WG2

2^{nd} General Meeting
University of Naples “L’Orientale”
Naples, Italy, 8-9 February 2024

https://unidive.lisn.upsaclay.fr/

An English-Bulgarian Comparable Corpus Annotated with FrameNet Valence Patterns

Ivelina Stoyanova | iva@dcl.bas.bg | https://dcl.bas.bg/
Department of Computational Linguistics
Institute for Bulgarian Language, Bulgarian Academy of Sciences

UniDive

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

Objectives

Practical:

- To build a bilingual corpus that demonstrates the syntactic realisation of the conceptual description of verbs in English and Bulgarian.
- To combine information from different resources for the extensive semantic and syntactic description of verbs.

Theoretical:

- To study universality and the possible cross-language linking and transfer of information (from English to Bulgarian).

Extraction of examples

For English:

➔ Annotated examples extracted from FN (annotated verb and frame elements).
➔ Supplemented with examples from SemCor (containing verbs of particular WN synsets).

For Bulgarian:

➔ Extracted examples from BulSemCor.
➔ Using the Bulgarian National Corpus.

Parallel:

➔ Extracted parallel example sentences from aligned Bulgarian-English corpus.

Annotation of examples

In each sentence the following are annotated:

- the verb – tagged, lemmatised, and linked to a WN synset (unique sense);
- the phrases corresponding to core frame elements (e.g., AGENT, SPEAKER, THEME, etc.)
- the syntactic function of the identified phrases (NP.Ext, NP.Obj, PP, AdvP, Clause, etc.).

Applications and future work

The conceptual description encoded in FN frames is largely language-independent.

Resources

- **Lexical-semantic resources:** WordNet, Bulgarian WordNet and FrameNet, linked by assigning FN frames onto WN synsets. The configurations of the (core) frame elements and their syntactic realisations for lexical units are extracted from FN and linked to WN synsets.
- **Corpora:** SemCor and BulSemCor (semantically annotated with WN glosses), parallel Bulgarian-English corpus, Bulgarian National Corpus.

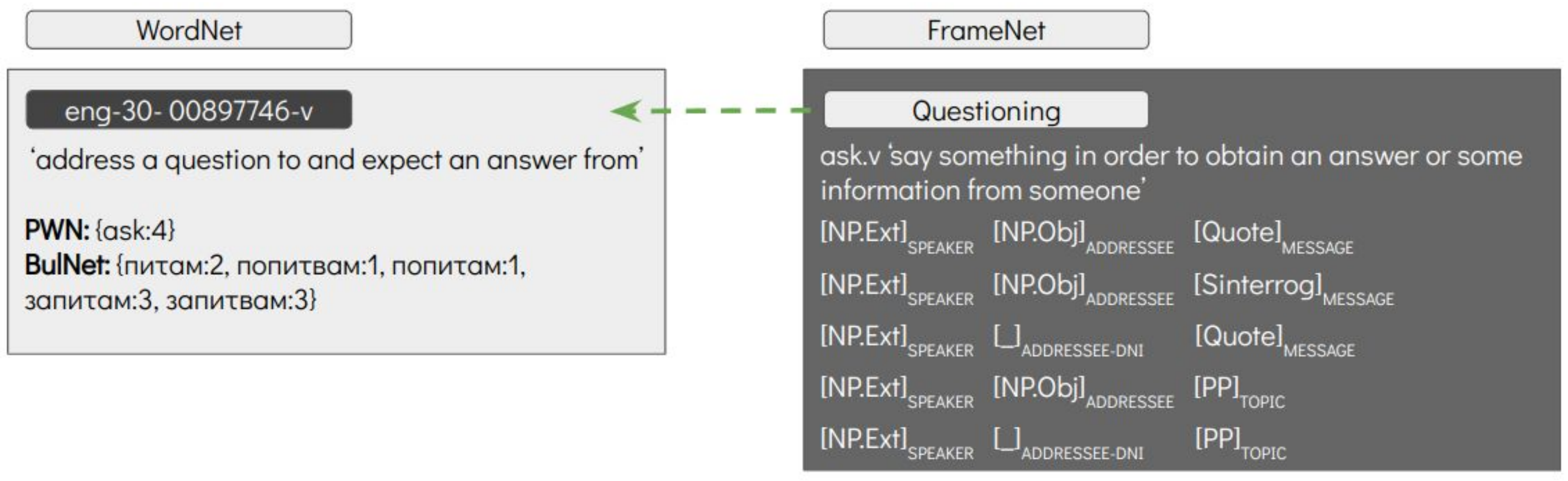
  Observations are made on 3 semantic classes: verbs of change, verbs of motion and verbs of communication.

For English:

➔ 211 verbs (lexical units in FN) aligned to 135 WN synsets.
➔ 13,295 annotated examples.
➔ 3,577 different valence patterns.

For Bulgarian (work in progress):

➔ 146 verbs aligned to 125 WN synsets.
➔ 2,050 annotated examples.
➔ 272 different valence patterns.



eng-30- 00897746-v
‘address a question to and expect an answer from’

PWN: {ask:4}
BulNet: {питам:2, попитвам:1, попитам:1, запитам:3, запитвам:3}

Questioning
ask.v ‘say something in order to obtain an answer or some information from someone’

$[\text{NP.Ext}]_{\text{SPEAKER}}$ $[\text{NP.Obj}]_{\text{ADDRESSEE}}$ $[\text{Quote}]_{\text{MESSAGE}}$
$[\text{NP.Ext}]_{\text{SPEAKER}}$ $[\text{NP.Obj}]_{\text{ADDRESSEE}}$ $[\text{Sinterrog}]_{\text{MESSAGE}}$
$[\text{NP.Ext}]_{\text{SPEAKER}}$ $[\sqcup]_{\text{ADDRESSEE-DNI}}$ $[\text{Quote}]_{\text{MESSAGE}}$
$[\text{NP.Ext}]_{\text{SPEAKER}}$ $[\text{NP.Obj}]_{\text{ADDRESSEE}}$ $[\text{PP}]_{\text{TOPIC}}$
$[\text{NP.Ext}]_{\text{SPEAKER}}$ $[\sqcup]_{\text{ADDRESSEE-DNI}}$ $[\text{PP}]_{\text{TOPIC}}$

[Престъпникът]$_{\text{SPEAKER:NP.Ext}}$	**попита**	[полицая]$_{\text{ADDRESSEE:NP.Obj}}$	[дали може да си купи цигари]$_{\text{MESSAGE.Clause}}$.
[Criminal-DEF]$_{\text{SPEAKER:NP.Ext}}$	**asked**	[policeman-DEF]$_{\text{ADDRESSEE:NP.Obj}}$	[whether can to himself buy cigarettes]$_{\text{MESSAGE.Clause}}$.
[The criminal]$_{\text{SPEAKER:NP.Ext}}$	**asked**	[the policeman]$_{\text{ADDRESSEE:NP.Obj}}$	[whether he could buy some cigarettes.]$_{\text{MESSAGE.Clause}}$.

Information can be transferred:

- between languages: from English to Bulgarian (and other less resourced languages);
- between resources: for mutually enriching both FrameNet and WordNet.

Acknowledgements:

This research is carried out as part of the project *Enriching the Semantic Network WordNet with Conceptual Frames* funded by the Bulgarian National Science Fund, Grant Agreement No. KP-06-H50/1 from 2020.