Generating sets of synonyms between languages

Ján Genči[1], Ondrej Dzurjuv[1],
Radovan Garabík[2]

[1]Department of Computers and Informatics,
Technical University of Košice

[2]Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava

Introduction

- Besides classical bi-/multi-lingual dictionaries on-line, there are EuroWordNet or Global WordNet exist
- Nor EuroWordNet neither Global WordNet cover the Slovak language
- There were several works dealing with automatic creation of synsets in the past
- A requirement to create computer application for building Slovak synsets

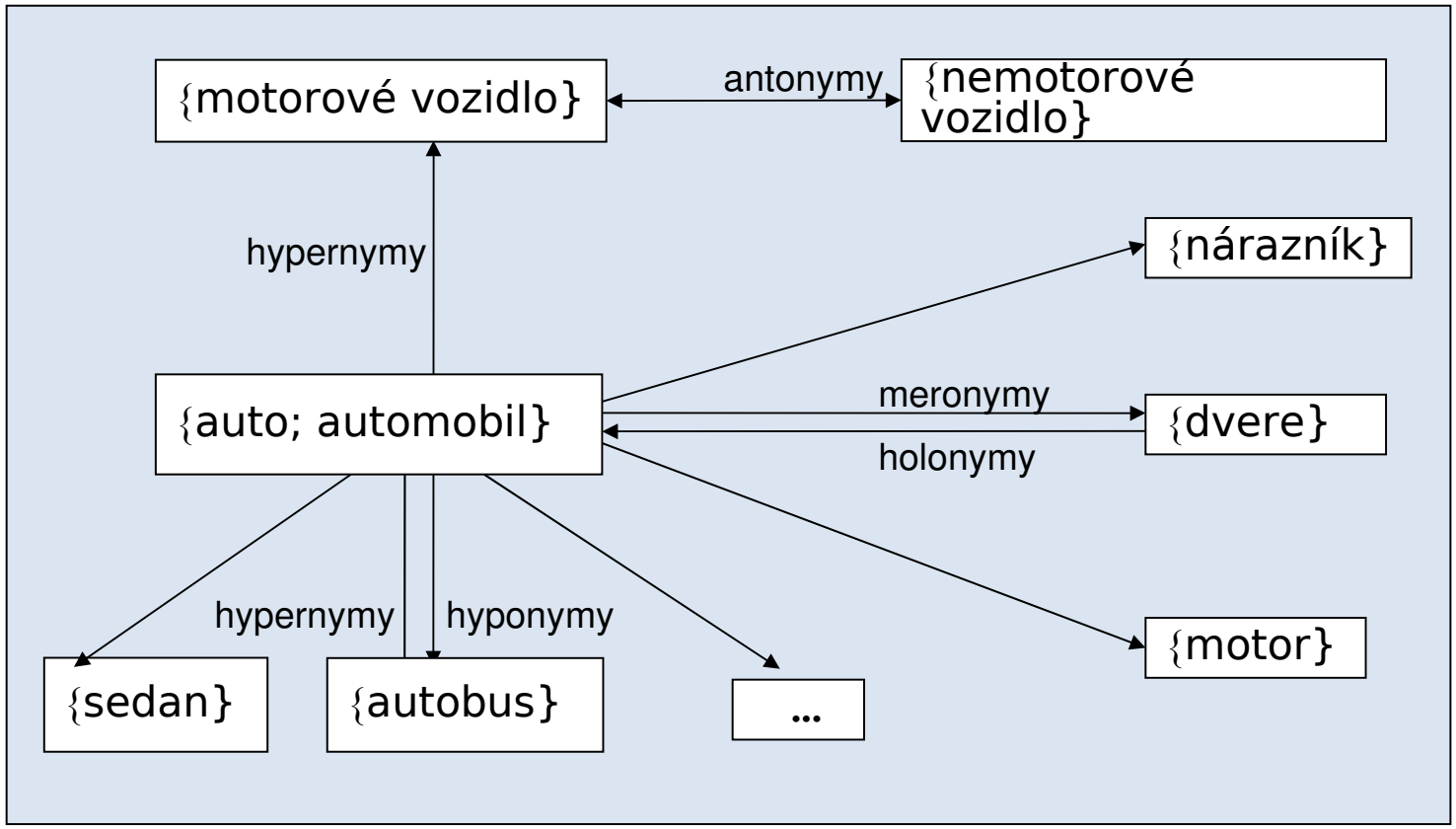
Our goals

- Design new methods of synsets generation
- Evaluate the quality of synset generation
- Design and implementation of computer application for assisted synset editing

WordNet

- Project WordNet started in 1990 at Princeton Univerzity
- Lexical database of English language; nouns, pronouns, verbs and adverbs
- Two base properties of WordNet
  - Words are organized into groups – sets of synonyms - synsets
  - Synsets are inteconnected

Relationships in WordNet



Method A

- Synonymy among words in synsets

$P_{S1}^{J_1}, P_{S2}^{J_1}, P_{S3}^{J_1}, \ldots, P_{Sn}^{J_1}$ – Set of words from synset

$t_{J_1 \to J_2}\left(P_{Si}^{J_1}\right) = \left\{P_{S1}^{J_2}, P_{S2}^{J_2}, \ldots, P_{Sp}^{J_2}\right\}, i = 1, \ldots, n$ – translation

$S := S \cup \left(t_{J_1 \to J_2}\left(P_{Si}^{J_1}\right) \cap t_{J_1 \to J_2}\left(P_{Sj}^{J_1}\right)\right)$

$pre\ i, j = 1, \ldots, p; i \neq j$ – result

{kind; sort; form; variety} ->

kind -> druh, rod, kategória

sort -> druh, akosť, trieda, typ, forma, chlap

form -> forma, tvar, podoba, formulár, blanketa, formula

variety -> rozmanitosť, odroda, výber, druh, rad, množstvo, mnohotvárnosť, rôznosť

∩

{druh, forma}

Method B

- Translation of univocal words

$P_{S1}^{J_1}, P_{S2}^{J_1}, P_{S3}^{J_1}, \dots, P_{Sn}^{J_1}$ - Set of words from synset

$P'^{J_1}_{S1}, P'^{J_1}_{S2}, \dots, P'^{J_1}_{Sm}$ - Subset of univocal words

$t_{J_1 \to J_2}\left(P'^{J_1}_{Si}\right) = \left\{P'^{J_2}_{S1}, P'^{J_2}_{S2}, \dots, P'^{J_2}_{Sk}\right\}, i = 1, \dots, m$ - translation

$S := S \cup t_{J_1 \to J_2}\left(P'^{J_1}_{Si}\right), i = 1, \dots, m$ - result

kind – 1 sense -> druh, rod, kategória
sort – 4 senses
form – 16 senses
variety – 6 senses

{druh, rod, kategória}

Method C

- It uses hyponymy and hyperonymy relationship for synset generation

$P_{S1}^{J_1}, P_{S2}^{J_1}, P_{S3}^{J_1}, \ldots, P_{Sn}^{J_1}$ - Words from synset

$P_{H1}^{J_1}, P_{H2}^{J_1}, P_{H3}^{J_1}, \ldots, P_{Hm}^{J_1}$ - Set of hyponym and hyperonym related to source synset

$$t_{J_1 \to J_2}\left(P_{Si}^{J_1}\right) = \left\{P_{S1}^{J_2}, P_{S2}^{J_2}, \ldots, P_{Sp}^{J_2}\right\}, i = 1, \ldots, n$$

$$t_{J_1 \to J_2}\left(P_{Hi}^{J_1}\right) = \left\{P_{H1}^{J_2}, P_{H2}^{J_2}, \ldots, P_{Hq}^{J_2}\right\}, i = 1, \ldots, m$$

- translation

$$S := t_{J_1 \to J_2}\left(P_{Si}^{J_1}\right) \cap t_{J_1 \to J_2}\left(P_{Hj}^{J_1}\right) = \left\{P_{S1}^{J_2}, P_{S2}^{J_2}, \ldots, P_{Sp}^{J_2}\right\} \cap \left\{P_{H1}^{J_2}, P_{H2}^{J_2}, \ldots, P_{Hq}^{J_2}\right\}$$

- result

$$pre\ i = 1, \ldots, n\ a\ j = 1, \ldots, m$$

{kind; sort; form; variety} -> {hyper,hypo}nyms: {category, type, brand, genus, species}

{kind; sort; form; variety} -> druh, rod, kategória, akosť, trieda, typ, forma, chlap, tvar, podoba, formulár, blanketa, formula, rozmanitosť, odroda, výber, rad, množstvo, mnohotvárnosť, rôznosť

{category, type, brand, genus, species} -> kategória, skupina, trieda, typ, symbol, litera, druh, odroda, značka, označenie, známka, kvalita, akosť, ohorok, rod, forma, tvar

∩

{druh; rod; kategória; akosť; trieda; typ; forma; tvar; odroda}

Method D

- It uses hyponymy and hyperonymy relationship for synset generation as source synsets

$P_{H1}^{J_1}, P_{H2}^{J_1}, P_{H3}^{J_1}, \dots, P_{Hk}^{J_1}$ - source synset hypernyms

$P_{Hr1}^{J_1}, P_{Hr2}^{J_1}, P_{Hr3}^{J_1}, \dots, P_{Hrl}^{J_1}$ - source synset hyponyms

$t_{J_1 \to J_2}\left(P_{Hi}^{J_1}\right) = \left\{P_{H1}^{J_2}, P_{H2}^{J_2}, \dots, P_{Hq}^{J_2}\right\}, i = 1, \dots, k$

$t_{J_1 \to J_2}\left(P_{Hri}^{J_1}\right) = \left\{P_{Hr1}^{J_2}, P_{Hr2}^{J_2}, \dots, P_{Hrr}^{J_2}\right\}, i = 1, \dots, l$

- translation

$S := t_{J_1 \to J_2}(P_{Hi}^{J_1}) \cap t_{J_1 \to J_2}(P_{Hrj}^{J_1}) = \left\{P_{H1}^{J_2}, P_{H2}^{J_2}, \dots, P_{Hq}^{J_2}\right\} \cap \left\{P_{Hr1}^{J_2}, P_{Hr2}^{J_2}, \dots, P_{Hrr}^{J_2}\right\}$

$pre\ i = 1, \dots, k\ a\ j = 1, \dots, l$

- target synset

{kind; sort; form; variety}:

hypernym: {category} -> kategória, skupina, trieda

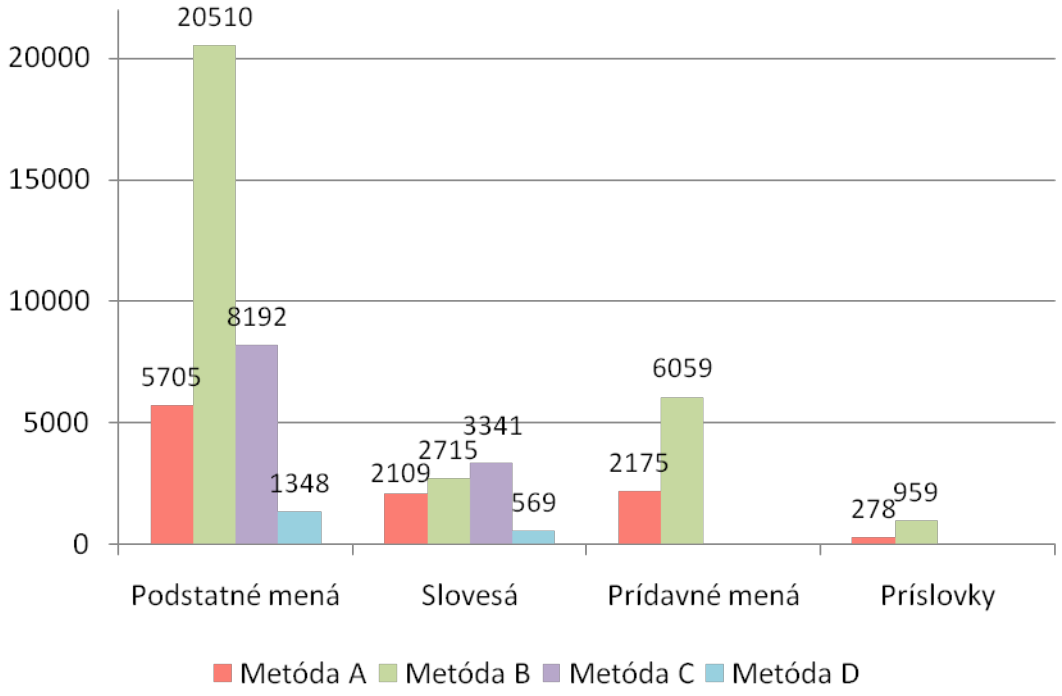
hyponym: {type, brand, genus, species} -> typ, symbol, litera, druh, odroda, značka, označenie, známka, kvalita, akosť, ohorok, druh, rod, skupina, trieda, forma, tvar

∩

{skupina, trieda}

Results

The WordNet contains 117659 synsets, Slovak equivalents were generated for 34.4 % (40521) of English synsets



Method A: 8.7 %

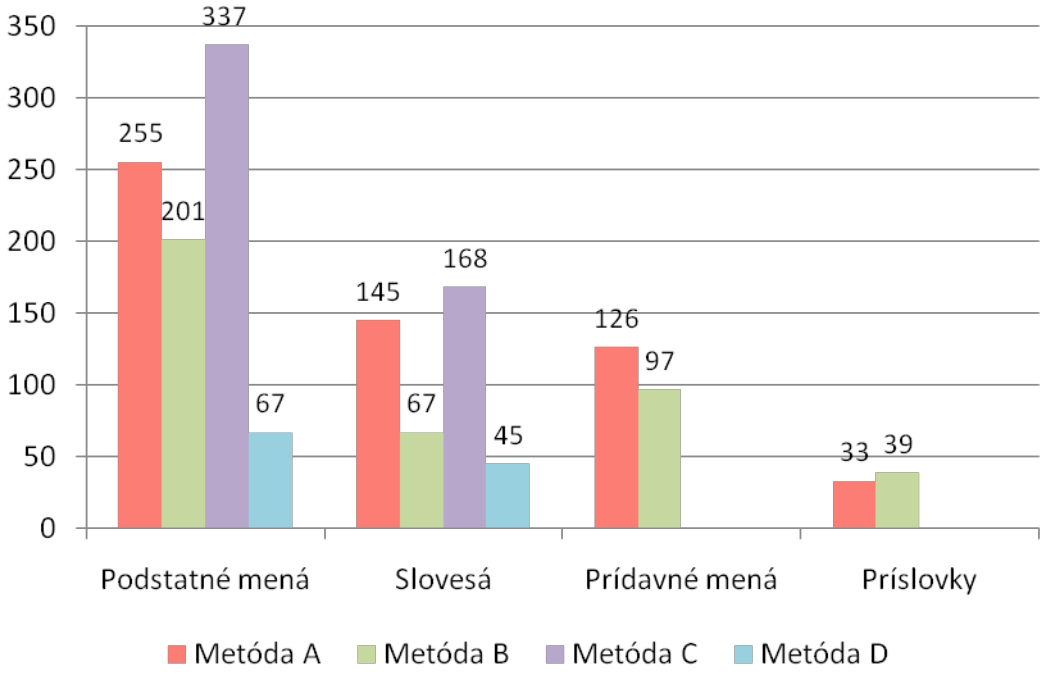
Method B: 25.7 %

Method C: 12 %

Method D: 1.4 %

Results for frequently used words

For set of 300 more frequently used English words there are 1709 synsets in WordNet, Slovak equivalents were produced for 55.4 % (946) synsets



Method A: 32.7 %

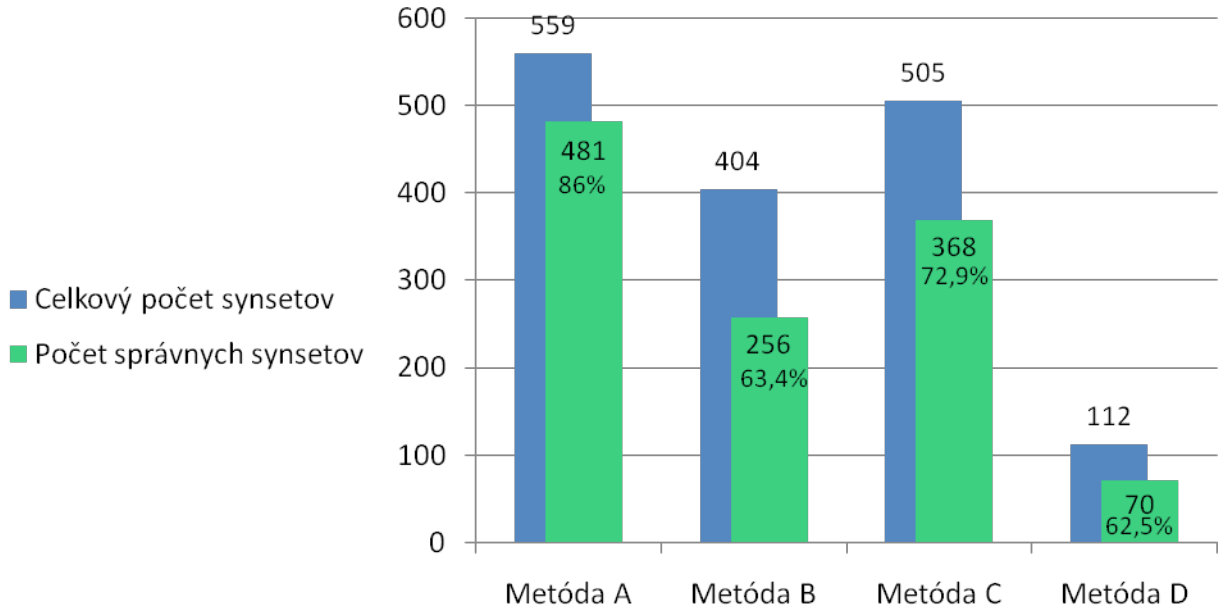
Method B: 23.6 %

Method C: 29.6 %

Method D: 6.6 %

Inspection of Slovak synsets

Part of speech and semantics of generated Slovak synsets were inspected.



Results

- New computer application for manual synset translation. Application was “fed” by generated data
- About 60% of synsets were generated
- About 75% of generated synsets were evaluated as “not incorrect“

Conclusion

- We were able to produce Slovak synsets for 34.4% of English synsets. Manual inspection was required.
- Process of synset generation is influenced by many factors – translation dictionaries used, frequency of words, number of words in synset.
- Application – used to create en-sk-pl-de-lt dictionary (with ontology relations)

Filter

Identifikátor anglického synsetu: 06410904 (identifikátor synsetu je zložený z 8 číslic)

Prechádzať: podstatné mená

Zobraziť prislúchajúce ☑ slovenské, ☑ nemecké, ☑ poľské, ☑ litovské synsety

Zobraziť

Zdrojový anglický synset				
Anglické synsety	**Slovenské synsety**	**Nemecké synsety**	**Poľské synsety**	**Litovské synsety**
EN synset: 06410904 { book } a written work or composition that has been published (printed on pages bound together); "I am reading a good book on economics" +SYN | -SYN	SK synset: 3538 { kniha; knižka } ↘ | +EN | -EN | × | ☑ | (i)	GER synset: 1737 { s Buch; s Büchlein } ↘ | +EN | -EN | × | ☐ | (i)	PL synset: 1686 { książka; książeczka } ↘ | +EN | -EN | × | ☐ | (i)	LAT synset: 1404 { knyga; knygelė } ↘ | +EN | -EN | × | ☐ | (i)
Antonym - počet anglických synsetov: 0				
Hypernym - počet anglických synsetov: 1				
Anglické synsety	**Slovenské synsety**	**Nemecké synsety**	**Poľské synsety**	**Litovské synsety**
EN synset: 06589574 { publication } a copy of a printed work offered for distribution +SYN | -SYN	SK synset: 835 { publikácia } ↘ | +EN | -EN | × | ☑ | (i)	GER synset: 1795 { e Publikation; e Veröffentlichung } ↘ | +EN | -EN | × | ☐ | (i)	PL synset: 1744 { publikacja } ↘ | +EN | -EN | × | ☐ | (i)	LAT synset: 4510 { leidinys } ↘ | +EN | -EN | × | ☐ | (i)

Part of speech	Nouns	Adjectives	Verbs	Adverbs	Totals
Unique strings	12941	3321	1150	982	18394
Strings with one sense	10239	2305	953	702	14199
Word-sense pairs	18740	5551	1400	1505	27196
Synsets	9317	2329	830	549	13025
Synsets of one word	3916	773	426	141	5256

Thank you for your attention