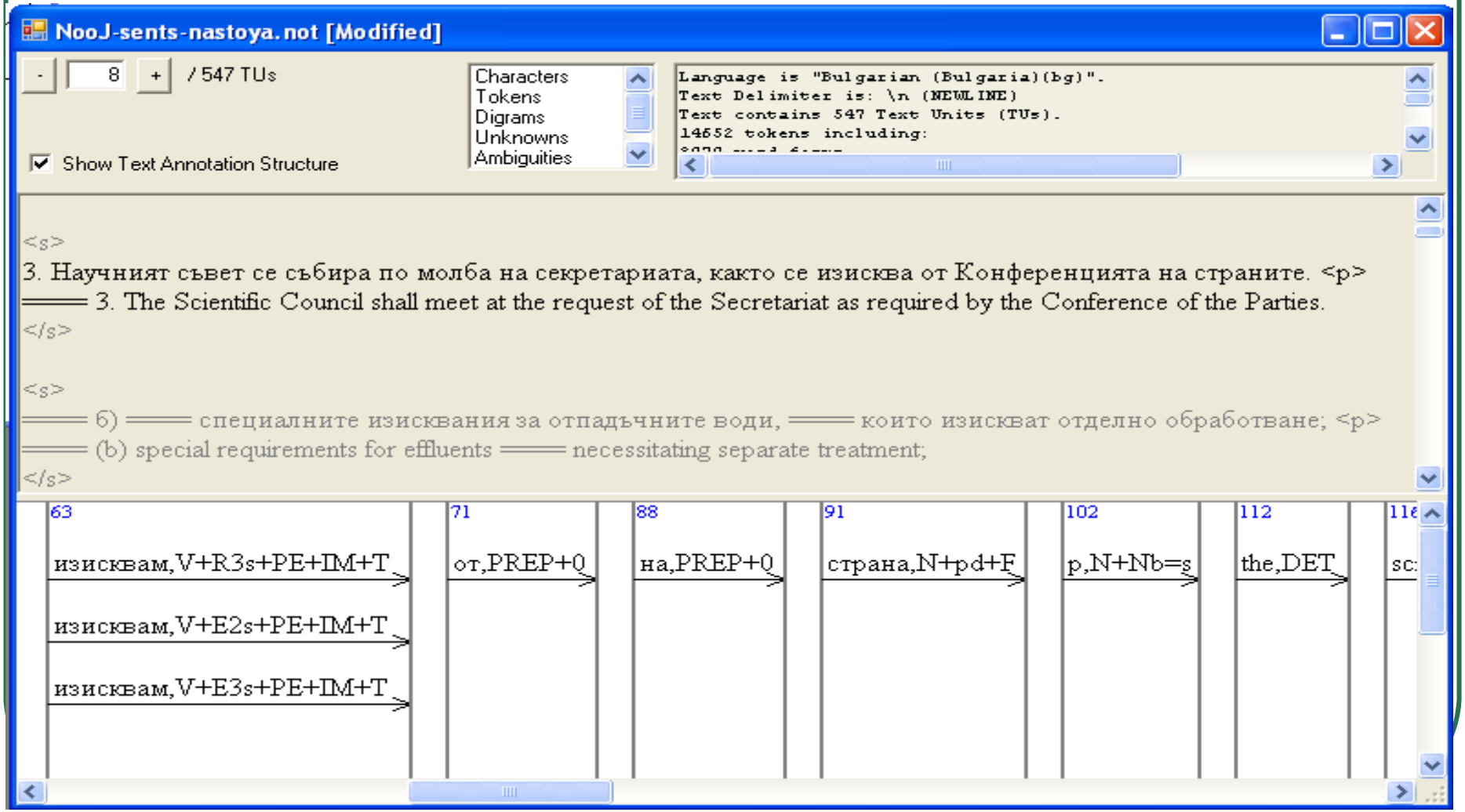


Analysis with NooJ

- 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 <*p*>



Analysis with NooJ

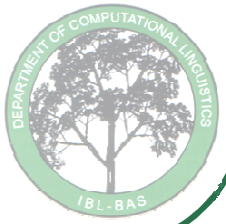


Analysis with NooJ

- 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

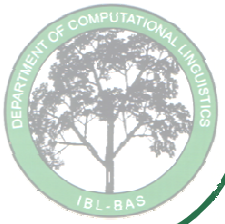
  insist, V+TR+SPEC+SYN=ID132



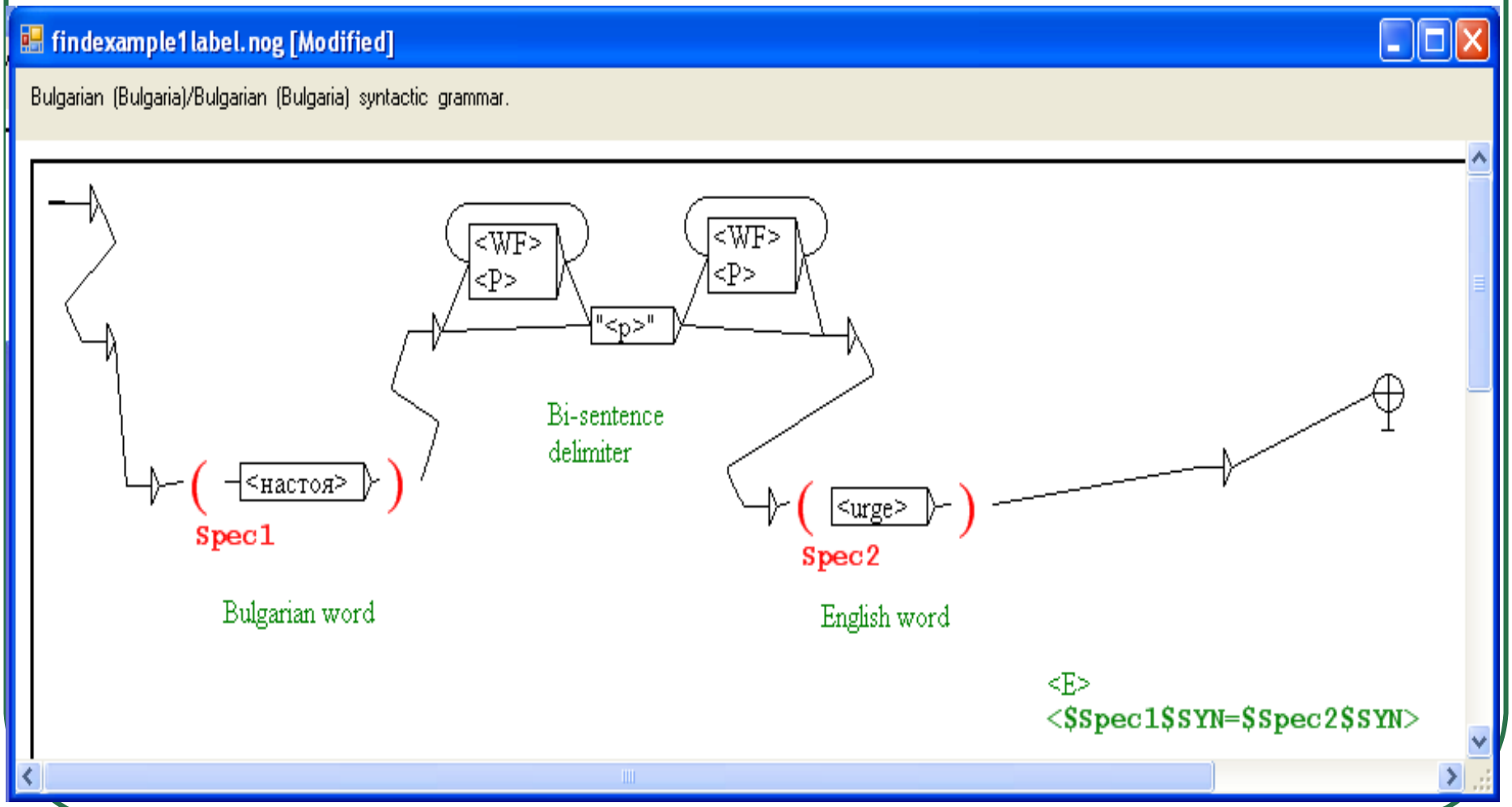
Analysis with NooJ

- We extract examples either using regular expressions or grammars

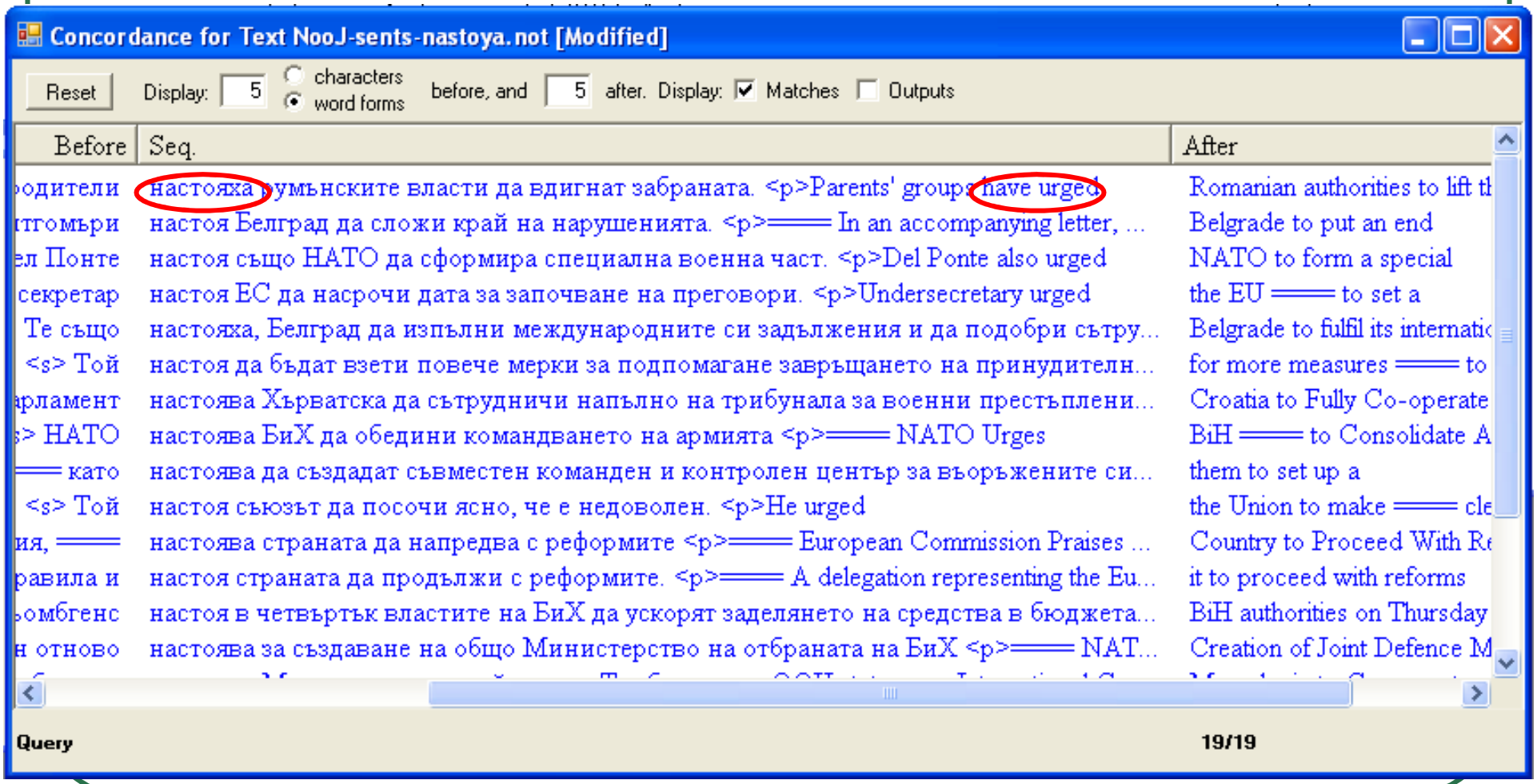
<настоя> (<WF>+<P>) "<p>" (<WF>+<P>)* <urge>*



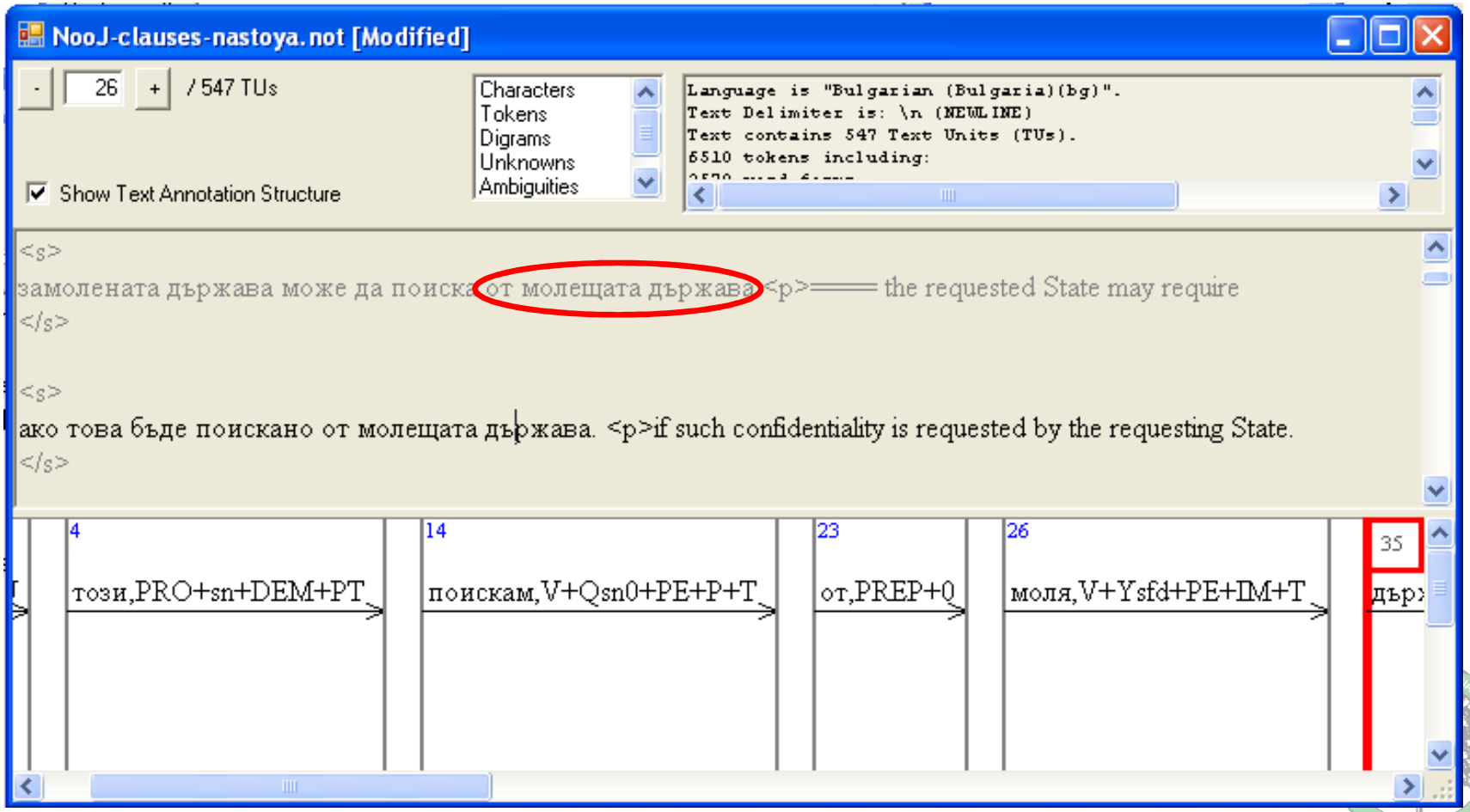
Analysis with NooJ



Analysis with NooJ



Analysis with NooJ

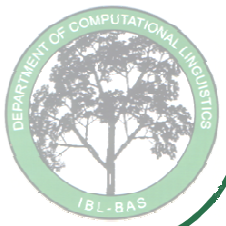


General conclusions

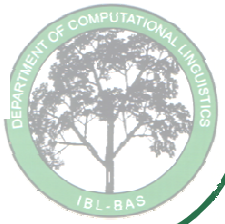
- These asymmetries are not single examples but a widely present relation between languages
- They need to be treated properly in terms of:
  - lexical description (dictionaries)
  - semantic relations (Wordnet)
  - argument structure (Framenet)

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.



Thank you for your attention!

{rosdek, iva}@dcl.bas.bg

http://dcl.bas.bg/