Annotation of metaphorical expressions in the Basic Corpus of Polish Metaphors

Elżbieta Hajnicz
E.hajnicz@ipipan.waw.pl

Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, Warsaw

Basic corpus of Polish metaphors
• 700 samples of the Polish Coreference Corpus (PCC) (206,031 tokens – 286 tokens per sample on average)
  PCC is a corpus of randomly selected samples of NKJP300M and the manually annotated corpus NKJP1M.
• 2000 samples of a fragment of NKJP1M considered in the Składnica treebank (144,087 tokens – 68 tokens per sample on average).
  The selection preserves balance rules for NKJP.
• NKJP (National Corpus of Polish) has two subsystems: the balanced corpus NKJP300M and the manually annotated corpus NKJP1M.
• Składnica is a treebank randomly selected from NKJP1M.

Identification of a metaphorical expression

The procedure
The procedure is based on the lexico-semantic annotation of the corpus by means of PlWorski lexical units (LU's).
• Reading the whole text (sample) in order to establish its general meaning and subject.
• Determining, whether another, more basic meaning of each phrase exists, adequate in different contexts (e.g. u uso – of a department, state etc. vs. body part).
• Staring at them, their common and distinct properties and checking, whether the new meaning can be interpreted through the prism of the old one, distinctly connected to it.

Furthermore:
• If the meaning adequate in context is not distinguished, but the corresponding “basic” meaning is used in a way that goes far beyond its normal usage, we treat it as metaphorical.

Example 1
Do Polski kapitalizm wjechał czolgiem : kompletne nas staranowali.
‘To Poland came capitalism on a tank and smashed us completely.’

There is no separate meaning for 'Capitalism drove into Poland on a tank and smashed us completely.'

The structure of a metaphorical expression

• vehicle – a part of an utterance used metaphorically, representing a source domain e.g. drive on a tank.
  - each vehicle phrase has its head (here mazu);
• topic – a part that refers to reality, that represents a target domain e.g. capitalism.
• both vehicle and topic need not be sequential,
• a vehicle should be included in the analysed utterance, whereas a topic could be even completely outside it (usually in the case of ellipsis);
• both vehicle and topic determines the scope of an metaphoric expression; phrases that can occur both in a literal or metaphorical context (e.g. Polano) are outside that scope.

The figurativeness of an expression emerges from the confrontation of its vehicle and its topic.

Classification of metaphorical expressions

• Text_form – a form the vehicle of a metaphor takes in a text:
  - word – the vehicle of a metaphor is composed of a single word, e.g. ‘ram’ in example 1;
  - phrase – the vehicle of a metaphor is a phrase, e.g. ‘drive on a tank’ in example 1;
  - text – if a metaphor has a narrative form.
        - Structure a conceptual structure of a metaphor:
        - simple – involves a single vehicle and a single (or none) topic, e.g. slodki (vehicle) czenika (topic) ‘sweet revenge’;
        - relational – differs from simple metaphors in that its vehicle relates two or more topics, e.g. the vehicle ‘built’ relates ‘organisms’ and ‘proteins’ in example 3;
        - elaborated – contains additional terms from a source domain emphasizing and expanding the metaphorical expression, e.g. proterotermiowny pasztec, lit. ‘expired pate’, an old, ugly woman in example 2;
        - mixed – a target domain is described by means of several source domains, cf. example 4;
        - layered – there are two topics from a different domain that one is applied to the other.
• Characteristics – specification of a typical source domain for the X is Y model:
  - personification – describing abstracts, objects and animals as people (solution providers);
  - animisation – describing abstracts, objects and sometimes people as animals (truth that bares teeth);
  - reflaction – describing abstracts, animals and sometimes people as objects (built a solution);
  - depersonification – describing people as objects or animals in a way depersonalising them (an expired pate).
• Contextuality – showing whether and to what extent the figurativeness of an utterance depends on its context:
  - contextual – an utterance can be interpreted literally and a topic of metaphor is located outside the utterance;
  - self-containing – an utterance can be completely, metaphorically interpreted regardless of the context.

All metaphorical expressions presented above are self-containing.

• Conventionality
  The conventionality of a metaphorical expression means that it is established in culture and language, and it is distinguished and represented in dictionaries.
  - standard – considered in PlWorski, our primary source of lexico-semantic info (e.g. pate’ used for an ugly woman in example 2);
  - external – considered in other dictionaries;
  - novel – outside dictionaries, usually used spontaneously (e.g. expired used in the context of palte in example 2).

Results of the annotation

Scope of the annotation
• 343 samples containing 98,336 tokens of PCC subcorpus,
  • each sample annotated by two annotators (of > 10),
  • the procedure processed by means of WebAnno tool.

Results of the annotation
• total number of metaphorical expression is 8547,
  • only 5,5% of tokens are considered metaphoric,
  • their average number in a sample is 16,
  • only 2410 (28%) words (or phrases) has been chosen as metaphorical by both Annotators.

<table>
<thead>
<tr>
<th>topics</th>
<th>equals</th>
<th>equals</th>
<th>overlaps</th>
<th>number</th>
<th>part</th>
<th>part</th>
<th>part</th>
<th>part</th>
<th>part</th>
</tr>
</thead>
<tbody>
<tr>
<td>4373</td>
<td>992</td>
<td>0.23</td>
<td>1397</td>
<td>0.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Inter-annotator agreement

The distribution of annotators’ choices for various ME features

<table>
<thead>
<tr>
<th>Feature</th>
<th>names of classes and their cardinality</th>
</tr>
</thead>
<tbody>
<tr>
<td>conventionality</td>
<td>2, external: 292, included: 214, novel: 552, standard: 3757</td>
</tr>
<tr>
<td>characteristics</td>
<td>1449, animisation: 366, depersonification: 58, personification: 635, reflaction: 2309</td>
</tr>
<tr>
<td>contextuality</td>
<td>2, contextual: 333, self-containing: 4482</td>
</tr>
<tr>
<td>text_form</td>
<td>phrase: 569, text: 23, word: 4226</td>
</tr>
<tr>
<td>The Smith’s t statistics is used ( t = \frac{P(A) - P(B)}{E} ) where ( P(A) ) is inter-annotator score and ( P(B) ) is random score.</td>
<td></td>
</tr>
<tr>
<td>The results of inter-annotator agreement</td>
<td>feature</td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>both</td>
<td>1551</td>
</tr>
<tr>
<td>P(A)</td>
<td>0.66</td>
</tr>
<tr>
<td>P(E)</td>
<td>0.56</td>
</tr>
<tr>
<td>( t )</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Strength of domination of dominating classes

<table>
<thead>
<tr>
<th>class</th>
<th>simple</th>
<th>standard</th>
<th>refl. self_cont</th>
<th>word</th>
</tr>
</thead>
<tbody>
<tr>
<td>number</td>
<td>3535</td>
<td>2517</td>
<td>2809</td>
<td>4492</td>
</tr>
<tr>
<td>part</td>
<td>0.73</td>
<td>0.78</td>
<td>0.48</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Example 2
Mieux opłacaloby mu się zakochać w przeterminowanych pasztecie z
perhaps pate’ed be as well mark fail in love as in expired pate’ with
own pate’.

Perhaps it would be profitable for him to fall in love with an old, ugly woman with her own palace.

Example 3
Wszystkie organizmy […] zdobawne są z białek.
All now, organisms now built from, are of protein.

All organisms are made of proteins.

Example 4
Zołę pate’ed moje swiata wczesnytya zęby okrutna prawda.
From behind ruins see my own, world see, bare teeth cruel truth.

From behind the ruins of my world the cruel truth was baring its teeth.'