Semantic Annotation for Ukrainian: Categorization Scheme, Principles, and Tools

Vasyl Starko[0000-0002-2530-2107]

Ukrainian Catholic University, 2a Kozelnytska Street,
Lviv, 79026, Ukraine

v.starko@ucu.edu.ua

Abstract. A semantic tagset for the semantic annotation of Ukrainian-language texts is presented, and the use of the taxonomic approach is substantiated. The categorization scheme implemented in the tagset takes into account the cognitive-linguistic perspective on categorization, specifically the basic level of categorization. Semantic tags are to be assigned to lemmas in the existing Large Electronic Dictionary of Ukrainian (VESUM) yielding a semantic lexicon that will be used by the TagText tagger (both tools developed by the r2u team) to add semantic annotation to the GRAC corpus. Used in conjunction with POS tags, semantic tags will serve as a powerful tool for the linguistic exploration of corpus data and for solving NLP tasks involving Ukrainian.

Keywords: semantic annotation, semantic lexicon, semantic tagset, categorization, corpus, r2u, GRAC, VESUM, TagText.

1 Introduction

Semantic annotation is an important type of annotation of natural-language texts. There are several different approaches to annotating texts with semantic labels: based on WordNet [14], FrameNet [4], hierarchical classification [12], and taxonomic classification [6], [7].

The ideographic, or hierarchical, approach to semantic tagging proceeds in the top-down fashion, from the most general notions to the most specific terms. While this system has its merits, it is not without its flaws. There is no one universal hierarchical classification scheme as is evidenced by discrepancies in the thesauri of different languages. The top layers of any such system are quite abstract and beyond the intuitive understanding of most users, while some categories appear to be not quite coherent, for example A10+ Open; Finding; Showing [12]. These systems involve a fine-grained semantic classification of vocabulary with numerous semantic features that transcend POS boundaries. It has been argued [7] that such a purely semantic approach is not well-suited for corpora for reasons of cumbersomeness, excessive ambiguity, and counterintuitive groupings. For example, the feature ‘motion’ would be assigned not only to verbs and deverbal nouns (*to run*, *running*) but also to adjectives (*quick*), object names (*feet*, *wheels*), and so on. Some of these words, such as *road* and *smoke*, would be highly surprising to an average corpus user.

Furthermore, hierarchical classification is dichotomous—a word may be assigned to one category only, while taxonomic classification allows for flexible attribution of several semantic features to a word. Taxonomic classification aligns well with cognitive linguistic research showing that speakers perceive and classify words relying on integral gestalts (a simultaneous, complex totality of their various features) rather than on discrete features [7].

For the Ukrainian corpus [1] developed by Nataliia Darchuk and her colleagues, a hybrid approach has been proposed involving a combination of taxonomic semantic classification for non-scientific texts and the construction of ontologies for various scientific domains to be applied to scientific texts [2]. However, the full specification of the semantic tagset has yet to be made publicly available by this group. The web interface to their corpus does not allow semantic searches so far, and semantically tagged words can only be seen in frequency wordlists generated on their website.

In the absence of a large-scale resource, such as WordNet, for Ukrainian, which would allow the attribution of specific sense to words in context, and given the advantages of the taxonomic system, it is reasonable to pursue the latter route. Thus, we propose a lexicon-based approach to semantic annotation for Ukrainian using taxonomic semantic tags. The lexicon will be publicly available together with a tagger, enabling researchers to semantically tag any Ukrainian-language texts.

While the proposed semantic annotation for Ukrainian has a predominantly practical focus, its scientific value lies in the domain of semantic description and formalization. It is, essentially, the first step in the direction of a formalized semantic description of Ukrainian.

2 Purpose and Principles of Semantic Annotation

The present paper is focused on developing a semantic tagset, while the overarching goal of the broader endeavor is to build a semantic lexicon for Ukrainian and add a semantic annotation layer to the General Regionally Annotated Corpus of Ukrainian (GRAC) [5]. With semantic tags in place, the user will be able to apply them separately or in combination with POS tags in constructing search queries. Researchers will be able to investigate, among other things, the linguistic behavior of semantically motivated classes of words: their combinatorial properties, patterns of government, usage and variation patterns, etc. An analysis of semantic phenomena, such as meaning shifts, polysemy, and word sense disambiguation, will have a firmer foundation. Semantic tags will also be useful for applications outside of corpus linguistics, for example, named entity recognition and information extraction.

GRAC has a wide audience—from school students to foreign students of Ukrainian to the NLP community and to seasoned linguists. Thus, the semantic annotation scheme for Ukrainian must be accessible and transparent to users with varying degrees of linguistic expertise. Moreover, it needs to be as unambiguous as possible. With proper explanation, each semantic tag should leave no doubt in the user's mind as to its content. This will require taking into account the so-called linguistic, or naïve, worldview possessed by the native speakers of the language.

The approach implemented in the Russian National Corpus [6], [7] served as a point of departure for the system of semantic annotation presented here, albeit there are significant differences in a number of aspects. Both systems employ the faceted, rather than hierarchical, approach to classification, allowing for one word to be placed in several taxonomic classes (assigned different semantic tags) at the same time, rather than focusing narrowly on those that are relevant for genus–species relations.

Taxonomy-based semantic annotation must meet three requirements: it must be generally understandable, linguistically meaningful [6], and cognitively motivated. The names of semantic tags need to be transparent and intuitive. Alphanumeric designations (such as those used in the USAS project, e.g. **S1.2.1**, which means ‘approachability and friendliness’) necessitate intimate knowledge of the classification scheme and, at least initially, frequent lookup. In contrast, the names of semantic tags for Ukrainian will be abbreviations of common English words (see examples below).

The linguistic relevance of semantic classes means that words belonging to the same semantic class should exhibit commonality of linguistic behavior by virtue of such membership. Semantics is viewed as having surface manifestations and, more precisely, as a motivating factor for surface lexical behavior. This view is widely accepted in cognitive linguistics and has been convincingly advocated, among others, by the prominent semantics scholar Anna Wierzbicka [13]. In selecting semantic features and establishing the overall classification scheme for such multifaceted feature attribution, it is important to consider what we know about human natural-language categorization and its principles [3], [9].

As far as semantic features labeled by tags are concerned, they should be independent, basic, forming sufficiently large classes, generating a minimum of noise, and yielding optimal results when searching for constructions (using complex corpus search queries) [7: 226]. Furthermore, the set of semantic tags needs to be constructed in such a way as to yield a linguistically relevant classification for the entire Ukrainian vocabulary and reflect, as much as possible, the semantic groupings involved in regular patterns of lexico-grammatical interactions in Ukrainian texts. In other words, semantic classes should be maximally homogeneous in terms of their linguistic behavior, while at the same time remaining sufficiently broad for an effective semantic description. A fine balance will need to be found between these competing requirements.

It has been found that it is easier for corpus users to formulate queries based on basic categories, i.e., basic groupings of words in a certain part of speech. For example, the basic categories for nouns are humans, animals, plants, tools, food, etc. Such classes are a reflection of the so-called “basic level” of human natural language categorization. This level is populated by categories in which perceptual, behavioral, and abstract features converge and which plays an especially prominent role in human categorization [11]. Basic semantic classes are not elementary and could theoretically be further decomposed, but this would result in the loss of their privileged status [7]. For example, the English verb *to melt* and its Ukrainian equivalent *танути* denote a change of state, which is a basic semantic class. Decomposing their meaning into 'begin to be different' would assign them to very broad low-level classes (inception, being, and difference), which they would share with many other words (*to start*, *existence*, *different*, etc.), but would miss the basic semantic class.

While moving one level up (to the superordinate level) or down (to the subordinate level) from the basic level is quite effortless and is regularly carried out by the speakers of the language, high levels of abstraction or, on the contrary, of detail are predominantly in the realm of science. Therefore, in developing the classification scheme for semantic annotation it is important to stay on, or close to, the basic level of categorization and strive to keep semantic classes psychologically real. This is highly relevant if semantic tags are to be used by a wide audience with uneven levels of linguistic competence and different backgrounds. The flexible taxonomic approach proposed here is compliant with the requirement of psychological and cognitive reality: indeed, numerous words ordinarily combine features of several classes, for example, the verb *break* means an action and a change of state, and it would be suboptimal to force it into just one of these categories.

Different word classes (verbs, adjectives, and adverbs, as well as concrete nouns and abstract nouns) have different sets of semantic tags. Nevertheless, several features are used consistently across POS boundaries to mark lexical items with similar meaning, for example, the tag **smell** applies equally to the verb *пахнути* 'to smell (good)', the noun *запах* 'smell', and the adjective *пахучий* 'fragrant'.

There are two independent parameters for semantic annotion expressing part–whole (or element–set) relations (mereology) and geometrical or spacial properties of objects (topology, e.g., **container** and **surface**). The list of topological tags may grow if further studies of usage patterns point to other relevant topological features.

Taxonomic classification is not necessarily flat, and in our case, it is complemented with elements of shallow hierarchical classification. For example, tools are subdivided into instruments, devices, means of transportation, weapons, etc., and qualities are divided into physical and abstract, while physical qualities are further subdivided into form, sound, color, light, taste, smell, temperature, and weight.

Semantic tags are assigned to a list of lemmas, thereby creating a semantic lexicon. In cases of ambiguity, i.e., when a lemma may have more than one set of semantic tags due to being used in multiple senses, all such sets are listed in the semantic lexicon, leaving the problem of semantic disambiguation for later stages. The semantic description of a word in a semantic lexicon is constructed so as to achieve certain explanatory and predictive power regarding various properties of this lexeme, such as its collocation patterns, valency, derivational potential, etc.

To sum up, faceted classification involves semantically cohesive categories that are flexibly combined as needed to best describe a given concept. This kind of classification stands closest to the naïve worldview of the users and their extralinguistic experience and is concordant with the principles of natural language categorization. The categorization scheme and the choice of semantic tags need to meet the criteria dictated by this approach and orientation.

3 Tools for Semantic Annotation

Each individual sense of a word singled out in explanatory dictionaries can potentially require a distinct set of semantic features. Even though lexicographic descriptions and semantic annotation differ in their respective objectives and the choice of semantic features, explanatory dictionaries are the most useful source of semantic information for the purposes of multifaceted semantic annotation. Other sources include dictionaries of specialized vocabulary (scientific and technical terms, slang, etc.), Wikipedia as a source of communal knowledge, and context in case of words lacking a lexicographic description.

The semantic annotation of GRAC is a natural complement to its morphological annotation. Morphological annotation enables the user to explore the corpus in a fairly complex manner, constructing search queries with the use of the Corpus Query Language (CQL) and employing a variety of grammar tags (POS and grammatical features such as number, gender, case, tense, etc.) [5]. For example, one may search for verbs in the past tense followed by singular nouns in the dative case within the span of three words.

Morphological annotation in GRAC is based on two key components developed by the r2u group: 1) VESUM, a large POS dictionary for Ukrainian currently containing over 400,000 lemmas from which over 6 million wordforms are generated [8]; 2) TagText, a POS tagger based on VESUM and also employing dynamic tagging for complex out-of-vocabulary words, dates, numbers, and punctuation symbols [10]. For a synthetic language, such as Ukrainian, it is convenient to add this layer when morphological (POS) annotation is already in place. This approach has several advantages. First, it makes use of lemmatization, which means that semantic tags can be added to lemmas instead of each individual wordform. Second, the TagText tagger will be modified to include semantic tags and tag texts for GRAC in one go. Third, semantic tags can be added incrementally and, for certain groups of words, automatically to the same dictionary (VESUM) which is already used by the tagger.

The semantic lexicon will be developed iteratively, enabling refinements of the semantic tagset, semantic classes, and features assigned to individual words. We envisage progression through the following stages:

1) developing the initial semantic tagset;
2) assigning semantic tags to the top 3,000 most frequent lemmas;
3) modifying the tagset if necessary;
4) semantic tagging of GRAC;
5) expanding the semantic lexicon, modifying the tagset, and tagging subsequent versions of GRAC.

4 Semantic Tagset

As mentioned above, each part of speech has its own set of semantic tags. Nouns are divided into the following large groups: **conc** (concrete nouns), **abst** (abstract nouns), and **prop** (proper names). Concrete and abstract nouns are treated separately as shown below. (For reasons of space, examples with glosses are provided selectively). Tags are shown in boldface, followed by a description and examples.

The taxonomic classification of concrete nouns is as follows: **hum** (human beings), **hum:group** (human groups, including groups of people based on ethnicity or place of origin or residence), **hum:kin** (kinship terms), **supernat** (supernatural beings), **animal** (animals), **plant** (plants), **mushr** (mushrooms), **stuff** (substances and materials), **loc** (locations and spaces), **build** (buildings and constructions), **vehicle** (vehicles), **furnit** (furniture), **dish** (plates, dishes, cups, and kitchen utensils), **cloth** (clothes and footwear), **food** (food and drinks), **org** (organizations), **event** (events), **work** (works, such as works of art and texts), and **tool** (tools and appliances in general). Tools are further subdivided into **tool:instr** (tools, implements, e.g., *ручка* ‘pen’), **tool:device** (devices), **tool:weapon** (weapons), and **tool:music** (musical instruments).

The tags **hum** and **supernat** can be quickly assigned based on a special mark (<) in the VESUM declension codes, which corresponds to these two categories. Likewise, animals are marked with <> and receive the tag **animal**. Thus, some 20,000 animate nouns in the lexicon can be efficiently supplied with semantic tags this way.

Mereology for concrete nouns is represented by the overarching label **part** (e.g., *середина* ‘middle part’), which is broken down into the following lower-level categories: **part:hum** (human body parts), **part:animal** (animal body parts), **part:plant** (parts of plants), **part:build** (parts of buildings and constructions), **part:tool** (parts of tools and appliances in general), **part:tool:instr** (parts of tools and implements, e.g., *ручка* ‘handle’), **part:tool:device** (parts of devices, e.g., *кнопка* ‘button’), **part:tool:weapon** (parts of weapons), **part:tool:music** (parts of musical instruments, e.g., *дека* ‘sounding board’), **part:tool:furnit** (parts of furniture, e.g., *ніжка* ‘leg’), **part:tool:dish** (parts of dishes, *носик* ‘spout’), **part:tool:cloth** (parts of clothes and footwear, e.g., *рукав* ‘sleeve’). Four tags describe mereology for concrete nouns: **quant** (quanta, i.e., particles and portions of substance, e.g., *крапля* ‘drop’), **set** (sets, e.g., *букет* ‘bouquet’), **collect** (collective nouns, e.g., *меблі* ‘furniture’, *селянство* ‘peasantry’), and **higher_class** (classes at the superordinate level, immediately above the basic level, e.g., *рослина* ‘plant’*, засіб* ‘tool’).

Concrete nouns have two topology tags, viz., **container** (containers, e.g., *зала* ‘hall’) and **surface** (surfaces, e.g., *майдан* ‘square’), and two evaluation tags, viz., **posit** (positive, e.g., *герой* ‘hero’) and **negat** (negative, e.g., *вайло* ‘sluggard’).

The taxonomy of abstract nouns is as follows: **move** (movement, motion, e.g., *біг* ‘running’, *переставляння* ‘rearrangement’), **move:body** (movement of a body part or change of position, e.g., *нахил* ‘bend’), **placing** (placement of an object, e.g. *завантаження* ‘loading’*, розташовування* ‘placement; ordering’), **impact** (physical impact, e.g., *удар* ‘strike’, *зішкрябування* ‘scraping off’), **creat** (creation of a physical object, e.g., *складання* ‘putting together’, *розроблення* ‘development’), **destr** (destruction, e.g. *руйнація* ‘ruination’, *злам* ‘breaking’), **change_state** (change of state or quality, e.g. *розігрів* ‘warming up’*, спрощення* ‘simplification’), **exist** (existence, e.g. *буття* ‘being’*, відсутність* ‘absence’), **appear** (beginning of existence, e.g., *постання* ‘emergence’, *народження* ‘birth’), **disappear** (end of existence, *смерть* ‘death’*, скасування* ‘cancellation’), **loc** (location or position, *урочище* ‘natural landmark’*, поза* ‘pose’), **loc:body** (special body position, e.g., *сидіння* ‘sitting’), **contact** (contact and support, e.g., *доторк* ‘touch’, *опертя* ‘support’), **poss** (possession, e.g., *придбання* ‘acquisition’, *втрата* ‘loss’), **ment** (mental domain, e.g., *думка* ‘thought’), **percept** (perception, e.g., *погляд* ‘look, view’), **psych** (psychological domain, e.g., *збудженість* ‘excitement’), **psych:emot** (emotion, e.g., *смуток* ‘sorrow’), **psych:vol** (volition, e.g., *бажання* ‘desire’), **speech** (speech acts, e.g., *обговорення* ‘discussion’), **physio** (physiology, e.g., *втома* ‘fatigue’), **weather** (weather phenomena, e.g., *дощ* ‘rain’), **sound** (sound, e.g., *дзенькіт* ‘ring(ing)’), **color** (color, e.g., *блакить* ‘blue color, azure’), **light** (light, e.g., *промінь* ‘ray’), **taste** (taste, e.g., *кислинка* ‘sour taste’), **smell** (smell, e.g., *пахощі* ‘’), **tempr** (temperature, e.g., *прохолода* ‘cool(ness)’), **weight** (weight, e.g., *тягар* ‘burden’), **time** (time, e.g., *прийдешнє* ‘future’), **time:period** (period, e.g., *строк* ‘period’), **time:moment** (moment in time, e.g., *реченець* ‘deadline’), **time:week** (day of the week), **time:month** (month), **time:age** (age, e.g., *молодість* ‘youth’), **quality:phys** (physical quality, e.g., *твердість* ‘hardness’), **quality:abst** (abstract quality, e.g., *непередбачуваність* ‘unpredictability’), **quality:abst:hum** (abstract quality of a person, e.g., *щедрість* ‘generosity’), **behave** (human behavior and acts, e.g., *халатність* ‘negligence’*, сварка* ‘quarrel’), **interact** (interaction and relationships, e.g., *допомога* ‘help’, *дружба* ‘friendship’), **event** (event, e.g., *збори* ‘meeting’, *фестиваль* ‘festival’), **disease** (disease, e.g., *грип* ‘influenza’), **game** (game, e.g., *шашки* ‘checkers’), **sport** (sports, e.g., *гімнастика* ‘calisthenics’), **param** (parameter, e.g., *довжина* ‘length’), and **unit** (unit of measurement, e.g., *секунда* ‘second’).

Abstract nouns have three tags for mereology: **part** (part, e.g., *початок* ‘beginning’), **quant** (quanta, e.g., *раз* ‘(one) time’, *момент* ‘moment’), and **set** (sets, e.g., *система* ‘system’). Two tags refer to evaluation: **posit** (positive, e.g., *насолода* ‘delight’) and **negat** (negative, e.g., *вульгарність* ‘vulgarity’).

Proper names receive four tags that are already implemented in the current system for POS tagging: **fname** (first name, e.g., *Тарас* ‘Taras’), **pname** (patronymic, e.g., *Григорович* ‘Hryhorovych’), **lname** (last name, e.g., *Шевченко* ‘Shevchenko’), and **geo** (geographical names, e.g., *Дніпро* ‘Dnieper’). Thus, a significant portion of the lemmas (over 53,000 proper names) in VESUM already have all the tags that are required for semantic annotation.

Adjectives are divided into nine taxonomic classes: **size** (size, e.g., *широкий* ‘wide’), **dist** (distance, e.g., *сусідній* ‘neighboring’), **quantit** (quantity, e.g., *нечисленний* ‘not numerous’), **orient** (orientation, direction, e.g., *правий* ‘right’, *прибережний* ‘coastal’*, зворотний* ‘reverse’), **time** (time, e.g., *майбутній* ‘future’), **dur** (duration, e.g., *короткочасний* ‘brief, short-lived’), **age** (age, e.g., *старий* ‘old’), **speed** (speed, e.g., *меткий* ‘quick, nimble’), and **quality**. Quality is not an independent feature; rather, it is qualified as either physical or abstract and supplied with a specific semantic tag: **quality:phys** (physical quality, e.g., *липучий* ‘sticky’), **quality:phys:form** (form, e.g., *вигнутий* ‘bent’), **quality:phys:sound** (sound, e.g., *дзвінкий* ‘resounding’), **quality:phys:color** (color, e.g., *зелений* ‘green’), **quality:phys:light** (light, e.g., *яскравий* ‘bright’), **quality:phys:taste** (taste, e.g., *гіркий* ‘bitter’), **quality:phys:smell** (smell, e.g., *духмяний* ‘fragrant’),

quality:phys:tempr (temperature, e.g., *гарячий* ‘hot’), **quality:phys:weight** (weight, e.g., *легкий* ‘light’), **quality:abst** (abstract quality, e.g., *непередбачуваний* ‘unpredictable’), **quality:abst:hum** (abstract quality of a person, e.g., *хитрий* ‘cunning’), **quality:abst:ment** (abstract quality in the mental domain, e.g., *зрозумілий* ‘understandable’).

Adjectives have three additional tags (**max**, **min**, and **absol** ‘absolute’) that are combined with the following tags: **size**, **dist**, **quant**, **dur**, **age**, and **speed**. For example, **size:max** (large size, e.g., *довгий* ‘long’), **size:min** (small size, e.g., *низький* ‘low’), and **size:absol** (absolute size, e.g., *триметровий* ‘three-meter long’). Two evaluation tags are also applied: **posit** (positive, e.g., *щасливий* ‘happy’) and **negat** (negative, e.g., *нечесний* ‘dishonest’).

Adverbs are divided into 12 large categories: **place** (location, e.g., *тут* ‘here’), **orient (**orientation, direction, e.g., *вниз* ‘down’*,* *праворуч* ‘to/on the right’), **dist** (distance, e.g., *недалеко* ‘not far’*,* *високо* ‘high’), **quantit** (quantity, e.g., *трішки* ‘a little’*,* *тричі* ‘three times’), **time** (time, e.g., *відтепер* ‘from now on’), **dur** (duration, e.g., *недовго* ‘not long’), **speed** (speed, *поволі* ‘slowly’), **manner** (manner, e.g., *по-українськи* ‘in Ukrainian’, *навприсядки* ‘in a squatting position’), **degree** (degree, *достатньо* ‘sufficiently’), **cause** (cause, *спересердя* ‘in anger’), **goal** (purpose, e.g., *навмисне* ‘intentionally’), and **quality**. Many qualities have essentially the same meaning whether expressed by adverbs or by adjectives. Thus, adverbs receive the same tags as adjectives to describe qualities: **quality:phys** (physical quality, e.g., *рівно* ‘smoothly’), **quality:phys:form** (form, e.g., *криво* ‘obliquely’), **quality:phys:sound** (sound, e.g., *дзвінко* ‘resoundingly’), **quality:phys:color** (color, *барвисто* ‘colorfully’), **quality:phys:light** (light, e.g., *яскраво* ‘brightly’), **quality:phys:taste** (taste, e.g., *гірко* ‘bitterly’), **quality:phys:smell** (smell, *духмяно* ‘fragrantly’), **quality:phys:tempr** (temperature, *холодно* ‘coldly’), **quality:phys:weight** (weight , *легко* ‘lightly’), **quality:abst** (abstract quality, *непередбачувано* ‘unexpectedly’), **quality:abst:hum** (abstract quality of a person, *хитро* ‘cunningly’), and **quality:abst:ment** (abstract quality in the mental domain (*зрозуміло* ‘understandably’).

The categorization scheme for adverbs includes two additional tags (**max** and **min**) that are combined with the tags **dist**, **dur**, **speed**, and **quantit**. For example, **dist:max** (large distance, e.g., *далеко* ‘far away’) and **dist:min** (short distance, e.g., *поблизу* ‘near’). Similar to adjectives and nouns, adverbs have two evaluation tags: **posit** (positive, e.g., *щасливо* ‘happily’) and **negat** (negative, *нечесно* ‘dishonestly’).

Verbs have ramified semantic classification with few hierarchical elements: **move** (movement, e.g., *іти* ‘to walk’, *штовхати* ‘to push’), **move:body** (movement of a body part or change of position, e.g., *нахилятися* ‘to bend’), **placing** (placement of an object, e.g., *завантажити* ‘to load’*,* *розташувати* ‘to place, to arrange’), **impact** (physical impact, e.g., *ударяти* ‘to strike’, *зішкрябувати* ‘to scrape off’), **creat** (creation of a physical object, e.g., *складати* ‘to put together’, *розробляти* ‘to develop’), **destr** (destruction, e.g., *руйнувати* ‘to destroy’, *зламати* ‘to break’), **change_state** (change of state or quality, e.g., *розігрівати* ‘to warm up’*,* *спростити* ‘to simplify’), **exist** (existence, e.g., *бути* ‘to be’), **appear** (beginning of existence, e.g., *народитися* ‘to be born’), **disappear** (end of existence, e.g., *померти* ‘to die’, *скасувати* ‘to cancel’), **loc** (location or position, e.g., *покласти* ‘to put’), **loc:body** (special body position, e.g., *сидіти* ‘to sit’), **contact** (contact and support, e.g., *торкатися* ‘to touch’, *спиратися* ‘to rest on’), **poss** (possession, e.g., *придбати* ‘to acquire’, *втратити* ‘to lose’), **ment** (mental domain, e.g., *думати* ‘to think’), **percept** (perception, e.g., *дивитися* ‘to look’*,* *побачити* ‘to see’), **psych** (psychological domain, e.g., *турбуватися* ‘to be concerned’), **psych:emot** (emotion, e.g., *засмучуватися* ‘to be sad’), **psych:vol** (volition, e.g., *бажати* ‘to desire’), **speech** (speech acts, e.g., *обговорювати* ‘to discuss’), **behave** (human behavior and acts, e.g., *дражнитися* ‘to tease’), **physio** (physiology, e.g., *втомлюватися* ‘to become tired’), **weather** (weather phenomenon, e.g., *дощити* ‘to rain’), **sound** (sound, e.g., *дзенькнути* ‘to tinkle’), **color** (color, e.g., *біліти* ‘to become white’), **light (**light, e.g., *засвітитися* ‘to light up’), **taste (**taste, e.g., *гірчити* ‘to taste bitter’), **smell (**smell, e.g., *духмяніти* ‘to smell pleasant’), **caus** (causative verbs, e.g., *скласти* ‘to put smth together’), and **noncaus** (non-causative verbs, e.g., *повертатися* ‘to return’).

While semantic tagsets are developed separately for each part of speech, it is important to uniformly tag similar semantic content. For example, physical qualities are consistently assigned the same tags for nouns and adjectives: **sound**, **color**, **light**, etc. Abstract nouns and verbs also share a number of semantic tags: **move**, **placing**, **impact**, etc. This way, a search query can be flexibly formulated to either zero in on a specific part of speech with a given tag or retrieve multiple parts of speech referring to the same physical quality.

5 Conclusion

The problem of semantic annotation for Ukrainian, specifically for the GRAC corpus, can be resolved using the taxonomic approach and relying on insights from human natural language categorization. The semantic tagset proposed here can be applied to create a semantic lexicon by assigning semantic tags to individual lemmas in VESUM, a POS dictionary for Ukrainian. TagText, the POS tagger for Ukrainian by the r2u group, can then be modified to carry out semantic tagging of texts.

Complex queries enabled by a combination of morphological (POS) and semantic tagging can be a powerful tool for corpus studies, linguistic research, and a variety of NLP applications. In particular, it has the potential to enhance the search functionality of the GRAC corpus and open up opportunities for the study of semantic classes in Ukrainian.

6 References

1. Corpus of the Ukrainian Language, available at http://www.mova.info/corpus.aspx
2. Darchuk, N.P.: Mozhlyvosti semantychnoyi rozmitky korpusu ukrainskoyi movy (KUM) [Possibilities of the Semantic Markup of the Corpus of the Ukrainian Language (KUM)]. In: Naukovyi chasopys Natsionalnoho pedahohichnoho

universytetu im. M.P. Drahomanova. Seriya 9: Suchasni tendentsiyi rozvytku mov, Vypusk 15, pp. 18-28. (2017) (in Ukrainian)

3. Evans, V., Green, M.: Cognitive Linguistics. An Introduction. Edinburgh. (2006)
4. FrameNet, available at http://framenet.icsi.berkeley.edu
5. Shvedova, M., von Waldenfels, R., Yarygin, S., Kruk, M., Rysin, A., Starko, V., Woźniak, M.: GRAC: General Regionally Annotated Corpus of Ukrainian. Kyiv, Oslo, Jena. (2017-2020), available at uacorpus.org.
6. Kustova G.I., Lyashevskaya O.N., Paducheva E.V., Rakhilina E.V.: Semanticheskaya razmetka leksiki v natsionalnom korpuse russkogo jazyka: printsipy, problemy, perspektivy [Semantic Markup of Vocabulary in the Russian National Corpus: Principles, Problems and Perspectives]. In: Nationalnyi korpus russkogo yazyka: 2003-2005 [Russian National Corpus: 2003-2005], Moskva, pp. 155–174. (2005) (in Russian)
7. Rakhilina E.V., Kustova G.I., Lyashevskaya O.N., Reznikova T.I., Shemanaeva O. Ju.: Zadachi i printsipy semanticheskoy razmetki leksiki v NKRJa [The Objectives and Principles of Semantic Markup of Vocabulary in the Russian National Corpus]. In: Nationalnyi korpus russkogo yazyka. Novyie rezultaty i perspektivy [Russian National Corpus. New Results and Perspectives]. Sankt-Peterburg, pp. 215-239. (2009) (in Russian)
8. Starko, V.: Kompiuterni linhvistychni proekty hurtu r2u: stan I zastosuvannia [Computational Linguistic Projects of the r2u Group: Progress and Applications]. In: Ukrainska mova, No. 3, pp. 86–100. (2017) (in Ukrainian) http://nbuv.gov.ua/UJRN/Ukrm_2017_3_9
9. Starko, V.: Paradyhma kohnityvnoyi linhvistyky j problema katehoryzatsiyi [Paradigm of Cognitive Linguistics and the Problem of Categorization]. In: Naukovi zapysky [Natsionalnoho universytetu “Ostrozka akademiya”] Seriya: Filolohichna, Vyp. 48, pp. 113–116. (2014) (in Ukrainian) http://nbuv.gov.ua/UJRN/Nznuoaf_2014_48_37.
10. Starko, V.: Rozviazannia kompiuternolinhvistychnykh zavdan zasobamy hurtu r2u [Solving Computational Linguistic Problems with Tools Developed by the r2u Group]. In: U poshukakh harmoniyi movy, Kyiv, pp. 367–373. (2020) (in Ukrainian) https://r2u.org.ua/data/other/Klymenko_2020/Klymenko_2020.pdf
11. Taylor, J.R.: Linguistic Categorizaton. 2nd Ed., Oxford. (1995)
12. UCREL Semantic Analysis System, available at http://ucrel.lancs.ac.uk/usas
13. Wierzbicka, A.: Lexicography and Conceptual Analysis. Tucson. (1995)
14. WordNet, available at http://wordnet.princeton.edu