Language Technology for Polish in Practice
Basic language resources for Polish

Maciej Piasecki, Marek Maziarz, Marcin Oleksy, Ewa Rudnicka
Wrocław University of Science and Technology
G4.19 Research Group
maciej.piasecki@pwr.wroc.pl
2017-01-17
Basic Resources

- **Morphological dictionary** (Woliński, 2014)
- Text corpus annotated on several levels
  - KPWr – Korpus Politechniki Wrocławskiej (Wrocław University of Technology Corpus) (Broda et al., 2012)
- Complex system of lexico-semantic resources
  - Proper Names – NELexicon 2
  - Lexical meanings – plWordNet 3.0 emo (Słowosieć 3.0 emo)
    - via bilingual mapping to a multiligual, global resource
    - mapped onto several other resources, e.g. SUMO ontology
  - **Syntactic-semantic valency lexicon** – Walenty (Przepiórkowski et al., 2014) IPI PAN
- Lexicon of Multiword Expressions (structurally described) (Kurc et al., 2012)
KPWr Corpus

- **Origin** (Broda et al., 2012)
  - in 2010 Polish National Corpus was in development and not accessible
  - manually annotated corpus was needed for LTs

- **Idea**
  - only texts on Creative Commons (CC BY-SA 3.0)
  - different sources, styles and genres
  - morpho-syntactically segmented and tagged (Morfeusz+WCFRT1)
  - manually annotated on several layers
    - several turns, verified
    - gradually expanded and converted to 2+1 annotation system
    - e.g. semantic roles (100 documents) and situations (150)
## KPWr Corpus: structure

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>Documents</th>
<th></th>
<th>Tokens</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>%</td>
<td>No</td>
<td>%</td>
</tr>
<tr>
<td>Blogs</td>
<td>171</td>
<td>10.48</td>
<td>52793</td>
<td>11.80</td>
</tr>
<tr>
<td>Longer press articles</td>
<td>132</td>
<td>8.09</td>
<td>41181</td>
<td>9.20</td>
</tr>
<tr>
<td>Dialogues</td>
<td>91</td>
<td>5.58</td>
<td>30070</td>
<td>6.72</td>
</tr>
<tr>
<td>Shorter press articles</td>
<td>221</td>
<td>13.55</td>
<td>34284</td>
<td>7.66</td>
</tr>
<tr>
<td>Science</td>
<td>87</td>
<td>5.33</td>
<td>28269</td>
<td>6.32</td>
</tr>
<tr>
<td>Popular-scientific and textbooks</td>
<td>73</td>
<td>4.48</td>
<td>22463</td>
<td>5.02</td>
</tr>
<tr>
<td>Older literature</td>
<td>86</td>
<td>5.27</td>
<td>36094</td>
<td>8.06</td>
</tr>
<tr>
<td>Contemporary literature</td>
<td>42</td>
<td>2.58</td>
<td>19101</td>
<td>4.27</td>
</tr>
<tr>
<td>Religion</td>
<td>9</td>
<td>0.55</td>
<td>5357</td>
<td>1.20</td>
</tr>
<tr>
<td>Parliamentary proceedings</td>
<td>79</td>
<td>4.84</td>
<td>32297</td>
<td>7.22</td>
</tr>
<tr>
<td>Technical</td>
<td>17</td>
<td>1.04</td>
<td>4373</td>
<td>0.98</td>
</tr>
<tr>
<td>Official</td>
<td>62</td>
<td>3.80</td>
<td>18890</td>
<td>4.22</td>
</tr>
<tr>
<td>Parliamentary bills</td>
<td>80</td>
<td>4.90</td>
<td>31620</td>
<td>7.06</td>
</tr>
<tr>
<td>Wikinews</td>
<td>123</td>
<td>7.54</td>
<td>28264</td>
<td>6.31</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>358</td>
<td>21.95</td>
<td>62520</td>
<td>13.97</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1631</strong></td>
<td><strong>447576</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
KPWr Corpus: functional styles

Functional styles in KWPr 1.2

Target functional styles in KWPr
## KPWr Corpus: annotations

<table>
<thead>
<tr>
<th>Name</th>
<th>Content</th>
<th>Index</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>index_chunks</td>
<td>Syntactic chunks (several types)</td>
<td>307</td>
<td>94831</td>
</tr>
<tr>
<td>index_chunks_rel</td>
<td>Relations between chunks</td>
<td>305</td>
<td>-</td>
</tr>
<tr>
<td>index_names</td>
<td>Named Entities (NE), mostly Proper Names</td>
<td>1343</td>
<td>28816</td>
</tr>
<tr>
<td>index_names_lemma</td>
<td>Lemmas for NEs</td>
<td>1343</td>
<td>-</td>
</tr>
<tr>
<td>index_names_rel</td>
<td>Semantic relations between NE</td>
<td>1118</td>
<td>-</td>
</tr>
<tr>
<td>index_wsd_nv</td>
<td>Disambiguated word senses</td>
<td>1179</td>
<td>13555</td>
</tr>
<tr>
<td>index_spatial</td>
<td>Spacial expressions (triggers)</td>
<td>1526</td>
<td>7873</td>
</tr>
<tr>
<td>index_spatial_rel</td>
<td>Spacial expressions (combinations)</td>
<td>1531</td>
<td>-</td>
</tr>
<tr>
<td>index_zero_verb</td>
<td>Verbs with subject ellipsis</td>
<td>969</td>
<td>7111</td>
</tr>
<tr>
<td>index_keywords</td>
<td>Textual keywords (extractive)</td>
<td>1629</td>
<td>10552</td>
</tr>
<tr>
<td>Name</td>
<td>Content</td>
<td>Index</td>
<td>Statistics</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------------------------------------------</td>
<td>-------</td>
<td>------------</td>
</tr>
<tr>
<td>index_timex</td>
<td>Temporal expressions</td>
<td></td>
<td>1630</td>
</tr>
<tr>
<td>index_timex_local</td>
<td>Local normalisation of temporal expressions</td>
<td></td>
<td>1606</td>
</tr>
<tr>
<td>index_timex_global</td>
<td>Global normalisation of temporal expressions</td>
<td></td>
<td>1630</td>
</tr>
<tr>
<td>index_events</td>
<td>Situations (triggers)</td>
<td></td>
<td>333</td>
</tr>
<tr>
<td>index_events_g0</td>
<td>Situations (triggers) gold set #0</td>
<td></td>
<td>101</td>
</tr>
<tr>
<td>index_events_g1</td>
<td>Situations (triggers) gold set #1</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>index_events_g2</td>
<td>Situations (triggers) gold set #2</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>index_events_g3</td>
<td>Situations (triggers) gold set #3</td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>index_serol</td>
<td>Semantic roles</td>
<td></td>
<td>1037</td>
</tr>
<tr>
<td>index_coref</td>
<td>Co-reference</td>
<td></td>
<td>50</td>
</tr>
</tbody>
</table>
KPWr Corpus: annotations

Liczba anotacji w dokumentach

- Wymażenia temporalne: 10552
- Wymażenia przestrzenne (w...): 12726
- Sytuacje (wyznaczki): 7873
- Sytuacje (wyznaczki) zbi...: 8686
- Sytuacje (wyznaczki) zbi...: 3077
- Sytuacje (wyznaczki) zbi...: 2357
- Sytuacje (wyznaczki) zbi...: 2364
- Role semanticzne: 1321
- Czasowniki z podmiotem d...: 2199
- Koreferencja: 4868
- Tekstowe słowa kluczowe: 1629
- Wymażenia temporalne: 1630
- Sytuacje (wyznaczki): 1526
- Sytuacje (wyznaczki) zbi...: 333
- Sytuacje (wyznaczki) zbi...: 101
- Sytuacje (wyznaczki) zbi...: 50
- Role semanticzne: 50
- Role semanticzne: 34
- Czasowniki z podmiotem d...: 969
- Czasowniki z podmiotem d...: 037
- Koreferencja: 50
Średnia liczba anotacji w dokumencie

- Tekstowe słowa kluczowe: 6,478
- Wyrażenia temporalne: 7,807
- Wyrażenia przestrzenné: 5,159
- Sytuacje (wyznaczniki): 26,084
- Sytuacje (wyznaczniki): 30,465
- Sytuacje (wyznaczniki): 47,140
- Sytuacje (wyznaczniki): 47,280
- Role semantyczne: 38,853
- Czasowniki z podmiotem: 2,121
- Koreferencja: 7,338
- Wszystko: 97,360
KPWrC inter-annotator agreement

- **Situations**

<table>
<thead>
<tr>
<th>Types</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>action</td>
<td>0.90</td>
</tr>
<tr>
<td>aspectual</td>
<td>0.88</td>
</tr>
<tr>
<td>perception</td>
<td>0.90</td>
</tr>
<tr>
<td>reporting</td>
<td>0.93</td>
</tr>
<tr>
<td>i_action</td>
<td>0.68</td>
</tr>
<tr>
<td>i_state</td>
<td>0.85</td>
</tr>
<tr>
<td>State</td>
<td>0.75</td>
</tr>
<tr>
<td>Light predicate</td>
<td>0.67</td>
</tr>
</tbody>
</table>
KPWrC inter-annotator agreement

- Semantic roles inside Noun Phrases

<table>
<thead>
<tr>
<th>Role</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>actor</td>
<td>0.89</td>
</tr>
<tr>
<td>agent</td>
<td>0.89</td>
</tr>
<tr>
<td>cause</td>
<td>0.89</td>
</tr>
<tr>
<td>undergoer</td>
<td>0.91</td>
</tr>
<tr>
<td>patient</td>
<td>0.92</td>
</tr>
<tr>
<td>theme</td>
<td>0.78</td>
</tr>
<tr>
<td>manner</td>
<td>0.84</td>
</tr>
<tr>
<td>attribute</td>
<td>0.92</td>
</tr>
<tr>
<td>amount</td>
<td>0.80</td>
</tr>
</tbody>
</table>
plWordNet (Słowosieć): Goal

To build a wordnet which provides a faithful and comprehensive description of the system of Polish lexical semantics

- its structure should represent accurately the lexico-semantic relations between lexical meanings in Polish
- and be motivated only by observations derived from Polish language data
- any form of translation from wordnets for other languages was excluded
- a resource with good coverage with respect to lemmas, word senses and instances of lexico-semantic relations
- in close correspondence to language data collected from very large corpora
A complex system of lexico-semantic resources (Maziarz et al. 2016)

- Walenty (valence lexicon) > 15,000 lemmas
- MWE lexicon (54,000)
- Proper Names (2.4 mln)
- Ontology SUMO (upper and intermediate level)
- enWordNet 1.0 (ext.: WordNet 3.1 plus 10k senses)

Lexico-syntactic structures
Methodological choices

- Transfer method
  - by translating WordNet – automatically or semi-automatically
  - can introduce some properties alien to the Polish lexical system
- Merge method – requires source dictionaries on an open licence
- WordNet synset definition
  - near synonymy or lexicalised concepts, appeared to be not clear enough to be used operationally
Corpus-based wordnet development process

- A large text corpus is primary data
  - Lemmas (starting with the most frequent)
  - Examples of use and senses
- Language tools and systems support corpus exploration
  - simple, e.g. concordances
  - advanced – extraction of semantic similarity, relations, sense clusters
  - combined – semi-automated wordnet expansion (Paintball algorithm, RANLP 2013)
- Process
  - systematic extraction of lemmas, acquisition of lexico-semantic knowledge, generation of suggestions, decisions of editors
  - supported by: dictionaries, encyclopaedias, intuition, team
plWordNet model

- Corpora contain words, senses discernible by context, not sets of synonyms
- Lexical unit (LU)
  - a triple: <part of speech, lemma, sense id>
  - the basic building block in plWordNet, belongs to one synset
- Synset - a group of LUs which share
  - lexico-semantic constitutive relations, e.g. hyper/hyponymy, mero/holonymy
  - and constitutive features: stylistic register, aspect, and semantic classes for adjectives and verbs
- A relation between two synsets is a shorthand for sharing relations between LUs
- Minimal Commitment Principle
Wordnet relations

- Lexico-semantic relations form a continuum of semantic distinctions
- Relations established in linguistics are subspaces of the continuum with fuzzy borders
  - e.g. hypernymy, meronymy etc.
- Wordnet is a **generalisation** derived from that relation continuum
  - some distinctions preserved, some deemphasised
Wordnet relations

1. Suitable for the **construction** of generalisations,
2. Suitable for the **application** of generalisations in NLP tasks,
3. **Compatible** with other wordnets
Minimal Commitment Principle

- **Minimal commitment principle:**
  construct a wordnet with as few assumptions as possible

- A wordnet should not favour any of the existing theories of meaning
  - little must be assumed to create even a very large wordnet
  - stay close to language data and simple observations

- Wordnet is a compromise between:
  - expressive power of such a system’s description
  - and the workload required to construct that description
Synset

- Describing synonymy without `synonymy`
- Feature sharing among lexical units
- Synset: a set of lexical units that cannot be distinguished
- Lexical units are described by instances of wordnet relations
Lexical units from the same synset should share instances of a selected subset of wordnet relations.

Synset comprises those LUs which share a set of lexico-semantic relation targets.

If synsets $S_1$ and $S_2$ are linked by relation $R$ is to say that any pair of LUs $l_1$ and $l_2$, such that $l_1 \in S_1$ and $l_2 \in S_2$, is an instance of $R$. 
Synset

- Example
  \{afekt 1 `passion’, uczucie 2 `feeling’\} \longrightarrow \text{hypernym} \rightarrow

  \{miłość 1 `love’, umiłowanie 1 `affection’, kochanie 1 ~ `loving’\}

- Synset as a notational convention
  - for a group of lexical units sharing certain relations

- What are wordnet relations?

- Are relations enough to define synsets?
Constitutive relations

- Required properties
  - well-established in linguistics
    - good understanding (e.g. paradigmatic relations)
    - existing descriptions
  - definable with sufficient specificity
  - and useful in generalisation
    - relatively frequent
    - should describe sets of LUs systematically - a sharing factor

- Level of generalisation of a wordnet vs selection of the constitutive relations
The quest for constitutive relations

- Linguistics - paradigmatic relations
  - (Cruse 2004) identity (synonymy), inclusion (hyponymy and meronymy) and exclusion/opposition (antonymy) and incompatibility (co-hyponymy, co-meronymy)

- Wordnets
  - hyper/hyponymy, holo/meronymy, different forms of entailment

- Sharing factor
  - the average size of the LU group such that they share the given relation
  - e.g. antonymy and derivational relations are seldom shared among groups of LUs
Constitutive relations: example

- **lustro** and **zwierciadło** both (Maziarz et al., 2013)
  - denote a mirror
  - are **hypernyms** of:
    - **lustro weneckie** ‘Venetian mirror’
    - and **tremo** ‘trumeau mirror, pier glass, pier mirror’
  - are **hyponyms** of **przedmiot** ‘object’
- **szkło** ‘glass’ is a **meronym** of both **lustro** and **zwierciadło**
  ⇒ a synset: {}\{lustro, zwierciadło\}⇒ **synonyms**
Constitutive relations: example

- **grać** ‘play_{impf}’
- **remisować** ‘draw_{impf}’
- **szachować** ‘check_{impf}’
- **matować** ‘checkmate_{impf}’
- **patować** ‘cause stalemate_{impf}’
- **(A) grać w szachy** ‘play chess_{impf}’
- **zwyciężyć** ‘win_{pf}’
- **dać szacha** ‘check_{pf}’
- **zaszachować**
- **zamiatować, dać mata** ‘checkmate_{pf}’
Constitutive relations: example

**splajtować, zbankrutować** ‘go bankrupt’

*bankrut* ‘bankrupt’

**wyłysieć** ‘go bald’

*bezwłosy* ‘hairless’, *łysy* ‘bald’
Constitutive features

- Wordnet structure as a basis for acceptable conclusions
  - lack of formal definitions
  - some conclusions based on properties of relations, e.g. transitivity of hypernymy
- Additional constraints on the relation definitions
  - meta-conditions
  - obligatory and built into the relation definitions
- plWordNet: stylistic registers, semantic verb classes and aspect
Stylistic register

- Not only denotation, but usage also matters
- Stylistic register
  “a variety of language with a particular situational context, such as an occupation or social activity” (Hartmann and James 1998)
- Example of a problem:
  \{cop, pig, policeman, police officer\} (Vossen, 2002)
  A police captain is a pig with certain properties. It is a police captain and therefore also a pig. If it is a police captain then it must be a pig.
Stylistic register

- Inter-register synonymy
  - links LUs of the same POS, near synonyms
  - share all constitutive relations, except hypernymy
  - differ in stylistic registers
  - described by classes of equivalence of register values

\Leftrightarrow \{\textit{kibel} ‘bog (Br.), loo (Am.)’, \textit{klop} ‘bog, loo’\}

-\text{hyper} → \{\textit{szalet} ‘public toilet’\}
-\text{hyper} → \{\textit{latryna} ‘latrine’\}
-\text{hyper} → \{\textit{pisuar} ‘toilet with urinal(s)’\}
Semantic verbal classes and aspect

- **Dictionaries**
  - verbs are explained by verbs (genus proximum) of the same class
    - e.g. achievement, stative, activity, etc.
  - the same aspect in definitions

- **Constraints imposed on relation definitions**
  - most relations: the identity of aspect and semantic class
  - verb classes as artificial lexical units
Assumptions and principles

- **Construction from scratch**
  - focus on lexico-semantic facts specific to Polish
  - system of relations adapted to Polish
  - stylistic register, aspect and derivations among primary considerations
  - not using translation, avoiding influence from English WordNet

- **Hypernymy forest instead a conceptual root**
  - only hypernymy links motivated by the linguistic definition
  - mapping of top synsets onto SUMO ontology (in progress)
Assumptions and principles

- Problematic Proper Names
  - a very large, open category, dynamically changing
  - priority: nearly comprehensive coverage of common nouns, verbs and adjectives
  - Named Entity Recognition techniques:
    - include gazetteers, provide semantic classification

- Very limited coverage of Proper Names in plWordNet
  - generally not included
  - one exception: geographical names from which a large class of common names is derived - inhabitants
  - e.g., warszawianin (Warsaw citizen) from Warszawa (Warsaw)
Assumptions and principles

- **Network density**
  
  average number of relation instances going from a LU to any other LU in the wordnet

- **Choice of relations**
  
  - good and balanced coverage = high network density
  - larger number of relations → larger density
  - but ... excessively fragmented description
  - constitutive relations as the basis
  - extended with:
    - relaxed condition of the sharing factor, e.g. antonymy
    - derivational relations
Assumptions and principles

- Limited resources at the starting point
  - translation ruled out
  - no electronic monolingual dictionaries to leverage
- Corpus-based Wordnet Development
  - a large corpus (~2 billion words) as a source of lemmas
    - including those not covered by dictionaries
  - dictionaries consulted to reduce corpus frequency bias
  - automated extraction of potential instances of lexico-semantic relations
  - editors supported by integrated WordnetWeaver and WordnetLoom systems
Synset relations

- **Hypernymy/hyponymy**
  - defined for all parts of speech
  - also for verbs
  - adjectives and adverbs - limited but surprisingly numerous

- **Inter-register synonymy**
  - nouns, verbs, adjectives and adverbs
  - \( \approx \) synonymy across different stylistic registers
  - links stylistically marked lexical units with their unmarked counterparts
  - e.g. *samochód* ‘a car’ - *fura* ‘a car (slang)’
Holonymy/meronymy
- between nouns, but also verbs
- typical subtypes: part, place, portion, element of a collection and substance
- taxonomic unit
  - e.g. *kotowate* (Felidae) — *kotokształtne* (Feliformia)
  - also connects graph of scientific terms with general-language LUs
- two special sub-types for verbs
Synset relations

- **Type-instance**
  - Proper Names linked to common nouns

- **Inhabitant**
  - based on the derivational relation, but expanded to synsets
  - at least two synset members must be derivationally associated
  - e.g., *Japończyk* ‘Japanese’ – *Japonia* ‘Japan’

- Fuzzynymy
Verb Synset relations

- Hyper/hyponymy
- Meronymy & holonymy
- Cause
- Process
- Inchoativity
- State

- Multiplicativity
  - iterativity
  - distributivity
- Presupposition
- Preceding
Motivation

- Verbs in wordnets
  - the second-largest component of any wordnet
  - properties going far beyond lexical semantics
  - variously described across wordnets

- What is the point between too much and too little verb semantics in a wordnet?

- Verbs in a wordnet in which the lexical unit is the basic building block

- Need: a set of relations for verbs which will ensure a clear account of all differences among lexical units
## Core verb relations

<table>
<thead>
<tr>
<th>Princeton WN</th>
<th>GermaNet</th>
<th>EuroWN</th>
<th>plWordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>synonymy</td>
<td>synonymy</td>
<td>synonymy</td>
<td>synonymy (V-V)</td>
</tr>
<tr>
<td>antonymy</td>
<td>antonymy</td>
<td>antonymy</td>
<td>antonymy (V-V)</td>
</tr>
<tr>
<td>troponymy</td>
<td>hyponymy</td>
<td>hyponymy</td>
<td>hyponymy (V-V)</td>
</tr>
<tr>
<td>entailment (proper inclusion)</td>
<td>subevent</td>
<td>subevent</td>
<td>meronymy (V-V)</td>
</tr>
<tr>
<td>entailment (backward presupposition)</td>
<td>entailment</td>
<td>cause</td>
<td>preceding (V-V)</td>
</tr>
<tr>
<td>cause</td>
<td>cause*</td>
<td>cause</td>
<td>causality* (V-V, V-N, V-Adj)</td>
</tr>
<tr>
<td>‘near’ relations*</td>
<td>cross-categorial synonymy*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fuzzynymy</td>
<td>fuzzynymy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* means a cross-categorial relation
Perspectives on verb hyponymy

- tropeonomy in PWN
  - “a manner relation”, temporally coextensive (Fellbaum, 98)
  - noun tests do not apply
- hyponymy in EWN
  - to X is to Y + AdvP/AdjP/NP/PP
  - co-extensive situations
- both: hyponym - the *definiendum*, and hypernym - the head of the *definiens*

plWordNet’s verb hyponymy

- X-ować to Y-ować w specjalny sposób
- ‘to X is to Y in a special way’
- *nawilżać `to moisten` -- moczyć ‘to make wet’
  and *nawilżać* = ‘to make wet slightly’
Assumption:
- verb denotation = a set of situations
- situation: participants, temporal span, constituents (sub-situations)
- e.g. *jeść* `eat': 〈gryźć〉 `chew\textsubscript{impf}', 〈połykać, przełykać〉 `swallow\textsubscript{impf}' and 〈połknąć, przełknąć〉 `swallow\textsubscript{pf}'

Verb meronomy and holonomy
Entailment

- Verb meronymy and holonymy
  - sub-situation
    - a composite situation and its component
    - the idea of a sub-situation is an integral and typical part of the situation of the holonym
  - accompanying situation
    - accounts for a “primary” situation, represented by the holonym,
    - typically supplemented by another situation, represented by the meronym
  - 9.4% of plWordNet’s verb synset relations
Extended Verb Synset Relations

- Core verb relations do not differentiate lexical units sufficiently well
  - low network density of verbs in plWordNet 1.0
- 19 relations not taken over from PWN, EWN or GermaNet
  - to further specify the relations noted in wordnets
  - to account for the intricacies of the Polish derivational system
Extended Verb Synset Relations

- **Converseness (LU, V-V)** - synonymy of expressions after shift of predicate’s actants:
  
  A kupił C od B = B sprzedał C A
  
  ‘A bought\text{pf} C from B’ = ‘B sold\text{pf} C to A’

- **Inter-register synonymy (synset, V-V)**
  
  identity of referential meaning - difference in registers
  
  In PWN such synset pairs are linked with hyponymy (with addition of domain relation)
  
  \{zjeść, ...\} `to eat\text{pf}' - \{zeżrzeć, wszamać\} `slang. to eat\text{pf}'
**Preceding** (synset, V-V)

~ presupposition of few alternatives

\{wstać, podnieść się, powstać\} `get up\textsubscript{pf}'

- \{siedzieć\} `sit\textsubscript{impf}' | \{leżeć, spoczywać\} `lie\textsubscript{impf}'

\textit{stać} `stand'

\textit{leżeć} | \textit{siedzieć}

`lie' | `sit'

\textit{wstać} `get up'
There is a productive schema of derivation

It could be generalized into semantic relation

Unconnected derivationally pairs hold tests of the relation

For instance,

- ze-starz-eć się `to become old’ < star-y `old’
- *stary* has synonyms *wiekowy* and *niemłody* which hold test for the relation
- ze-starzeć się = to become *wiekowy*, *niemłody*
Relations Motivated Derivationally

- **State** (synset, V-N, V-Adj)
  describes a state by a noun or an adjective
  
  \{królować, tronować\} `to reign_{impf}' - \{król\} `king'
  \{żyć\} `live' - \{żywy\} `alive'*)

  *) the V-Adj type is equivalent to EWN XPOS Synonymy (state-denoting verb - adjective)

- **Inchoativity** (synset, V-V)
  describes a beginning of a situation
  
  \{zasnąć, usnąć\} `fall_{pf} asleep' - \{spać\} `sleep_{impf}'
Processuality (synset, V-N, V-Adj)

change of a situation (described by a noun or adjective)

\{zgłupieć\} `become\textsubscript{pf} a fool’ or `become\textsubscript{pf} stupid’
- \{głupek, debil, dureń, …\} n. `fool’
- \{głupi, nierozumny, ciasny, …\} adj. `stupid’

`not to be stupid’

\begin{center}
\textit{processuality}
\end{center}

\begin{center}
\textit{zgłupieć `become stupid’}
\end{center}

\begin{center}
\textit{głupi `stupid’}
\end{center}
Multiplicativity (V-V)
repeated actions, iterative and distributive aspect
derivational relation possible but not necessary
Iterativity `from time to time’
\{jeść, spożywać, \ldots\} `eat_{impf}` - jadać `eat_{impf} from time to time’
Distributivity `done on many things or about action performed by many subjects’
\{pozabijać\} `to kill_{pf} many man or animals’ - \{zabić\} `to kill_{pf}`
Derivational Relations

- Relations between LUs
- Frequent in dictionaries’ definitions
- **Reflexivity** (V-V) `oneself’
  - Action directed by a subject on herself/himself
  - *myć się* `wash oneself’ - *myć* `wash’
- **Reciprocality** (V-V) `each other’
  - Action performed by two (or many) subjects directed on each other
  - *bić się* `to bit each other’ - *bić* `to bit’
Derivational Relations

- **Role inclusion & Role**
  - Agent, Patient, Instrument, Location, Time
  - relations similar to EWN Role & Involved
  - but between LUs (not synsets)
  - they are **not** inverse relations
  - Role = deverbal nouns
    
    zszywacz `stapler' < zszywać `to push staples through pieces of paper' [Instrument]

  - Role inclusion = denominal verbs
    
    solić `to salt' < sól `salt' [Instrument]
Aspect Semantics

- Aspectual derivation
  - among most prominent derivational relations in Slavic languages
  - difficult to solve using wordnet relations
    - e.g., CzechWN: HAS-SUBEVENT but also HAS IMPF, HAS PERF, HAS ITER
    - BulNet: aspectual pair in one synset
  - encodes lexical meaning shift
- plWordNet: semantic aspect relation
  - pure aspectuality
  - secondary aspectuality
Aspect Semantics

Pure aspectuality
  Perfective ~ finished situations
  Imperfective ~ ongoing situations

Secondary aspectuality
  Perfective/imperfective aspect + Aktionsart

_Aktionsarten_ = kinds of actions - causative, inchoative, iterative, momentary etc.
Expressed with suffixes or suffixes, or root alternations
Verb classes

- States (*imperfectiva tantum*, primarily atelic, static):
  e.g., spać ‘sleep’, kosztować ‘cost’,
- Activities (*imperfectiva tantum*, primarily atelic, dynamic, changeless):
  e.g., tańczyć ‘dance’, jeść ‘eat’,
- Accomplishments (*imperfectives and perfectives*: primarily telic),
  e.g., budować ‘build₉’ - wybudować ‘build₉’
Verb classes

- Achievements (*perfectiva tantum*, momentary verbs):
  - e.g., *zgubić* ‘lose\(\text{pf}\) (something)’, *klęknąć* ‘kneel\(\text{pf}\)’
- Dynamic - atelic - non-momentary perfectives (*perfectiva tantum*)
  - e.g., *posiedzieć* ‘sit (for a while)’,
    *przesiedzieć* ‘sit (for some time)’,
    *nakłamać* ~‘lie (a lot)’,
    *pozabijać* ‘kill a number of people, animals’
Aspect and verb classes vs relations

- Aspect and verb classes influence the distribution of relations
- Examples:
  - Hyponymy is limited only to verbs of the same aspect and class
  - Processuality links accomplishments with adjectives and nouns
  - Inchoativity links inchoatives (achievements and activities) with states and activities
  - Distributivity (a sub-type multiplicativity) links distributive verbs (a sub-class of dynamic - atelic - non-momentary perfectives) with perfective accomplishments or achievements
Lexical unit relations

- **Antonymy**
  - all parts of speech
  - *complementarity* - polar pairs of LUs with opposite and mutually exclusive meanings
  - *gradable opposition* - non-exhaustive oppositions

- **Converseness**
  - nouns and verbs
  - mutually opposite roles assigned to the arguments
  - for nouns:
    If A is X (Prep) B, then B is Y (Prep) A
    e.g., *If A is a husband of B, then B is a wife of A*
Derivationally motivated relations

- Cross-categorial synonymy
- Feature bearer & State
- Femininity
- Markedness
  - diminutive
  - augmentative
  - young being
- Semantic role & Role inclusion
- Derivational
Derivational & derivationally motivated relations

- Relation between a derivative and its base
- In some cases expanded into pure semantic relation
- Clear semantics
- Important in defining lexical units and synsets
- Only most frequent phenomena in Polish
- Most productive derivational schemas
- >100,000 instances in plWordNet 2.0
Derivational relations

- Cross-categorial synonymy
  - `NEAR’ relations in EWN
  - ‘transpositional’ or ‘syntactic’ derivation
  - POS shift without any significant semantic change
  - *pis-anie* `writing (gerund)’ < *pisać* `write’ (N-V)
  - *pisz-qcy* `writing (part.)’ < *pisać* (Part.-V)
  - *czerwon-ość* `redness’ < *czerwony* `red’ (N-Adj)
Derivational relations

- **Feature | State bearer** (N-Adj)
  - ślepi-ec `a blind man’ < ślep-y `blind’
  - starz-ec `an old man’ < star-y `old’
  Feature | State is an inverse relation

- **Femininity** (N-N)
  - feminine derivatives from masculine bases
  - pisar-ka `writeress’ < pisarz `writer’
  - kot-ka `female-cat’ < kot `cat’
Derivational relations

- **Markedness (N-N)**

**diminutives** `small`, `tiny`, `nice`

- `dom-ek` `small or nice house` $<$ `dom` `house`
- `książecz-ka` `tiny or nice book` $<$ `książ-ka`

**augmentatives** `big`, `large`, `awful`

- `ptasz-ysko` `big or awful bird` $<$ `ptak` `bird`
- `noch-al` `big or awful nose` $<$ `nos` `nose`

**young being** `offspring child of an animal`

- `koci-ę` `kitten` $<$ `kot` `cat`
Semantics roles & Role inclusion
Similar to EWN Role & Involved Relations

Roles (N < V)
- Agent: pływ-ak `swimmer’ < pływ-ać `swim’
- Instrument: pis-ak `marker’ < pis-ać `write’
  etc.

Role inclusion (V < N)
- Object: kartk-ować `to leaf through’
  < kartka `a sheet of paper’
- Instrument: pieprz-yć `to pepper’ < pieprz `pepper’
Derivationally motivated relations

- Derivational relations expanded into pure semantic relations
- Among nouns:
  - inhabitant
- Among verbs (*Aktionsarten*, `kinds of action`):
  - inchoativity, state, processuality, causation, iterativity, distributivity
Derivationally motivated relations

- Inhabitant (N-N) `a citizen of…’
  
  \{\text{Trojańczyk, Trojanin}\} `Troy citizen’ - \{\text{Troja, Ilion}\} `Troy’

  \text{Troj-ańczyk, Troj-anin < Troja}

  \text{Troj-ańczyk, Troj-anin < Ilion} (but they pass tests)

- Inchoativity (V-V) `the beginning of a situation’
  
  \{\text{wstać, podnieś się, powstać}\} `get up’ - \{\text{stać}\} `stand’

  \text{w-stać < stać}

  \text{podnieść się, powstać < stać} (but they pass tests)
**plWordNet content**

<table>
<thead>
<tr>
<th></th>
<th>synsets</th>
<th>lemmas</th>
<th>LUs</th>
<th>avs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GermaNet</td>
<td>101,371</td>
<td>119,231</td>
<td>131,814</td>
<td>–</td>
</tr>
<tr>
<td>Princeton WordNet 3.1</td>
<td>117,659</td>
<td>155,593</td>
<td>206,978</td>
<td>1.74</td>
</tr>
<tr>
<td>enWordNet 1.0</td>
<td>125,500</td>
<td>165,712</td>
<td>218,611</td>
<td>1.74</td>
</tr>
<tr>
<td>plWordNet 3.0 emo</td>
<td>197,721</td>
<td>179,125</td>
<td>260,214</td>
<td>1.32</td>
</tr>
</tbody>
</table>

- LUs – lexical units (= senses)
- avs – average synset size
plWordNet content

- >40 different relation types (>100 when counting subtypes)
  - including many relations linking lexical units of different PoS
- Semantic domains (*lexicographer files* from WordNet)
- Semantic verb classes – originating from Vendler’s verb classification
- Stylistic labels (11 in total)

<table>
<thead>
<tr>
<th>Description layer</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>lexico-semantic relations</td>
<td>&gt;700K</td>
</tr>
<tr>
<td>glosses</td>
<td>&gt;100K</td>
</tr>
<tr>
<td>usage examples</td>
<td>83K</td>
</tr>
<tr>
<td>links to Wikipedia</td>
<td>55K</td>
</tr>
<tr>
<td>sentiment annotation</td>
<td>31K</td>
</tr>
</tbody>
</table>
plWordNet emotional annotation

- Basic emotions
  - joy, trust, fear, surprise, sadness, disgust, anger, anticipation (Plutchik, 1980)

- Fundamental human values (Puzynina, 1992)
  - (positive) utility, another's good, truth, knowledge, beauty, happiness
  - (all negative) futility, harm, ignorance, error, ugliness, misfortune

- Sentiment polarity
  - +strong +weak, neutral, -weak, -strong

- Details: cf (Zaśko-Zielińska et. al, 2015) from RANLP’2015
plWordNet emotional annotation

<table>
<thead>
<tr>
<th>PoS</th>
<th>#</th>
<th>-s</th>
<th>-w</th>
<th>n</th>
<th>+w</th>
<th>+s</th>
<th>amb</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>19,625</td>
<td>11.29</td>
<td>8.78</td>
<td>69.06</td>
<td>3.24</td>
<td>2.88</td>
<td>4.74</td>
</tr>
<tr>
<td>Adj</td>
<td>11,573</td>
<td>9.89</td>
<td>11.22</td>
<td>58.85</td>
<td>9.21</td>
<td>5.60</td>
<td>5.24</td>
</tr>
<tr>
<td>Both</td>
<td>31,198</td>
<td>10.77</td>
<td>9.69</td>
<td>65.27</td>
<td>5.46</td>
<td>3.89</td>
<td>4.92</td>
</tr>
</tbody>
</table>

- **Annotation coverage**
  - 27% of adjectival LUs
  - 12% of noun LUs
    - the domains most likely to include LUs with non-neutral sentiment polarity
- A scale several times larger than, e.g. SentiWordNet
Network volume and density

<table>
<thead>
<tr>
<th>WordNet 3.1</th>
<th>verbs N</th>
<th>verbs ρ</th>
<th>nouns N</th>
<th>nouns ρ</th>
<th>adverbs N</th>
<th>adverbs ρ</th>
<th>adjectives N</th>
<th>adjectives ρ</th>
<th>all N</th>
<th>all ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>LU relations</td>
<td>24,840</td>
<td>0.99</td>
<td>44,185</td>
<td>0.28</td>
<td>720</td>
<td>0.13</td>
<td>21,636</td>
<td>0.72</td>
<td>91,381</td>
<td>0.42</td>
</tr>
<tr>
<td>synset relations</td>
<td>16,827</td>
<td>1.22</td>
<td>145,338</td>
<td>1.62</td>
<td>109</td>
<td>0.03</td>
<td>23,491</td>
<td>1.29</td>
<td>185,765</td>
<td>1.48</td>
</tr>
<tr>
<td>all relation types</td>
<td>80,280</td>
<td>3.20</td>
<td>492,457</td>
<td>3.12</td>
<td>1,015</td>
<td>0.18</td>
<td>86,221</td>
<td>2.87</td>
<td>659,973</td>
<td>3.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>plWordNet 3.0</th>
<th>verbs N</th>
<th>verbs ρ</th>
<th>nouns N</th>
<th>nouns ρ</th>
<th>adverbs N</th>
<th>adverbs ρ</th>
<th>adjectives N</th>
<th>adjectives ρ</th>
<th>all N</th>
<th>all ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>LU relations</td>
<td>48,744</td>
<td>1.50</td>
<td>98,376</td>
<td>0.58</td>
<td>12,542</td>
<td>1.14</td>
<td>48,894</td>
<td>1.02</td>
<td>208,556</td>
<td>0.80</td>
</tr>
<tr>
<td>synset relations</td>
<td>36,616</td>
<td>1.66</td>
<td>219,266</td>
<td>1.75</td>
<td>19,716</td>
<td>2.18</td>
<td>48,258</td>
<td>1.17</td>
<td>323,856</td>
<td>1.64</td>
</tr>
<tr>
<td>all relation types</td>
<td>127,065</td>
<td>3.92</td>
<td>494,893</td>
<td>2.94</td>
<td>43,551</td>
<td>3.94</td>
<td>118,574</td>
<td>2.47</td>
<td>784,083</td>
<td>3.02</td>
</tr>
</tbody>
</table>

ρ is the relation density measured either for LUs, or synsets, or for all relation types
plWordNet Corpus vocabulary

- plWordNet Corpus – merged corpora
- vocabulary in lemmas
- v. 7.0 – plWordNet corpus of ~2G words
- v. 10.0 – plWordNet corpus of ~4G words

Frequency bin

<table>
<thead>
<tr>
<th>[in thousands]</th>
<th>v. 7.0</th>
<th>v. 10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-9</td>
<td>1217</td>
<td>412</td>
</tr>
<tr>
<td>10-99</td>
<td>966</td>
<td></td>
</tr>
<tr>
<td>100-999</td>
<td>105</td>
<td>186</td>
</tr>
<tr>
<td>1000-9999</td>
<td>31</td>
<td>49</td>
</tr>
<tr>
<td>10000&lt;</td>
<td>10</td>
<td>17</td>
</tr>
</tbody>
</table>
plWordNet coverage

- Percentage of PLWNC vocabulary covered by plWN 3.0
- v. 7.0 – merged corpus of ~2G words
- v. 10.0 – merged corpus of ~4G words
plWordNet as a small world

- Compared graphs
  - plWN1-plWN3 – different version of plWordNet
  - PWN – WordNet 3.1
  - WN-plWN – PWN and plWN 3.0 combined by mapping

- Average path length
  - expected: short average path length
  - average path length for the random graph: 11
plWordNet as a small world

- Global clustering coefficient
  - higher - denser
plWordNet as a small world

- Connectivity
  - how often a path can be established between two synsets randomly chosen
plWordNet mapping

- Manual mapping strategy (Rudnicka et. al, COLING 2012)
  - briefly, a comparison of the two relation structures in order to find the corresponding nodes of synset graph structures
  - 8 types of interlingual relations for synsets
  - plWordNet 3.0 ➔ Princeton WordNet 3.1
- Laborious process, but resulted in a remarkable opportunity to run a comparative analysis!
plWordNet mapping

- Recognise the sense of a source synset by:
  - its position in the network structure,
  - existing relations, commentaries (glosses),
  - comparison to other synsets containing the given lemma
- Search for candidates for a target synset:
  - intuitions, automatic prompting and dictionaries
- Verify candidates:
  - by comparing hypernymy and hyponymy structures
  - by exploring existing inter-lingual relations;
  - by comparing definitions, commentaries; dictionaries
- Link the source synset with the target synset
plWordNet mapping in WordnetLoom system
### Interlingual relations: plWordNet 3.0 ➔ Princeton WordNet 3.1

<table>
<thead>
<tr>
<th>I-relation</th>
<th>Noun</th>
<th>Adjective</th>
<th>Adverb</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-Synonymy</td>
<td>36,367</td>
<td>4,077</td>
<td>448</td>
<td>40,892</td>
</tr>
<tr>
<td>I-Hyponymy</td>
<td>74,394</td>
<td>29,216</td>
<td>781</td>
<td>104,391</td>
</tr>
<tr>
<td>I-Hypernymy</td>
<td>4,121</td>
<td>167</td>
<td>51</td>
<td>4,339</td>
</tr>
<tr>
<td>I-Meronymy</td>
<td>6,982</td>
<td>-</td>
<td>-</td>
<td>6,982</td>
</tr>
<tr>
<td>I-Holonymy</td>
<td>3,471</td>
<td>-</td>
<td>-</td>
<td>3,471</td>
</tr>
<tr>
<td>I-Partial synonymy</td>
<td>4,339</td>
<td>1,544</td>
<td>4</td>
<td>5,887</td>
</tr>
<tr>
<td>I-Inter-register synonymy</td>
<td>1,672</td>
<td>54</td>
<td>22</td>
<td>1,748</td>
</tr>
<tr>
<td>I-Cross-categorical synonymy</td>
<td>-</td>
<td>19,286</td>
<td>-</td>
<td>19,286</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>131,346</td>
<td>54,344</td>
<td>1,306</td>
<td>186,996</td>
</tr>
</tbody>
</table>
plWordNet 3.0 vs WordNet 3.1

- Synset granularity
  - plWordNet synsets are smaller
  - distinct synsets for feminine, masculine and neuter forms, diminutive, augmentative and stylistically marked forms,
  - mixed WordNet synsets grouping together marked and unmarked forms,
  - singular and plural, mass and count,
- Different multiple hyponymy: always `and` in plWN, `or/and` in WN
- Understanding of relations: meronymy vs hyponymy
- Significantly different structures for adjectives and adverbs
  - vertical structure in plWN with many relations
  - horizontal dumbbell model in WN
- Errors in WN and lexical gaps between languages
enWordNet 1.0

- Motivations
  - I-Synonymy is more useful than I-Hyponymy, …
  - Goal: to exploit the existing I-Hyponymy links to extend WordNet’s coverage

- Result
  - enWordNet 1.0 – an extended version of WordNet 3.1
  - focus given to translations of the plWordNet leaf synsets:
    - equivalents whose lemmas were not present in WordNet
    - no equivalents;
    - equivalents whose lemmas were already present in WordNet
  - verification with English corpora and dictionaries
  - other missing lexical units from hypernymy branches added
  - 7,841 new English synsets, 11,633 LUs, 10,119 lemmas
A complex system of lexico-semantic resources

- **Walenty** *(valence lexicon)*  
  > 15 000 lemmas

- MWE lexicon  
  (54 000)

- Proper Names  
  (2.4 mln)

- plWordNet 3.0 emo

- enWordNet 1.0  
  (ext.: WordNet 3.1 plus 10k senses)

- Ontology SUMO  
  (upper and intermediate level)

(Hajnicz, 2014)  
ICS PAS Warsaw

Lexico-syntactic structures

Samsung R&D Institute
Invit. Lecture
2017-01-17
CLARIN-PL
Structured description of MWE

```xml
<mwegroup type='fix'
    name='SubstSubstGenGapFix'
    class='subst'>
  <condition>
    and(
      inter(base[0],$s:S),
      inter(class[0],subst),
      rlook(1,end,$G, and(
        inter(base[$G],$s:SG)
        inter(class[$G],subst),
        inter(cas[$G],gen),
      )),
    only(1,$G-1,$N, or(
      inter(class[$N],adv,qub),
      and(
        or(
          inter(class[$N],
            adj,ppas,pact),
          inter(base[$N],"mój") ),
          agrpp($G,$N,nmb,gnd,cas)
        )
      )))
    setvar($Pos1, 0),
    setvar($Pos2, $G)
  </condition>

<instances>
  <MWE base='kobieta życia'>
    <head>
      in(class[0],
        subst,ger,depr)</head>
    <var name='S'>kobieta</var>
    <var name='SG'>życie</var>
  </MWE>
</instances>
</mwegroup>
```
plWordNet 3.0 emo applications

- Wide coverage inspires a lot of applications
- plWordNet is a pivotal element of a system of language and knowledge resources
- An anchor to Linked Open Data via mapping to WordNet
- Monolingual and bilingual dictionary
  - Web-based: [http://plwordnet.pwr.edu.pl](http://plwordnet.pwr.edu.pl)
  - Android application
  - WordnetLoom for visual, graph-based browsing
  - included in a very large and popular Polish multilingual Web dictionary Ling
- WordTies (Pedersen et al., 2012), Open Multilingual WordNet (Bond and Foster, 2013)
plWordNet 3.0 emo applications

- Numerous research applications, for instance
  - Classification of gestures based on the verb categorisation in plWordNet (Lis and Navarretta, 2014)
  - Referred to in the resource for textual entailment (Przepiórkowski, 2015)
  - Language correction
  - Relation extraction
  - Text classification
  - Open Domain Question Answering
  - A quasi-ontology in document structure recognition
- An exceptional case is the practical use of plWordNet during the medical treatment of aphasia
plWordNet 3.0 emo applications

- A large number of declared applications:
- Education (at different levels) including Polish language teaching,
- Building dictionaries, extraction of synonyms and semantically related words, detection of loanwords,
- Cross-linguistic study on phonesthemes, classification of metaphorical expressions,
- Corpus studies, grammar development, comparative and contrastive studies,
- Language recognition, parsing disambiguation, semantic analysis of text, document similarity measures, semantic indexing of documents, semantic information retrieval,
- Recommendation systems, construction of chatbots and dialogue systems,
- Plagiarism detection,
- Translation evaluation, data visualisation, research on complex networks and ontologies, …
Bibliography


Thank you very much for your attention!

www.clarin-pl.eu

Supported by the Polish Ministry of Science and Higher Education [CLARIN-PL]