Implementing Semantic Annotation in a Ukrainian Corpus

Vasyl Starko

Ukrainian Catholic University, 2a Kozelnytska Str., Lviv, 79026, Ukraine

Abstract

The paper describes the first phase of semantic annotation implemented in the General Regionally Annotated Corpus of Ukrainian (GRAC) using the Ukrainian Semantic Lexicon (USL) and the TagText tagger for Ukrainian. Over 1,000 most frequent lemmas were supplied with semantic tags, creating the foundation for the lexicon. In the process of developing the USL, the original semantic tagset underwent changes and was expanded. The revised tagset is presented, and the linguistic aspects of practical semantic annotation are analyzed. The TagText tagger was updated to enable both morphological and semantic annotation of Ukrainian texts. The current versions of the USL and TagText are released and available for download. Text coverage by semantic tags in GRAC is discussed, and examples of semantic and complex searches in the GRAC corpus are provided. Plans for future work on the USL are outlined.

Keywords 1

Semantic annotation, corpus, Ukrainian Semantic Lexicon, USL, semantic tagset, GRAC, TagText, categorization

1. Introduction

Processing natural language at the semantic level is an important part of computational linguistics and NLP. Several well-known approaches to annotating texts with semantic labels have been developed and implemented, including WordNet [27], FrameNet [9], USAS [13], [26], and taxonomic classification [12]. All of these projects involve the construction of a lexicon or a lexical database as a key resource for semantic tagging.

A hierarchical approach, such as USAS, has a number of merits: it is lexicon-based; tagging is efficiently performed by a semantic tagger; most notably, it employs a universal taxonomy that can be applied to a variety of structurally different languages [14]. However, it has been argued [15], [24] that this kind of system has certain drawbacks that make it suboptimal for semantic annotation in corpora: a high degree of arbitrariness in the overall classification; strict adherence to the dichotomous, one-category-per-word principle; excessive abstraction at the top layers; a lack of coherence within some categories, and counterintuitive groupings. In contrast, the taxonomic classification is more flexible and non-dichotomous, allowing a word to be assigned to more than one semantic class. It is also well-aligned with the categorization that is inherent in natural language and affects language use as observed in the speech production of a linguistic community. Furthermore, taxonomic tags have descriptive, easy-to-remember names and are stated explicitly, which is an advantage for non-professional corpus users.

As far as the formal semantic description of Ukrainian is concerned, Darchuk [5], [6] offers an overview of approaches to semantic annotation in corpora and beyond, discusses semantic resources based on synsets and frames, and presents models of several thesauri and ontologies for Ukrainian. Darchuk and her colleagues later have also proposed a differential approach to the semantic annotation of a Ukrainian corpus [4]: ontologies are to be used for scientific texts, while other types of

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: v.starko@ucu.edu.ua (V. Starko)
ORCID: 0000-0002-2530-2107 (V. Starko)

texts are to be annotated based on taxonomic classification [7]. In this latter approach, each individual sense described in a monolingual dictionary is treated as a separate entry and is assigned tags in a semantic lexicon, which is the key resource for the semantic annotation of a corpus [8]. Then, in the process of manual disambiguation, a linguist uses a web interface tool to select the appropriate sense for a given occurrence of a word in the corpus. One of the unfortunate but inevitable consequences of this approach is the multiplication of semantic entries: for example, the authors report that 16,000 nouns drawn from a news subcorpus require nearly three times as many entries in the semantic dictionary [8: 99–100]. This may complicate the challenge of manual disambiguation, to say nothing of automatic disambiguation. To date, only a demo version of the semantic dictionary (30 lemmas) developed by this team is available online [18], while the corpus itself [4], as accessed via its web interface, does not (yet) have the semantic annotation layer. The complete semantic tagset is not available either.

The framework for semantic annotation developed in the Ukrainian Linguistic Information Fund also follows the structure of lexicographic definitions adopted in a large general-purpose monolingual dictionary of Ukrainian, including division into senses and subsenses. The dictionary data is parsed and stored in a database from which it can then be drawn for tagging corpora [21].

A project to create WordNet for Ukrainian [1] was developed by a group of researchers in Kyiv, but it is now defunct with apparently no resulting resource available to the public. Other efforts along this line [11] have not come to fruition. However, several resources are available for sentiment analysis in Ukrainian [10], [19]. In general, there are no publicly available large-scale resources and tools for processing Ukrainian-language texts at the semantic level. The Ukrainian Semantic Lexicon and the TagText tagger described in this paper are aimed precisely at filling this void.

2. Previous work

It has been argued that, if a general-purpose and large-scale semantic resource similar to WordNet is lacking for a language, it makes sense to take advantage of the benefits offered by the taxonomic approach to semantic annotation [24]. To this end, a lexicon needs to be developed as a key resource from which semantic data can be drawn. In the project presented here, this role is played by the Ukrainian Semantic Lexicon.

As the first step in this direction, the principles of semantic annotation were laid down and the initial classification scheme (semantic tagset) was developed and presented in [24]. The following properties were defined for the semantic annotation scheme for Ukrainian:

- faceted, rather than hierarchical, approach to classification, allowing multiclass assignment of words;
- accessibility and transparency to various group of users, including nonlinguists;
- absence of ambiguity in the form and content of semantic tags;
- taxonomic classes must be linguistically meaningful, psychological real, and cognitively motivated;
- classification should represent the linguistic worldview of native speakers and natural language categorization inherent in the language with a special focus on the basic-level categories, which have a privileged status [2];
- semantic tags must be independent and basic, form sufficiently large classes, generate little noise, and be highly suitable for search queries;
- semantic classes picked out by semantic tags should be maximally homogeneous and, at the same time, sufficiently broad;
- distinct semantic tagsets are to be developed for verbs, adjectives, adverbs, concrete nouns, and abstract nouns, but similar semantic content has to be designated by identical tags across POS boundaries;
- taxonomic classification may involve elements of shallow (two-level) hierarchy;
- semantic tags are to be assigned to individual monosemous and polysemous lemmas in a semantic lexicon.

The initial classification scheme for Ukrainian presented in [24] consists of distinct tagsets for six large word classes (concrete nouns, abstract nouns, proper nouns, verbs, adjectives, and adverbs) and employs semantic tags of four types (for taxonomy, mereology, topology, evaluation, and causativity). The composition of the current semantic tagset is shown in Table 1. The figures represent the number of tags of a specific type available for a given word class.

Table 1
Composition of the semantic tagset

Tag types\Classes	Concrete nouns	Abstract nouns	Proper nouns	Verbs	Adjectives	Adverbs
Taxonomy	33	54	6	50	48	38
Mereology	24	4				
Topology	4					
Evaluation	2	2			2	2
Causativity				2		

This classification scheme was used as the starting point for assigning semantic tags to Ukrainian lemmas and adding them to the Ukrainian Semantic Lexicon. The semantic description of lemmas in the Ukrainian Semantic Lexicon has the form of ordered tag sequences in which individual tags follow the shown order—from taxonomy to causativity. For example, a concrete noun may be assigned a taxonomic tag followed by a topological tag. In the process of this work, it became apparent that certain tags needed to be changed, while others had to be added. Thus, the initial tagset was modified, refined, and re-applied to the selected vocabulary. After multiple iterations, over 1,000 lemmas were annotated and added to the USL, and the resulting semantic tagset was published online in Ukrainian and English in the documentation for the General Regionally Annotated Corpus of Ukrainian [20] (see the menu Metatextual Annotation > Semantic Annotation). An overview of the modifications to the original semantic tagset is provided in section 3.

3. Modifications to the initial semantic tagset

All of the modifications that have been made to the semantic tagset fall into one of the following categories:

- addition—a new semantic tag was added;
- renaming—an existing tag was renamed without changes to its content;
- rearrangement—existing tags were reordered, for example, to create two-level hierarchy;
- reinterpretation—an existing tag was reinterpreted as having a broader or narrower scope.

Eleven taxonomic tags have been added to the tagset for concrete nouns:

conc:doc documents (*акція* 'share of stock'*, квиток* 'ticket'*, диплом* 'diploma');

conc:food&fruit edible fruit (*вишня* '(sweet) cherry'); here **fruit** is used in the scientific, rather than naïve, sense;

conc:form form (*лінія* 'line'*, гора* 'mountain');

conc:hum:prof profession (*вчителька* '*fem.* teacher'*, журналіст* 'journalist');

conc:loc:room rooms in buildings (*офіс* 'office', *кухня* 'kitchen');

conc:money money (*долар* 'dollar', *грн* 'UAH');

conc:org&&build organizations and buildings (*школа* 'school', *лікарня* 'hospital');

conc:poss possession (*майно* 'possessions'); this tag was initially envisaged only for abstract nouns but has been found useful also in relation to concrete vocabulary;

conc:speech speech unit (*слово* 'word', *склад* 'syllable');

conc:text textual objects (*лист* 'letter'*, договір* 'contract');

conc:thing unspecified individual objects (*річ* 'thing', *об'єкт* 'object'*, продукт* 'product'). This tag has been introduced for nouns that designate concrete objects but are highly general nature.

Rearrangement took place in the mereology part of the tagset for concrete nouns as it was decided to move the mereological tag **part** from the front to the end of tag sequences, e.g., **conc:tool:music:part**. This was done to increase the convenience of constructing search queries. For example, the query **conc:build.*** would now match references to both entire buildings and their parts, such as hallways and domes. The following tags were rearranged or added in the mereology section:

conc:body:animal:part animal body parts (*хвіст* 'tail', *кіготь* 'claw');

conc:body:hum:part human body parts (*ніготь* 'nail', *мізинець* 'pinky');

conc:body:part body parts of humans and animals (*нейрон* 'neuron', *печінка* 'liver');

conc:food:part parts of food or meals (*скорина* 'crust', *друге* 'second course');

conc:loc:part parts of locations and spaces (*дно* 'bottom', *поверхня* 'surface');

conc:loc:room:part parts of rooms (*вікно* 'window', *батарея* 'radiator');

conc:org:part parts of organizations (*відділ* 'department'*, кафедра* 'university department');

conc:quantum particles and portions of substance (*крихта* 'crumb', *уламок* 'fragment');

conc:text:part parts of texts (*зміст* 'table of contents');

conc:vehicle:part parts of vehicles (*кермо* 'steering wheel', *педаль* 'pedal');

Two tags were added to the topology section:

conc:ball spheres, globes (*м'яч* 'ball'*, сонце* 'sun');

conc:line lines (*кордон* 'border'*, стрічка* 'ribbon').

The taxonomic part of the tagset for abstract nouns has undergone mostly superficial modifications. Several tags have been renamed without any changes in content:

abst:abst abstract quality (*непередбачуваність* 'unpredictability', *якість* 'quality');

abst:abst:humqual abstract quality of a person (*доброта* 'kindness', *щедрість* 'generosity');

abst:chstate change of quality or state (*розширення* 'expansion', *сповільнення* 'slowing down');

abst:create creation of a physical object (*виробництво* 'production', *складання* 'assembly');

abst:put placement of a physical object (*встановлення* 'installation', *запис* 'recording').

A handful of tags referring to physical qualities have been reorganized:

abst:physqual physical quality (*м'якість* 'softness', *слизькість* 'slipperiness');

abst:physqual:color color (*синява* 'blueness', *відтінок* 'shade of color');

abst:physqual:form form (*вигнутість* 'curvature'*, опуклість* 'convexity');

abst:physqual:hum human quality (*дужість* 'strength'*, моторність* 'agility');

abst:physqual:smell smell (*чад* 'smoke', *аромат* 'aroma');

abst:physqual:sound sound (*луна* 'echo', *плюскіт* 'squish');

abst:physqual:taste taste (*терпкість* 'tartness', *солодкавість* 'sweetness');

abst:physqual:tempr temperature (*спека* 'heat', *мороз* 'frost');

abst:physqual:vis visual appearance (*вигляд* 'look', *тьмяність* 'dimness');

abst:physqual:weight weight (*ноша* 'burden', *баласт* 'ballast').

Five new taxonomic tags have been added for abstract nouns:

abst:interact:conflict conflict, confrontation (*дуель* 'duel', *війна* 'war'*, боротьба* 'fight');

abst:quantit quantity (*тисяча* 'thousand', *млн* 'million');

abst:quantit:max maximum quantity (*сила* 'great many'*, море* 'a sea (of smth.)');

abst:state state, condition (*безпека* 'safety; security', *цілісність* 'integrity');

abst:vis abstract representation (*образ* 'image of an epoch, etc.').

The mereological tag **quant** has been renamed as **quantum** and the new mereological tag **collect** has been added:

abst:collect collection of different entities (*інститут* 'institute'*, механізм* 'mechanism');

abst:quantum quantum (*випадок* 'case', *раз* 'time, instance', *момент* 'moment');

The taxonomic part of the tagset for adjectives has been extensively rearranged and expanded and now includes the following modified tags:

abst abstract quality (*безпечний* 'safe', *невпинний* 'relentless');

abst:hum abstract quality of a person (*розумний* 'smart', *добрий* 'good', *хитрий* 'cunning');

abst:ment abstract quality in the mental domain (*незрозумілий* 'incomprehensible');

abst:sim similarity (*однаковий* 'same'*, інший* 'other, different'*, аналогічний* 'analogous');

abst:vis visual appearance of humans and objects (*згорблений* 'stooped'*, усміхнений* 'smiling');

degree degree (*помірний* 'moderate');

degree:max maximum degree (*видатний* 'outstanding, prominent', *всесильний* 'omnipotent');

degree:min minimum degree (*мізерний* 'meager');

hierar hierarchical attribute (*рядовий* 'rank-and-file'*, центральний* 'central');

ord ordinality (*третій* 'third'*, наступний* 'next');

physio physiological quality (*хворий* 'sick');

physqual physical quality (*слизький* 'slippery', *м'який* 'soft');

physqual:color color (*бірюзовий* 'turquoise', *золотистий* 'golden');

physqual:form form (*рівний* 'smooth', *круглий* 'round');

physqual:hum human physical quality (*дужий* 'strong'*, моторний* 'agile');

physqual:smell smell (*ароматний* 'fragrant');

physqual:sound sound (*лункий* 'hollow', *щебетливий* 'chirping');

physqual:taste taste (*пряний* 'spicy', *смачний* 'delicious', *терпкий* 'tart');

physqual:tempr temperature (*прохолодний* 'cool', *гарячий* 'hot');

physqual:vis visual appearance (*тьмяний* 'dim', *блискучий* 'shiny', *іскристий* 'sparkling');

physqual:weight weight (*масивний* 'massive', *тяжкий* 'heavy');

poss possession (*Андріїв* 'Andrii's'*, власний* 'own'); this tag means belonging to somebody in a broad sense rather than material possession per se;

psych:emot emotion (*злий* 'angry', *радісний* 'happy').

Adverbial tags referring to physical and abstract qualities have been substantially reorganized and several new taxonomic tags have been added:

abst abstract quality (*безпечно* 'safely, securely', *невпинно* 'relentlessly');

abst:hum abstract quality of a person (*суворо* 'strictly', *чесно* 'honestly', *хитро* 'cunningly');

abst:ment abstract quality in the mental domain (*уважно* 'carefully', *чітко* 'clearly');

degree:max maximum degree (*сильно* 'strongly'*, максимально* 'maximally');

degree:min minimum degree (*нітрохи* 'not in the least'*, ледве* 'hardly'*, трішки* 'a little');

freq frequency (*часто* 'frequently'*, іноді* 'sometimes');

modal modality (*треба* 'need to'*, безумовно* 'undoubtedly'*, звичайно* 'of course');

ord order (*насамперед* 'first of all'*, по-друге* 'secondly');

physqual physical quality (*чисто* 'purely', *м'яко* 'softly', *цілком* 'wholly');

physqual:color color (*зеленаво* 'greenishly', *квітчасто* 'colorfully');

physqual:form form (*тупо* 'bluntly', *круто* 'steeply; abruptly');

physqual:hum human physical quality (*моторно* 'in an agile manner');

physqual:smell smell (*затхло* 'mustily', *духмяно* 'fragrantly');

physqual:sound sound (*гучно* 'loudly', *тихо* 'quietly');

physqual:taste taste (*смачно* 'deliciously', *пікантно* 'spicily', *гірко* 'bitterly');

physqual:tempr temperature (*гаряче* 'hotly', *холодно* 'coldly');

physqual:vis visual appearance (*світло* 'lightly', *темно* 'darkly', *видно* 'visibly');

physqual:weight weight (*важко* 'heavily').

The class of verbs remains the most difficult one for semantic annotation. Despite having an extensive initial set of available tags, verbal taxonomy has been subjected to multiple rounds of rearrangment and modification. On the surface level, several tags have been renamed:

chstate change of state or quality (*лікувати* 'to heal', *спростити* 'to simplify');

create creation of a physical or non-physical object (*встановити* 'to establish');

put placement of an object (*ставити* 'to put smth. somewhere', *розсадити* 'to plant').

Tags referring to qualities have been rearranged in line with other parts of speech:

physqual physical quality (*тужавіти* 'to become solid, to harden');

physqual:color color (*червоніти* 'to turn red');

physqual:form form (*рівнішати* 'to become more even', *вигнутися* 'to bend about smth.');

physqual:hum human physical quality (*підрости* 'to grow up');

physqual:smell smell (*духмяніти* 'to smell nicely');

physqual:sound sound (*звучати* 'to sound'*, щебетати* 'to chirp');

physqual:taste taste (*смакувати* 'to taste'*, гірчити* 'to taste bitter');

physqual:tempr temperature (*холоднішати* 'to become colder');

physqual:vis visual appearance (*маяти* 'to flutter'*, виникати* 'to appear in sight').

More significantly, a number of tags have been added to improve coverage of verbal semantics:

able ability (*могти* 'to be able'*, уміти* 'to know how to; to be able to');

act action without specification (*діяти* ‘to act’*, виконувати* ‘to fulfill’);

begin begin (doing) smth. (*започаткувати* ‘to start, to launch’*, відкрити* ‘to open’);

effect non-physical effect, influence on smth. (*допомагати* ‘to help’*, сприяти* ‘to facilitate’);

effort exertion of effort (*старатися* ‘to make an effort, to exert oneself’*, намагатися* ‘to try’);

end end, finish, stop, terminate (*зупинитися* ‘to stop doing smth.’*, закінчуватися* ‘to end’);

func function (*робити* ‘to work (about smth.)’, *функціювати* ‘to function’);

grasp grasp (*взяти* ‘to take’*, схопити* ‘to grab’);

limit reaching or moving towards a limit (*наїстися* ‘to eat one’s fill’);

modal modality (*мовляти* in the form *мовляв* ‘so to say, as it were’);

move&loc movement and change of location (*зайти* ‘to enter’, *побувати* ‘to visit (some place)’);

orient move, change, etc. in a certain direction, orientation, physical or non-physical (*направляти* ‘to direct’*, вести* ‘to lead’);

phase phasal verb (*починати* ‘to begin’, *продовжувати* ‘to continue’, *закінчувати* ‘to end’);

prof profession (*вчителювати* ‘to work as a teacher’, *теслювати* ‘to carpenter’);

psych:want wanting, allowing, willing (*хотіти* ‘to want’, *дозволяти* ‘to allow’);

use use (*використовувати* ‘to use’*, застосовувати* ‘to apply’).

In the course of semantica annotation of Ukrainian vocabulary, it was established that a number of verbal senses lacked taxonomic semantic tags. Whenever they were identified in a range of verbs, new tags were added to the tagset. However, blanks still remain where such senses were not (easily) generalizable. As work on the Ukrainian Semantic Lexicon continues, the taxonomic tagset is likely to be expanded for verbs—more so than for the other parts of speech, which appear to have near-optimal tagsets that are sufficiently extensive and rich for the description of their semantics.

The latest version of the semantic tagset involves a greater number of elements organized in the form of two-level hierarchy. For example, the semantic feature **physqual** is a higher-order tag for color, form, smell, sound, taste, temperature, weight, visual appearance, and physical human qualities. Abstract qualities (**abst**) are also subdivided into several types—human, mental, similarity, visual. A researcher can, therefore, operate at the level of physical or abstract qualities or zero in at the lower level and specify particular types that are of interest.

4. Developing the Ukrainian Semantic Lexicon

The Ukrainian Semantic Lexicon has been filled with 1,000+ entries in the descending order of frequency. The motivation behind this approach has been twofold: first, this is efficient as higher-frequency words make greater contribution to lexical coverage; second, the top part of the frequency list includes a diverse selection of basic nouns, verbs, adjectives, and adverbs, which are an ideal testing ground for the classification scheme and a solid foundation for the rapid annotation of semantically related lower-frequency vocabulary.

Following the principle of multifaceted tag assingment, semantic tags in the USL are combined in a number of cases to reflect the complex semantics of words they are applied to. The double ampersand **&&** is used to indicate regular polysemy and can be read as “also used in the sense”. For example, the tag combination **org&&build** is assigned to words that designate an organization but can also be used to refer to a building in a particular context: *лікарня* ‘hospital’*, музей* ‘museum’, and *міністерство* ‘ministry’. The single ampersand symbol **&** marks simultaneity of semantic features. One example here is the tag sequence **abst:time:period&unit** found in the entries for the following words to indicate simultanenous expression of a period and unit of time: *хвилина* ‘minute’*, година* ‘hour’, *день* ‘day’, *тиждень* ‘week’, *місяць* ‘month’, *рік* ‘year’, and *століття* ‘century’.

A special effort was made to secure uniformity of semantic description. This was achieved by refining the tagset to make sure that identically named tags are used for similar semantic content across word classes. A case in point here is the domain of physical qualities (color, smell, sound, etc.) that have the same tags among nouns, adjectives, adverbs, and verbs (see the examples of tags involving **physqual** in section 3).

In line with the structure and format adopted in VESUM, homonyms in the USL are identified as xp1, xp2, etc. Whenever there are morphological honomyms in VESUM, they are treated as separate entries also in the USL. Currently, two graphical words (*сон* ‘sleep; pasqueflower; sone’ and *чин* ‘a high-ranking person; rank; deed, a way of doing') have three USL entries and 39 words have two (see an example below), while all other entries are unambiguous at this level.

In the USL, each entry is assigned a string of semantic tags. Senses are numbered (1, 2, 3, etc.), with the most frequent sense in the first position. In the absence of a semantically annotated and disambiguated corpus of Ukrainian, the ordering of senses was based on data from several sources: a monolingual dictionary of Ukrainian, corpus data, and the compiler's linguistic intuition. It should be emphasized that sense division was dictated by the classification scheme adopted for a given word class in the USL rather than by sense distribution in a monolingual dictionary. Because the USL has a higher level of semantic granularity, dictionary senses were often merged into one USL sense, whereas the splitting of dictionary senses had to be done only in a few cases. Let us consider an example of a noun entry in the USL:

голова 'head (*person*)' xp1 **1:conc:hum&hierar**

голова 'head' xp2 **1:conc:body:part:2:abst:ment:3:abst:unit**

Two morphological honomyms are distinguished in VESUM for this word: the first one refers to a person (e.g., head of an organization), while the second one corresponds to all inanimate senses. The hononymy is preserved in the USL. The first entry is tagged as as concrete noun (**conc**) referring to a human being (**hum**) and, at the same time, a place in a hierarchy (**hierar**). Individual semantic tags are separated by colons, except for such combinations as **hum&hierar**, where ampersands are used to conjoin semantic features. The second homonym (xp2) has three senses. It is most frequently used as a concrete noun designating a body part (**1:conc:body:part**) and can also refer to mental abilities in which sense it is abstract (**2:abst:ment**). In its least frequent sense, this noun refers to a unit in numbering cattle (cf. *thirty head of cattle*).

The following adjectival entry in the USL

останній 'last' **1:ord:2:time:3:degree:max:negat**

signifies that the adjective most often describes a characteristic related to sequential order (**ord**), less frequently to temporal sequence (**time**), and is also used as an intensifier with negative evaluation (e.g., *останній негідник* 'dirty scoundrel').

Adverbs are supplied annotations that cover purely adverbial senses and, additionally, predicative and modal uses. For example, the Ukrainian *звичайно* most often occurs in texts as a modal word but is also used as an adverb of manner:

звичайно 'of course; ordinarily' **1:modal:2:manner**

The most frequent use of the adverb *важко* is that of a predicate (*Мені важко зрозуміти.* It's hard for me to understand.), but it also expresses the adverbial senses of manner and physical quality:

важко 'heavily' **1:abst:hum:2:manner:3:physqual:weight**

Verbal semantics is notoriously hard to capture in terms of semantic features. High-frequency verbs exhibit an extremely complex network of meanings and semantic nuances. Here is an example of a moderately complex verbal USL entry:

взяти 'take; grasp' **1:grasp:caus:2:poss:caus:3:phase:4:ment:caus**

This verb is most often used to refer to an act of physical grasping and taking in possession but can also express the beginning of an action (**phase**) and a mental act (**ment**). All senses, except for **phase**, are causative, meaning that the action affects the object it is directed at.

One key distinction of the approach presented here is that it represents coarse-grained semantic annotation, which reduces the average number of senses per word as compared to approaches strictly based on sense division in monolingual dictionaries. Similar observations at this level of semantic granularity (the so-called supersenses) have been made regarding English nouns and verbs in comparison with WordNet synsets [3]. A comparison of sense division in the USL and a monolingual dictionary of Ukrainian [22] for the words discussed above, which are typical representatives of their respective classes, shows that the USL has fewer senses: 4 vs. 9 for the noun; 3 vs. 9 for the adjective; 5 vs. 6 for the two adverbs; 4 vs. 19 for the verb. (We have not counted here the shades of meaning, which are singled out within senses in [22] and may be treated as separate senses in some formal systems.) Thus, the coarse-grained semantic annotation in the USL leads to a reduction in the number of senses by a factor of approximately 2.5. Admittedly, this is just a very small part of the vocabulary, but it is indicative of the overall tendency. A bigger sample that includes words from various frequency bands needs to be analyzed for a more precise evaluation.

An important consideration for producing any sizable lexicon is organizing the work of its compiler(s) in an efficient manner. One way to achieve efficiency in the compilation of the USL is to leverage the already provided semantic annotations and apply them, *mutatis mutandis*, to similar items. This has been done for aspectual pairs of verbs as the imperfective and perfective forms of the same verb often require identical semantic description. For example, *вдатися* 'to succeed; to resort to; etc.' was annotated as one of the top 1,000 lemmas and its tags were then copied to its imperfective counterpart *вдаватися*, which occurs much less frequently. Both lemmas were added to the USL, and the procedure was repeated for all such cases. VESUM has more than 24,000 perfective verbal lemmas and a little over 20,000 perfective verbs, which promises substantial economy of effort in their semantic description. As a practical rule, aspectual verbal pairs are best annotated together regardless of the frequency of each of the two forms.

Furthermore, close-knit semantic clusters of words can be identified and annotated as a group, again cutting across the frequency continuum, as this will require few modifications in terms of semantic tags. The working hypothesis in the project has been that after high-frequency words, which are a diverse set that includes many basic-level categories, are semantically annotated, it will be possible to utilize them as precedents and rapidly describe numerous semantically similar lexical items. Considering that Ukrainian has an extensive system of prefixal derivation where prefixes are often employed to convey fine nuances of meaning without affecting the core semantic import of a word, this hypothesis was tested on two sets of prefixed verbs. One is associated with the core verb *думати* 'to think' annotated as **1:ment:noncaus**, which means that it has one sense related to mental activity (**ment**) and is noncausative. Incidentally, the monolingual dictionary used as a reference source for semantic description [22] singles out five senses and several shades of meaning for this verb, each specifying a slightly different aspect of thinking. The "thinking group" comprises 49 words and includes verbs with additional suffixes, the reflexive postfix -*ся*, and aspectual pairs. Forty of them were ultimately assigned the same tag sequence as the core verb; eight others were annotated with **ment&create:caus** (mental activity combined with creation, causative in nature, e.g., *видумати* 'to make up, invent, devise smth.'); two have the additional sense **ment&end:caus**, which points to the completion of the thinking process (e.g., *додумати* 'finish thinking'). Altogether, this group of prefixed verbs exhibits an extremely high degree of semantic homogeneity and shares the common **ment** tag in all senses. The effort invested in the description of one core lexical item has been leveraged to great effect.

Even where homogeneity is lower and variation higher, a group of words may be worth approaching as a semantic cluster. One major reason is that in this case semantic distinctions come to the fore move vividly, facilitating the work of the compiler and contributing to the high accuracy of the description. One such group consists of verbs derived from the basic verb *дивитися* 'to look; to consider; etc.'. In the USL, two senses are recognized for this verb, the primary one referring to perception and the second one to mental consideration: **1:percept:noncaus:2:ment:noncaus**. To compare, the monolingual dictionary [22] singles out as many as 11 senses and 17 shades of meaning for this verb. Its perfective counterpart *подивитися* also belongs to the top 1,000 words in the USL and has the same annotation. All other derivatives, 29 in total, are less frequent. Five of these copy the exact semantic annotation of the basic verb; nine others have only the first, perceptual sense. Several others combine the perceptual sense with an additional feature (**percept&end**, **percept&limit**) or merge the two senses (**percept&ment** as in *піддивитися* 'to peep, to spy'). Finally, four verbs exhibit additional senses that depart from the semantics of *дивитися*. The advantage of having this group considered as a whole has been evident: similar senses were rapidly annotated; distinct senses became more transparent in contrast; the overall speed and efficiency were higher than if these verbs were described separately, possibly with a significant period of time in between. This method is highly advisable for processing semantic clusters of words.

The level of semantic granularity adopted in the USL leads to a compact semantic description with a relatively low degree of ambiguity. The USL presently contains 1,192 entries that have a total of 2,147 senses. The average number of senses per entry is, therefore, 1.8. Nearly half of the USL entries (594) are unambigous as they have only one sense; 31% (364) have two senses, 12% (147) have three, and 7% (87) have four or more senses. The majority of these words belong to the high-frequency vocabulary, which is known for extensive semantic structures. Moving down the frequency list, the number of average senses per entry is expected to decrease.

5. Practical implementation

The system of semantic annotation described here has been implemented in the General Regionally Annotated Corpus of Ukrainian (GRAC) [20]. GRAC is constantly developing and has the parameters that make it the most diverse corpus of Ukrainian: 652 million tokens; more than 80,000 texts in various genres; 20,000 authors; texts published from 1816 to 2020; different spelling systems. All versions of GRAC have been tagged using VESUM, a large morphological dictionary of Ukrainian [17], and the TagText tagger. Both resources are available for download and non-commercial use.

VESUM is a large dictionary with a wide coverage of the Ukrainian word stock, including proper names, abbreviations, non-standard wordforms and lemmas, slang, alternative spellings, and dialect and archaic words [23]. In version 5.2.5., it has over 402,000 lemmas from which more than 6.5 million wordforms are generated. VESUM has a dynamic tagging module to cover out-of-vocabulary items and a toolkit for the morphological analysis of Ukrainian.

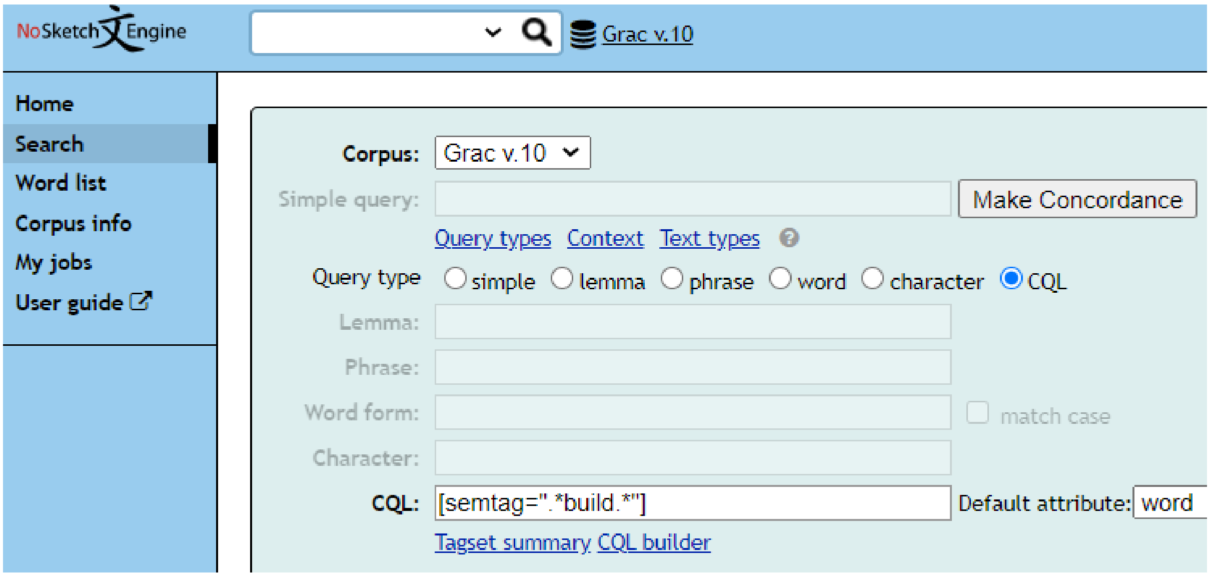
The TagText tagger, which is part of the LanguageTool API NLP UK project [16] comprising several utilities for processing Ukrainian-language texts, was updated to enable semantic annotation to be performed in one pass together with morphological annotation. Semantic annotation of any Ukrainian-language text can now be carried out by running TagText with the following one-line command:

```
groovy TagText.groovy -i <input text file> -e
```

The -i key points to the input text file, while the -e key adds the semantic annotation layer to the output text. Each time TagText is run with the -e key, it draws the semantic data from the current version of the Ukrainian Semantic Lexicon stored on GitHub [25] and assigns semantic tags to every word in the text which has a corresponding entry in the USL.

Using the USL data available at the time (USL 1.0, October 2020) and the updated TagText tagger, semantic annotation was added to every lexical token expressing 1,000+ most frequent Ukrainian lemmas (nouns, verbs, adjectives, and adverbs) in version 10 of the GRAC corpus. The strategic decision to populate the USL in the order of frequency made it possible to achieve high text coverage with numerically modest initial input as discussed in section 6. Naturally, as the USL grows, the contribution of each subsequent annotated lemma to the total output (coverage of text by semantic annotation) will follow a downward curve according to the law of diminishing returns.

An example of a simple semantic search query in GRAC-10 via the Manatee/Bonito interface is provided in Figure 1.



Corpus: Grac v.10

Query type ○ simple ○ lemma ○ phrase ○ word ○ character ● CQL

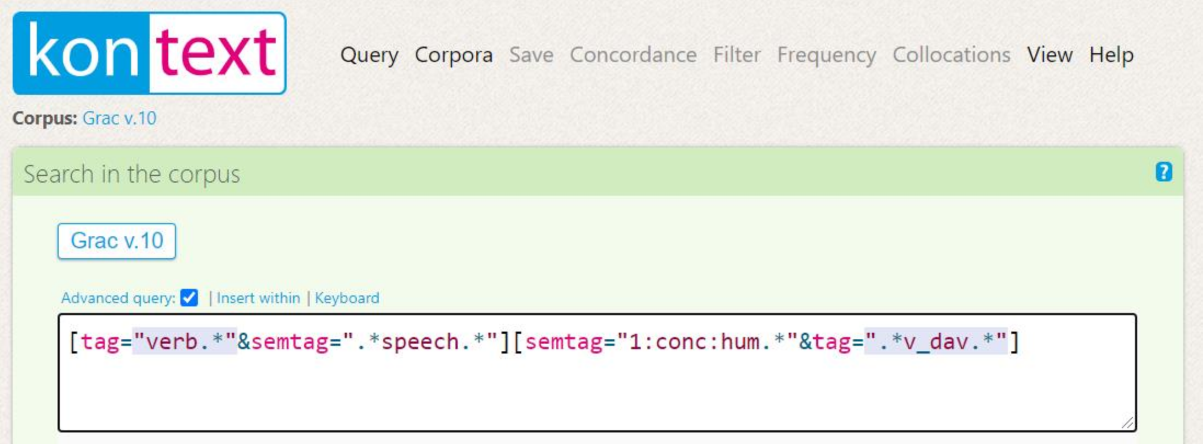
CQL: [semtag=".*build.*"]

Default attribute: word

Figure 1: A simple search query in GRAC-10.

This query searches for the semantic feature **build** occurring anywhere in tag sequences. It will match the following semantic tags, all assigned to concrete nouns: **conc:build** (buildings and constructions), **conc:org&&build** (organizations and buildings), and **conc:build:part** (parts of buildings). In GRAC-10, the query will pick out the following lemmas (in all their wordforms): *база* 'base', *будинок* 'building', *дім* 'house', *інститут* 'institute', *квартира* 'apartment', *клуб* 'club', *комплекс* 'complex', *кімната* 'room', *лікарня* 'hospital', *музей* 'museum', *міністерство* 'ministry', *пункт* 'post, station', *суд* 'courthouse', *театр* 'theater', *університет* 'university', *установа* 'establishment', *хата* 'house', *церква* 'church', and *школа* 'school'.

Figure 2 provides an example of a complex (morphological and semantic) search query run in GRAC-10 via the KonText interface:


kon text
Query Corpora Save Concordance Filter Frequency Collocations View Help
Corpus: Grac v.10
Search in the corpus
Grac v.10
Advanced query: ☑ | Insert within | Keyboard

```
[tag="verb.*"&semtag=".*speech.*"][semtag="1:conc:hum.*"&tag=".*v_dav.*"]
```

Figure 2: A complex search query in GRAC-10.

This query matches verbs of speech immediately followed in text by concrete nouns in the dative case that refer to individuals or groups of people in their primary sense, e.g., *показати народові* 'to show to the people' and *скажуть синові* '(they) will tell the son'.

Purely semantic searches allow researchers, among other things, to study the properties of the entire class of words subsumed under a specific tag. Owing to the uniform description of similar semantic content across parts of speech, it is possible to explore entire semantic domains, such as color, regardless of whether it is expressed by adjectives, adverbs, nouns, or verbs. Complex searches afford scholars studying Ukrainian a unique opportunity to analyze the linguistic behavior of semantically related words in an efficient and convenient manner.

6. Lexical coverage

In order to estimate the lexical coverage by the current version of the USL, data has been taken from version 10 of the GRAC corpus, which is the version where semantic annotation was implemented for the first time. Since semantic tags can only be assigned to words (tokens that begin with a letter of the alphabet), lexical coverage is calculated here with respect to words only. GRAC-10 contains some 506 million words. Using frequency lists automatically generated by the Manatee corpus management tool for GRAC and adding up the absolute frequencies of lemmas described in the USL, it has been established that USL entries account for a total of 167 million words (33%) in GRAC-10. Thus, semantic annotation covers one third of the corpus. However, this figure requires some correction. One factor here is that the semantic lexicon contains 12 ambiguous words (*все, усе, де, коли, його, так, також, тому, ще, як, його, її*) that are used in the corpus also as other parts of speech that are outside the scope of semantic annotation. For example, *коли* (when) can function as an adverb, conjunction, or particle. In GRAC, it is not disambiguated and occurs a total of 1.2 million times. In the semantic lexicon, this word is annotated as an adverb of time, and the **time** tag is applied to all of its 1.2 million occurrences in the corpus. The 12 words in this group account for roughly 3% of the total words, so the lexical coverage figure needs to be reduced by this amount. Once the corpus is morphologically disambiguated, it will be possible to apply semantic annotations only to the appropriate parts of speech and make more precise calculations.

One further correction, in the opposite direction, has to be made regarding proper nouns. The identification of proper nouns is considered to be part of semantic annotation. However, they are assigned the **prop** tag in the morphological, rather than semantic, layer in GRAC for historical reasons—this tag has been used in the morphological dictionary VESUM for years now. GRAC-10 contains 22.4 million (4.4%) instances of proper nouns, and this amount needs to be added to the previously obtained figure. Thus, the lexical coverage of GRAC-10 by semantic tags is above 34%.

There are three more tags (**number**, **date**, and **time**) that are semantic in nature and are part of GRAC's annotation: they occur 8.2 million times (1.6%) in version 10 of the corpus. However, these are assigned by the tagger to numerical text tokens (non-words) and do not contribute to our calculation of lexical coverage.

Following the approach adopted in the project, the entire Ukrainian lexicon is divided into two parts, one of which (nouns, verbs, adjectives, and adverbs) is subject to semantic annotation, while the other (all other parts of speech) is not to be semantically tagged. This latter part includes such high-frequency classes as pronouns, conjunctions, and particles. The 60 most frequent lemmas from this group account for 30% of lexical tokens in GRAC-10, which points to the theoretical upper limit of coverage by semantic tags—70% of words in the corpus. The real limit is, no doubt, lower due to, inter alia, misspellings and foreign words present in the corpus, as well as the incomplete lemmatization of valid forms.

7. Conclusions and future work

The semantic tagset for Ukrainian has been revised and now consists of a total of 229 taxonomic tags across four word classes, as well as additional tags for mereology, topology, evaluation, and causativity. The tagset is available online for reference purposes. It may be further refined in the future, but the overall classification scheme is now stable. The semantic tagset is transferable, with the necessary modifications, to other languages.

The Ukrainian Semantic Lexicon has been filled with the initial selection of 1,000 lemmas picked in the order of frequency and complemented with less frequent but semantically similar lexical units. The annotated part of the USL provides a solid ground for further work. The level of granularity chosen for semantic annotation leads to a significant decrease in polysemy as compared to monolingual dictionaries. USL entries have an average of 1.8 senses. The current version of the USL is available online for reference and download.

The TagText tagger has been modified to draw semantic data from the USL and perform both morphological and semantic annotation in one pass. The updated version has been made available for download, enabling users to semantically tag their Ukrainian texts. Semantic annotation has been implemented in the General Regionally Annotated Corpus of Ukrainian (version 10) with the help of TagText and using USL data. Lexical coverage by semantic tags in the corpus is over 34% with the theoretical upper limit being close to 70% as some of the most frequent vocabulary items are outside the scope of semantic annotation. It is now possible to construct purely semantic, as well as complex (morphological and semantic) search queries in GRAC via both the Manatee/Bonito and KonText interfaces.

The USL will be enlarged following the approach adopted for the initial phase. By adding the most frequent vocabulary first, it will be possible to rapidly increase lexical coverage in GRAC and other Ukrainian texts. At the same time, annotating semantically similar clusters of words has proved to be an efficient method. Further ahead lies the challenge of semantic disambiguation, which should be preceded by morphological disambiguation in GRAC.

8. Acknowledgements

This project was supported by a research grant from the Believe in Yourself Charitable Foundation at the Ukrainian Catholic University.

9. References

[1] A. Anisimov, O. Marchenko, A. Nikonenko, E. Porkhun, V. Taranukha. Ukrainian WordNet: Creation and Filling, in: FQAS 2013: Proceedings of the 10th International Conference on Flexible Query Answering Systems, volume 8132 (2013): 649–660. doi:10.1007/978-3-642-40769-7_56

[2] M. Ciaramita, S. Sloman, M. Johnson, E. Upfal. Hierarchical Preferences in a Broad-Coverage Lexical Taxonomy, in: Proceedings of CogSci (2005): 459–464.

[3] M. Ciaramita, Y. Altun, Y. (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger, in: EMNLP (2006): 594–602.

[4] Corpus of the Ukrainian Language. 2021. URL: http://www.mova.info/corpus.aspx

[5] Nataliia Darchuk. Computational Linguistics / Kompiuterna linhvistyka. Kyiv University, Kyiv, 2008.

[6] N. Darchuk. Kompiuterne anotuvannia ukrainskoho tekstu: rezultaty i perspektyvy [Computational Annotation of Ukrainian Texts: Results and Prospects]. Osvita Ukrainy, Kyiv, 2013.

[7] N. Darchuk Mozhlyvosti semantychnoyi rozmitky korpusu ukrainskoyi movy (KUM) [Possibilities of the Semantic Markup of the Corpus of the Ukrainian Language (KUM)]. Naukovyi chasopys Natsionalnoho pedahohichnoho universytetu im. M.P. Drahomanova [Academic Journal of the Drahomanov National Pedagogical University]. Seriya 9: Suchasni tendentsiyi rozvytku mov [Series 9: Modern Trends in Language Development], 15 (2017): 18–28.

[8] N. Darchuk, O. Zuban, M. Langenbakh, Ya. Khodakivska. AHAT-semantyka: semantychna rozmitka Korpusu ukrainskoi movy [AGAT-Semantics: Semantic Annotation of the Corpus of the Ukrainian Language]. Ukrainske movoznavstvo [Ukrainian Linguistics] 1(46) (2016): 92–102.

[9] FrameNet. 2021. URL: http://framenet.icsi.berkeley.edu

[10] O. Kanishcheva. WordNet-Affect-UKR. 2021. URL: https://github.com/olgakanishcheva/WordNet-Affect-UKR

[11] I. Kulchytsky, A. Romaniuk, Kh. Khariv. Rozroblennia Wordnet-podibnoho slovnyka ukrainskoi movy [Developing a WordNet-like Dictionary of Ukrainian], in: SISN 673.1 (2010): 306–318.

[12] G. I. Kustova, O. N. Lyashevskaya, E. V. Paducheva, E. V. Rakhilina. 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, 2005, pp. 155–174.

[13] S. Piao, F. Bianchi, C. Dayrell, A. D'Egidio, P. Rayson. Development of the multilingual semantic annotation system. Proceedings of NAACL HLT 2015 (2015): 1268–1274.

[14] S. Piao, P. Rayson, D. Archer, F. Bianchi, C. Dayrell, M. El-Haj, R.-M. Jiménez, D. Knight, M. Kren, L. Löfberg, R. Muhammad Adeel Nawab, J. Shafi, Ph. Lee The, O. Mudraya. Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages, Proceedings of the 10th Edition of the Language Resources and Evaluation Conference (LREC2016) (2016): 2614–2619.

[15] E. V. Rakhilina, G. I. Kustova, O. N. Lyashevskaya, T. I. Reznikova, O. Ju. Shemanaeva. Zadachi i printsipy semanticheskoy razmetki leksiki v NKRJa [The Objectives and Principles of the Semantic Markup of Vocabulary in the Russian National Corpus], in: Nationalnyi korpus russkogo yazyka. Novyie rezultaty i perspektivy [Russian National Corpus. New Results and Prospects]. Saint Petersberg, 2009, pp. 215–239.

[16] A. Rysin. LanguageTool API NLP UK Project. 2021. URL: https://github.com/brown-uk/nlp_uk

[17] A. Rysin, V. Starko. Large Electronic Dictionary of Ukrainian (VESUM), 2005–2021. URL: https://github.com/brown-uk/dict_uk

[18] Semantic Dictionary of the Ukrainian Language. 2021. URL: http://www.mova.info/semvoc.aspx?l1=193

[19] S. Shekhovtsov, O. Petriv, D. Chaplinsky, V. Dyomkin. Sentiment Dictionary of Ukrainian. URL: https://lang.org.ua/uk/dictionaries/

[20] M. Shvedova, R. von Waldenfels, S. Yarygin, A. Rysin, V. Starko, M. Woźniak, M. Kruk et al., GRAC: General Regionally Annotated Corpus of Ukrainian. Kyiv, Lviv, Jena, 2017–2021. URL: http://uacorpus.org

[21] V. Shyrokov et al. Korpusna linhvistyka [Corpus Linguistics]. Dovira, Kyiv, 2005.

[22] I. Bilodid (Ed.) Slovnyk ukrainskoi movy [A Dictionary of the Ukrainian Language]. In 11 vols. Naukova Dumka, Kyiv, 1970–1980.

[23] V. Starko, A. Rysin. Velykyi elektronnyi slovnyk ukrainskoi movy (VESUM) iak zasib NLP dlia ukrainskoi movy [Large Electronic Dictionary of Ukrainian (VESUM) As an NLP Tool for Ukrainian], in: Halaktyka Slova. Halyni Makarivni Hnatiuk. [A Galaxy of Words. To Halyna Makarivna Hnatiuk]. Vydavnychyi dim Dmytra Burago, Kyiv, 2020, pp. 135–141.

[24] V. Starko, Semantic Annotation for Ukrainian: Categorization Scheme, Principles, and Tools. Computational Linguistics and Intelligent Systems. Proc. 4th Int. Conf. COLINS. Vol. I. (2020): 239–248.

[25] V. Starko, Ukrainian Semantic Lexicon. 2021. URL: https://github.com/brown-uk/dict_uk/tree/master/data/sem

[26] UCREL Semantic Analysis System (USAS). 2021. URL: http://ucrel.lancs.ac.uk/usas

[27] WordNet. 2021. URL: http://wordnet.princeton.edu