

****Standardization Codebook****

Rules for Lemmatizing Metaphorical Expressions

****Note:** ****These are general standardization rules that may require additional, domain-specific guidelines depending on the dataset (e.g., a corpus of animal metaphors may need extra rules for breed names, fictional creatures, etc.). The annotation aims to provide a general picture of the metaphorical expression, which may omit certain nuances that affect interpretation (e.g., a specific dog breed carries different connotations than the general term *dog*, but the standardized form captures the broader category).**

**Part 1: General Standardization Rules**

**A. Grammatical Form**

- ****Nouns:** ****Standardize to the **nominative singular** form.**
- ***nosorożce (plural)* → ***nosorożec*****
- ****Verbs:** ****Standardize to the **infinitive**, preserving prefixation and aspect. If the dictionary/resource does not recognize the prefixed or aspectual form, remove the aspect marker.**
- ***wylatywać* → ***wylatywać*** (prefix and aspect preserved)**
- **If the dictionary does not recognize the form, fall back to the base verb.**
- ****Relational adjectives used as metaphors:** ****Convert to the corresponding **noun**.**
- ***szczurzy (adjective, ***'rat-like'***)* → ***szczur (*****'*****rat*****')*****

**B. Diminutives**

- ****Affectionate/stylistic diminutives:** ****Standardize to the base (non-diminutive) form.**
- ***misiu / miś (affectionate)* → ***niedźwiedź (*****'*****bear*****')*****
- ***motylek* → ***motyl (*****'*****butterfly*****')*****

***Note:** Remember the diminutive form when interpreting the metaphor; the affective dimension may be relevant for analysis.*

- ****Diminutives denoting young animals:** ****Do **not** standardize to the adult form. The researcher must determine from context whether a diminutive refers to an affectionate form of the adult or to a young/different entity.**
- ***kotek* (meaning 'kitten') ≠ *kot* ('cat'). Remains ***kotek***.**
- ****Diminutives with distinct meanings (e.g., food names):** ****Do **not** standardize to the animal/base form. These are separate lexical items.**
- ***gołąbek* (a stuffed cabbage dish) ≠ *gołąb* ('pigeon'). Remains ***gołąbek***.**
- ***kopytka* (a type of dumpling) ≠ *kopyto* ('hoof'). Remains ***kopytka***.**

**C. Gender**

- ****Preserve grammatical/biological gender. **Do not neutralize gendered animal**

terms to a default form.

- *łania* (female deer) ≠ *jeleń* (male deer). Remains ***łania***.

Rationale: Gendered forms may be relevant for gender-related analysis.

D. Breed, Variety, and Subcategory Names

- Standardize **specific breed/variety names to the general category** they represent.

- *owczarek niemiecki (**'German Shepherd**')* → ***pies
(*****dog*****)

Note: The specific breed may carry distinct connotations (e.g., a German Shepherd vs. a Poodle), but the standardized form captures the broader category for systematic annotation.

E. Trademarks and Brand Names (Eponyms)

- Standardize **brand names and eponyms to the generic product** they represent.

- *Pampers* → ***pielucha (*****diaper*****)

- *adidas* (colloquial eponym) → ***buty sportowe (*****sports shoes*****)

F. Fictional Characters and Proper Names

- **Fictional characters: Standardize to the **most salient feature** of the character that is relevant to the metaphorical use.

- *Król Lew (**'The Lion King**')* → ***lew
(*****lion*****)

- *Świnka Peppa (**'Peppa Pig**')* → ***świnia
(*****pig*****)

- *Kosogłós (**'Mockingjay**')* → ***symbol
(*****symbol*****)

– the metaphor draws on its role as a symbol of rebellion, not on the bird itself

- If the character cannot be reduced to a single salient feature, **retain the original name** and explain in a comment.

G. Multi-Word Phrases (Compounds and Collocations)

- **If the multi-word phrase exists as a dictionary entry, retain the full phrase and base the annotation on the dictionary meaning.

- *pies ogrodnika* ('dog in the manger') – has a dictionary entry referring to a person; retain as ***pies ogrodnika***.

- If the phrase is not in the dictionary, **standardize to the most salient feature** (typically the head noun; see Fuoli 2022 on premodifiers).

- *zmokła kura* ('a drenched hen') – not a fixed dictionary entry → ***kura*** ('hen')

- *bezpański pies* ('stray dog') – not a fixed dictionary entry → ***pies*** ('dog')

H. Definitions: Everyday vs. Scientific

- When selecting the definition for the standardized form, always prefer the ****common, everyday understanding**** of the word over a scientific or taxonomic definition.

- ***ogórek***: use the common meaning 'cucumber (vegetable)', not the botanical definition ('a plant of the Cucurbitaceae family...').

***Rationale:** Metaphors operate on conventional, embodied knowledge (cf. embodied cognition in cognitive linguistics). The basic meaning (BM) should reflect sensory and everyday experience.*

****I. Language-Specific Awareness****

- When working with bilingual or cross-linguistic data, ensure that definitions and standardized forms correspond to the ****correct language****. Do not conflate terms across languages.

- Polish ***puzzle*** vs. English ***jigsaw puzzle*** – these may differ in scope and usage.

This is especially important in automated pipelines that draw definitions from multilingual resources.

****Part 2: Additional Notes for plWordNet (Słowosieć) Users****

The following notes apply specifically to researchers using the Polish WordNet (Słowosieć) for retrieving definitions and hypernym chains.

****A. Plural-Only Nouns****

- ***śmieci*** ('rubbish') – Słowosieć does not recognize the singular ***śmieć*** as a noun; it parses it as a verb (meaning 'to dare'). In such cases, use the plural form as it appears in Słowosieć and note the discrepancy.

****B. Removing the Standardized Form from Hypernym Chains****

- When constructing a hypernym chain from Słowosieć, ****remove the standardized form itself from the chain**** to avoid redundancy. The chain should begin with the first hypernym above the standardized form.

****C. Basic Meaning (BM) and Hypernym Gaps****

- Some definitions in Słowosieć that qualify as the Basic Meaning (BM) – i.e., the most fundamental, sensory-grounded sense of a word (in line with embodied cognition) – may use formulations that do not appear in the hypernym chain. This means that certain conceptual information from the BM definition may not be visible in the annotation based on hypernyms alone. ****Researchers should be aware of this gap**** and consult BM definitions alongside hypernym chains.

****D. Missing Definitions in Słowosieć****

- When a word has no definition in Słowosieć, check the ****cross-lingual synonym (interlingual link)**** for an English equivalent and use the English definition.

- If neither the Polish nor the English definition is available in Słowosieć, consult an ****external dictionary****: PWN (Słownik Języka Polskiego) for Polish, Merriam-Webster for English. The specific dictionary choice may vary depending on the researcher's needs.

****E. Colloquial / Informal Entries****

- Słowosieć may contain colloquial or informal entries. Treat them according to the same standardization rules, but be aware that their hypernym chains or definitions may be less complete.

F. Automated Data Collection

- When using automated tools to retrieve definitions from Słowność, ensure that the tool pulls definitions in the **correct language** matching the dataset. Cross-linguistic resources may return English glosses for Polish queries or vice versa.