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)
  (200,031 tokens – 286 tokens per sample on average)
  PCC is compose of randomly selected samples of NKJP300M (balanced subcorpus)

- 2000 samples of a fragment of NKJP1M considered in the *Składnica* treebank
  (144087 tokens – 68 tokens per sample on average);
- The selection preserves balance rules for NKJP;

- NKJP (National Corpus of Polish) have two subcorpora:
  the balanced subcorpus NKJP300M and the manually annotated subcorpus 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 PlWordnet lexical units (LUs).

- 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. HEAD – *of a department, state etc.* vs. *body part*).
- Stating 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** **czołgiem** i kompletnie nas **staranował**.
To Poland.GEN capitalism.NOM drive.PAST into tank.INST and completely we.ACC ram.PAST
'Capitalism drove into Poland on a tank and smashed us completely.'

There is no separate meaning for DRIVE INTO (in particular, *drive on a tank*) or for RAM, but *capitalism* is not a living being or an object that can drive or ram anything.

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 DRIVE);
- ***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 ***topics*** determine the scope of an metaphoric expression; phrases that can occur both in a literal or metaphorical context (e.g. POLAND) 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. ***słodka*** (vehicle) *zemsta* (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 emphasising and expanding the metaphorical expression, e.g. *przeterminowany pasztet*, lit. 'expired pâté', 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* ***provides***);
  - `animisation` – describing abstracts, objects and sometimes people as animals (*truth* that ***bares teeth***;
  - `reification` – 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 pâté***.
- `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_contained` – an utterance can be completely, metaphorically interpreted regardless of the context.

  All metaphoric expressions presented above are self_contained.

- `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 PlWordnet, our primary source of lexico-semantic info (e.g. *pâté* 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 *pâté* 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.

topics' number	equals number	equals part	overlaps number	overlaps part
4373	992	0.23	1397	0.32

Inter-annotator agreement

The distribution of annotators' choices for various ME features

feature	names of classes and their cardinality
structure	elaborated: 461, layered: 29, mixed: 285, relational: 500, simple: 3535, unknown: 7
conventionality	*: 2, external: 292, included: 214, novel: 552, standard: 3757
characteristics	*: 1449, animisation: 366, depersonification: 58, personification: 635, reification: 2309
contextuality	*: 2, contextual: 333, self_contained: 4482
text_form	phrase: 569, text: 23, word: 4226

- The Scott's π statistics is used $\pi = \frac{P(A) - P(E)}{1 - P(E)}$,
  where $P(A)$ is inter-annotator score and $P(E)$ is random score.

The results of inter-annotator agreement

feature	struct.	convent.	charact.	context.	text_form
both	1591	1807	1231	2206	2068
$P(A)$	0.66	0.75	0.51	0.92	0.86
$P(E)$	0.56	0.63	0.34	0.87	0.78
π	0.23	0.33	0.26	0.35	0.35

Strength of domination of dominating classes

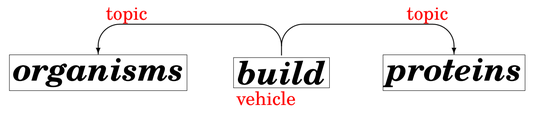
class	simple	standard	reific.	self_cont.	word
number	3535	3757	2309	4482	4226
part	0.73	0.78	0.48	0.93	0.88

Example 2

Może opłacałoby mu się zakochać w **przeterminowanym** **pasztecie** z własnym pałacem.
perhaps pay.COND he.DAT refl. mark fall in love.INF in expired.LOC pâté.LOC with own.INST palace.INST
'Perhaps it would be profitable for him to fall in love with an old, ugly woman with her own palace.'

Example 3

Wszystkie organizmy [...] ***zbudowane*** *są z białek*.
All.NOM organisms.NOM [...] built.NOM.PL are of proteins.GEN
'All organisms are made of proteins.'



Example 4

Zza ruin mojego świata ***szczerzyła*** ***zęby*** *okrutna prawda*.
from behind ruins.GEN my.GEN world.GEN bare.IMPERF.PAST teeth cruel.NOM truth.NOM
'From behind the ruins of my world the cruel truth was baring its teeth.'