

APPLICATIONS OF NOOJ IN WSD AND MT

Svetlozara Leseva

DCL - IBL

zara@ibl.bas.bg

OVERVIEW

- WSD- identifying a word's sense/definition/meaning in a given context
- Types of approaches to WSD:
 - **Information based/Knowledge driven WSD:** external knowledge source (dictionary, thesaurus, etc.)
 - **Corpus based/Data driven WSD:** context of the word

WORD SENSE DISAMBIGUATION

- Corpus-based approach
- Types of techniques:
 - Supervised WSD – makes use of the language data derived from hand annotated corpora to develop probability models
 - Unsupervised WSD – calculates probability of co-occurrences of word sense without using the context
 - Hybrid techniques - Bootstrapping

WORD SENSE DISAMBIGUATION

- Hybrid approach

Annotation of a corpus using sense inventory from a knowledge base – dictionaries, thesauri, etc.

Lexical-semantic resources such as wordnet are largely employed in word-sense disambiguation

The Bulgarian database BulNet

SENSE-TAGGED CORPUS

- Assignment of certain linguistic + metalinguistic information in an established format. Tokens are

- sense disambiguated:

A word is associated with a number of senses: $w_1 \{s_1, s_2, \dots, s_n\}$ from which the most appropriate one is selected in the context: $w_1 \{s_k\}$;

- identified and supplied with specific annotation (named entities, syntactic structures, etc.);

ANNOTATION

- Annotated tokens look like:

лекарят{лекар **ENG20-09380179-n**
1126537557 8 1}

етичния{етичен **0 2000000000 7**
0} кодекс{кодекс **BUL-**
159035502 1141032506 6 1}



SYNSET ID
in BULNET

ANNOTATION

- Tokens and units with respect to annotation:
 - single words;
 - multiword expressions: “idiosyncratic interpretations that cross word boundaries (or spaces)” (Sag et al. 2002);
 - Syntactic structures- identifiable and reproducible syntactic patterns.

TARGET ENTITIES FOR AUTOMATED PROCEDURES

- Simple words and MWE
 - subject to WS-tagging:
entity names, temporal expressions,
number expressions;
predictably derived compound-words;
 - subject to identification and
metalinguistic annotation:
entity names, temporal expressions,
number expressions
- Treating words differently is dictated by
their linguistic and extralinguistic
significance

TARGET ENTITIES FOR AUTOMATED PROCEDURES

Compound-word derivation:

Cardinal and ordinal numerals

*I told him **one thousand forty-four** times to stop.*

- Adjective and noun derivation
 - Numeral + adjective
 - Numeral + noun

***n-year-old** boy, **n-storey** house, **[A-z]-shaped** box;*

***25-five-year-old** woman, **2-storey** house, **U-turn**;*

TARGET ENTITIES FOR AUTOMATED PROCEDURES

- Productive derivation roots/affixes and models:

Contracted form of an adjective + another category

{A+CONTR}+ N/A/V

***Euro**integration, **Euro**parliament;*

Compounds:

***top** cop, **top** reporter, **top** news;*

- Derivation from proper names:

Suffixes: **-ov**, **-ev**, **-ski**, **-ianski**, etc.
(**`s/-ian/-ean**);

TARGET ENTITIES FOR AUTOMATED PROCEDURES

- Hyphenated compound words
kulturno-istoricheski = *cultural and historical*;
- Syntactic structures;
- Named entities – proper names, dates, hours, per cents/portions:
22 May 2006/**<DATE_MONTH_YEAR>**
24 %/**<NUM_PERCENT>**

ENHANCED SEMANTIC ANNOTATION

- Introducing semantic relations - synonyms, hyperonyms into the dictionary (Silberztein, Koeva 2005):

Assisting the manual annotation of a corpus by introducing:

- Semantic relations;
- Explanatory definitions;
- Translation equivalents.

DICTIONARY

Dictionary entries:

LEMMA:

SUPER-LEMMA: translation equivalent derived from PWN (or any other target language);

CATEGORY: Categorical information of the lemma;

INFL: Inflectional description of the lemma.

двайсет, twenty, NU+CA+M+FLX=NU_CA_M_2

MORPHOLOGICAL GRAMMAR

The screenshot displays a software interface for morphological analysis. The main window, titled "Untitled (Modified)", shows the text "Актуализирах плана за евроинтеграцията" (I updated the plan for European integration). Below the text, a morphological analysis window shows the following information:

Bulgarian (Bulgaria) morphology

HM+T+EЗр → /LMA.N+M+sl → за.PREP → евроинтеграцията.N+R_F-eurointegration →
DEF:integration within Europe or the EU

The analysis window also displays a morphological tree diagram. The root node is a circle with a cross, which branches into a triangle node labeled "Pref" and a circle node labeled "N". The "Pref" node branches into a box containing "eво" and a circle containing "<L>". The "N" node branches into a circle containing "<S>" and a circle containing "R". Below the diagram, the following text is shown:

```
{ $0, # $1 F = euro $1 L  
DEF: $1 L within Europe or the EU }
```

Abstract definition of nouns beginning with the contracted form Euro (European)

FROM WSD TO MT

Dictionaries and grammars may be used to associate a word with its translation variant:

The **super-lemma**:

- Orthographic variants;
- Semantic relations;
- Translation variants.

TOWARDS MT

Two levels:

- Assignment of translation equivalents (TE) in the annotated corpus:

Lemmas are already associated with the word form, and the wordnet key is assigned.

Hence, the combination of the lemma and the corresponding ID may serve as dictionary entry:

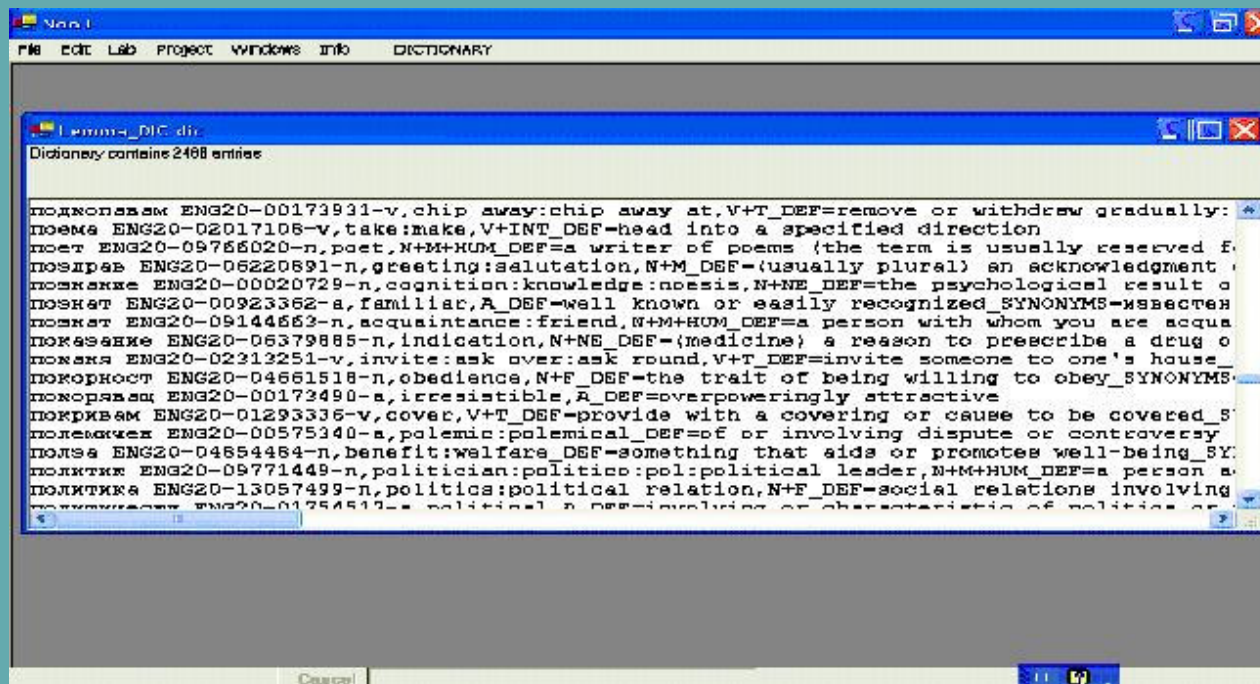
пасажер ENG20-09729204-n

DICTIONARY OF LEMMAS

The dictionary of lemmas includes the following parts:

- Ordered pair LEMMA ID;
- SUPER-LEMMA: synonym sets derived from PWN (or any other target language);
- CATEGORY: Categorical information for the lemma;
- DEF: the corresponding definition derived from wordnet;
- SYNONYMS: the members of the synonym set to which the lemma pertains.

DICTIONARY OF LEMMAS



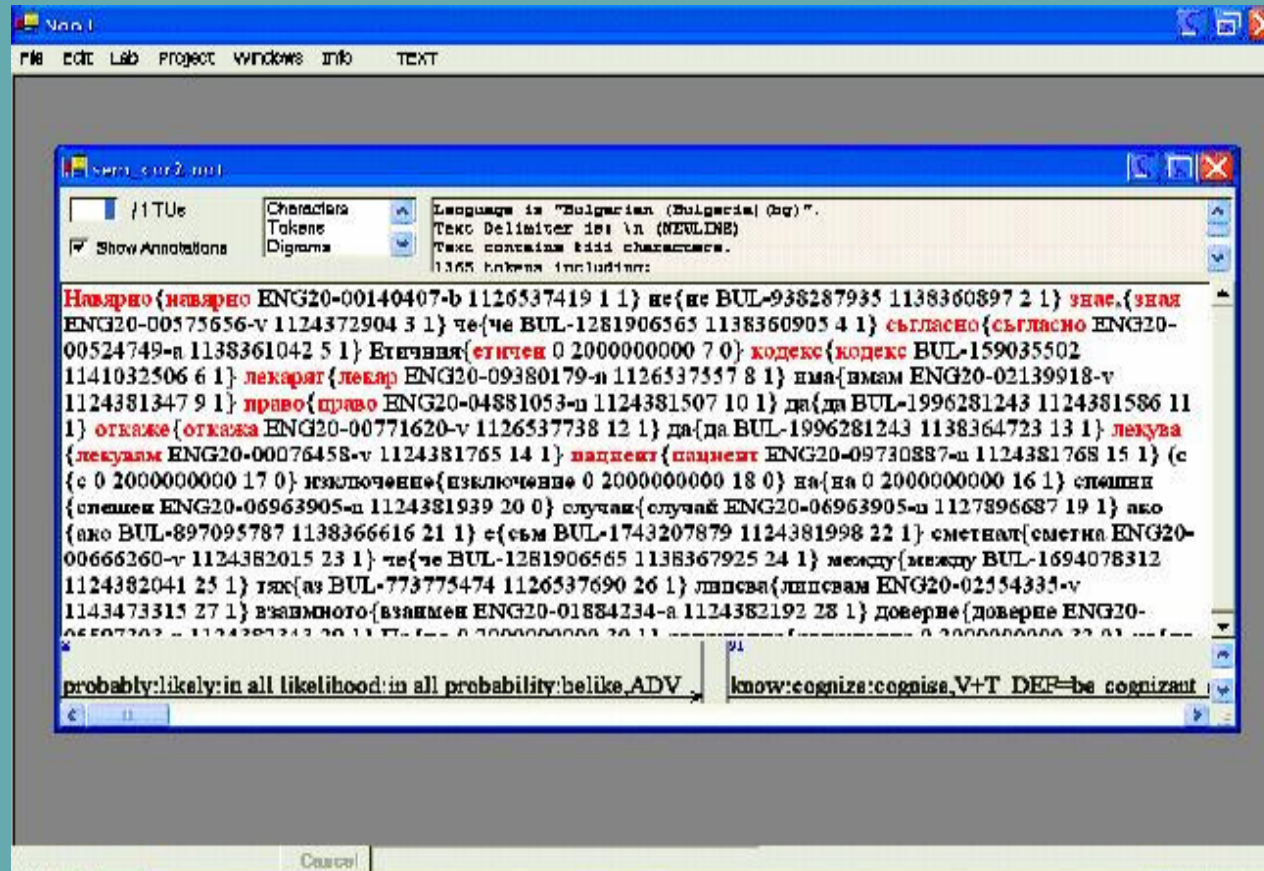
*пасажер ENG20-09729204-n, passenger:rider,
N+M+HUM_DEF=a traveler riding in a vehicle (a boat or bus or car or
plane or train etc) who is not operating it_ SYNONYMS=пътник*

TOWARDS MT

WS ANNOTATED INPUT

съгласно{съгласно ENG20 00524749- а}
Етичния{етичен} кодекс{кодекс BUŁ
159035502} **лекарят**{**лекар ENG20-**
09380179 н} има{имам ENG20
02139918- в} право{право ENG20
04881053- н} да{да BUŁ 1996281243}
откаже{откажа ENG20 00771620 в} да{да
BUŁ 1996281243} лекува{лекувам
ENG20 0076458 в} пациент{пациент
ENG20 0730887 н}

OUTPUT AFTER LEMMAS' DICTIONARY IS APPLIED



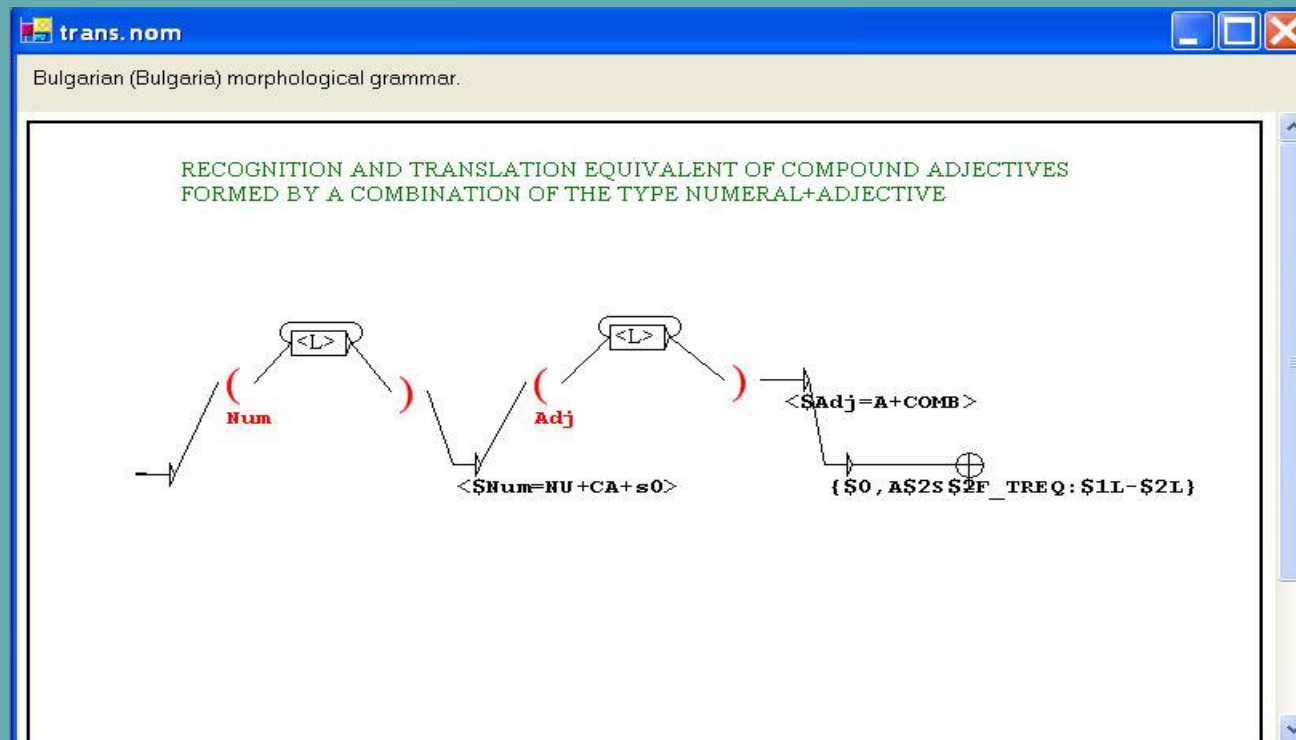
TOWARDS MT

- Assignment of translation equivalents on raw data:

The morphological grammars for recognition of derived words are enhanced to provide the translation equivalent through the super lemma.

This procedure is applicable to words not featuring in the dictionary derived from words which are dictionary entries – the already defined targets for AP.

GRAMMARS GENERATING TE



Grammar assigns TE to Bulgarian adjectives such as nine-meter, five-pound, etc.

DIFFICULTIES

- Ambiguities
 - Lexical;
 - Grammatical;
 - Structural.

trigodishen plan (three year plan)

trigodishno dete (three-year-old child)

- Disambiguation of head word and application of derivation on the disambiguated item.

FUTURE PERSPECTIVES

- Experiments;
- Recognition and translation of multiword-expressions (constituting more than one token) – 2 200 out of 46 000 annotated words are MWE;
- Rendition of syntactic patterns.