On WordNet Semantic Classes: Is the Sum Always Bigger?

Tsvetana Dimitrova
Institute for Bulgarian Language "Prof. Lyubomir Andreychin"
Bulgarian Academy of Sciences
cvetana@dcl.bas.bg

Abstract

The paper offers an approach to the validation of the data resulted from a previous effort on expansion of WordNet noun semantic classes by mapping them with the semantic types within the Corpus Pattern Analysis (CPA) ontology employed by the framework of the Pattern Dictionary of English Verbs (PDEV). A case study is presented along with a set of conditions to be checked when validating the combined data.

Keywords: WordNet, semantic classes, Corpus Pattern Analysis, semantic types

1. Introduction

The present work discusses the validity of the results of an effort ((Koeva et al., 2018)) on enriching WordNet through expansion of the WordNet noun semantic classes by mapping the WordNet data ((Miller, 1990)) with the data in another resource – the semantic types within the Corpus Pattern Analysis (CPA) ontology that is used by the Pattern Dictionary of English Verbs (PDEV) (Hanks, 2004); (Hanks and Pustejovsky, 2005); (Hanks, 2008). The discussion builds on results of work described in (Koeva et al., 2018), (Koeva et al., 2019b), (Koeva et al., 2019a), where the PDEV verb patterns were further automatically mapped to WordNet sentence frames thus adding information about the character of the arguments – the resulting patterns are considered conceptual frames whose arguments were specified for a set of lexical units – the semantic types assigned to the WordNet noun synsets. As far as we are aware, there were no previous attempts at mixing WordNet and CPA ontology, although there are proposals at mixing up information in WordNet and other resources (Longman Dictionary of Contemporary English (Dorr, 1997); (Korhonen, 2002); VerbNet and FrameNet (Shi and Mihalcea, 2005); VerbNet and PropBank ((Pazienza et al., 2006).

The main relation among words in the lexical-semantic network WordNet ((Miller et al., 1993), (Fellbaum, 1998)) is synonymy (or near-synonymy; synonyms are defined as words which denote the same concept and are interchangeable in many (but not all) contexts). The synonyms (called 'literals') are grouped into unordered sets (synsets) which are linked via the so-called 'conceptual relations'. Most relations between synsets connect words of the same part-of-speech (POS). Noun synsets are linked via hypernymy / hyponymy (superordinate) relation, and meronymy / holonymy (part-whole) relation. Verb synsets are arranged into hierarchies via hypernymy / hyponymy relation. Adjectives are organised in terms of antonymy and similarity, and relational adjectives ('pertainyms') are linked to the nouns they are derived from. Adverbs are linked to each other via similarity and antonymy relations.

In addition, nouns in WordNet are organised within the superordinate / subordinate (hypernymy / hyponymy) hierarchy which is limited in depth. Distinguishing features are added to create lexical inheritance system where each word inherits the distinguishing features (attributes (modification), parts (meronymy), functions (predication) from its superordinates (Miller, 1990). An example would be {chef:1}[1], which, as a hyponym of {cook:6}, could be an Agent of the verb synsets {cook:1}, {cook:3},

[1]Throughout the paper, the numbers of the literals follow those applied in the database used by the viewer Hydra available at: http://dcl.bas.bg/bulnet/. We do not give all literals and definitions due to space limitation. There may be changes to semantic

and {cook:4} just like its hypernym.
Noun synsets in WordNet are classified into 26 semantic classes (primitives or primes, (Miller, 1990)), namely nouns denoting humans (*noun.person*), animals (*noun.animal*), actions (*noun.act*), feelings and emotions (*noun.feeling*), etc.
The present paper discusses possible ways to check the validity of the data resulted from a previous effort to enrich the data in WordNet through expansion of the noun semantic classes by merging the WordNet data with the data in the CPA hierarchy used in the Pattern Dictionary of English Verbs (PDEV). The WordNet was enriched through merging the WordNet concepts and the Corpus Pattern Analysis (CPA) semantic types. The 253 CPA semantic types were mapped to the respective WordNet concepts. As a result of the mapping, the hyponyms of a synset to which a CPA semantic type was mapped, inherit not only the respective WordNet semantic class but also the CPA semantic type. The resources and the mapping are described in section (2) following (Koeva et al., 2018) – the resulting data on which the discussion in section (3) is based on, is publicly available at: http://dcl.bas.bg/PWN_CPA/ [2].

2. The mapping

2.1. WordNet noun hierarchy vs. CPA ontology

As already mentioned, noun synsets in WordNet are organized into the following 26 semantic classes (primitives or primes, (Miller, 1990)) – given in Table 1:

Nouns denoting	**semantic class**	**Nouns denoting**	**semantic class**
humans	*noun.person*	animals	*noun.animal*
plants	*noun.plant*	foods and drinks	*noun.food*
actions	*noun.act*	natural processes	*noun.process*
feelings and emotions	*noun.feeling*	cognitive processes	*noun.cognition*
spatial position	*noun.location*	time and temporal relations	*noun.time*
man-made objects	*noun.artifact*	natural objects	*noun.object*
body parts	*noun.body*	substances	*noun.substance*
possession and transfer of possession	*noun.possession*	quantities and units of measure	*noun.quantity*
relations between people or objects or ideas	*noun.relation*	natural phenomena	*noun.phenomenon*
groupings of people or objects	*noun.group*	two and three dimensional shapes	*noun.shape*
goals	*noun.motive*	communicative processes and contents	*noun.communication*
natural events	*noun.event*	attributes of people and objects	*noun.attribute*

Table 1: WordNet noun semantic classes.

The synsets labeled *noun.Tops* are the top-level synsets in the hierarchy – these are the so-called unique beginners which divide the nouns into (sub-)hierarchies as illustrated in Table 2:

classes between the PWN and the version on http://dcl.bas.bg/bulnet/, for detail see (Leseva et al., 2015).

[2] To every noun <SYNSET> element there are <CPA> elements assigned.

noun.Tops	**hyponyms**	**semantic class of the hyponym(s)**
{entity:1}	{physical entity:1} {abstraction:1; abstract entity:1} {thing:4}	*noun.Tops* *noun.Tops* *noun.artifact*
{physical entity:1}	{thing:1} {object:1; physical object:1} {causal agent:1; cause:1; causal agency:1} {matter:1} {substance:7} {process:1; physical process:1}	*noun.Tops* *noun.Tops* *noun.Tops* *noun.substance* *noun.substance* *noun.process*
{thing:1}	*varying...*	*noun.object*
{object:1; physical object:1}	*varying...*	*noun.object*, *noun.artifact*
{causal agent:1; cause:1; causal agency:1}	*varying...*	*noun.person*, *noun.phenomenon*, *noun.state*, *noun.object*, *noun.substance*
{matter:1}	*varying...*	*noun.substance*, *noun.object*
{abstraction:1; abstract entity:1}	{psychological feature:1} {attribute:1} {group:1; grouping:1} {relation:1} {communication:1} {measure:1; quantity:1; amount:1} {otherworld:1} {set:41}	*noun.attribute* *noun.attribute* *noun.group* *noun.relation* *noun.communication* *noun.quantity* *noun.cognition* *noun.group*

Table 2: (Sub-)hierarchies under *noun.Tops*.

Different entities may inherit information for their features from different sub-hierarchies as they may have more than one hypernym, as in (1) where the two hypernyms are members of two sub-hierarchies which can be followed down to two main opposite concepts – {physical entity:1} and {abstract entity:1}:

(1)
{substance:1} *noun.substance*
hypernym: {matter:1}
...hypernym: {physical entity:1}
hypernym: {part:18; portion:7; component part:1; component:3; constituent:6} *noun.relation*
...hypernym: {abstract entity:1}

The second resource that was mapped, is the PDEV framework which employs the so-called semantic types which are members of the Corpus Pattern Analysis (CPA) ontology. The CPA semantic types refer to properties shared by a number of nouns that are found in the argument positions of verb patterns

and are formulated if repeatedly observed in verb patterns in a corpus – so these are corpora-based. The CPA semantic types are organized into a relatively shallow ontology (up to 10 sublevels for some types), where the top-level type is [Anything] which has six (sub)types: [Entity], [Eventuality], [Group], [Part], [Property], and [Not_Connected]. These are further subcategorised as follows:

Type	**Subtypes**
[Entity]:	[Abstract_Entity], [Energy], [Physical_Object], [Particle], [Self]
[Eventuality]:	[Event], [State_of_Affairs]
[Group]:	[Human_Group], [Vehicle_Group], [Animal_Group], [Physical_Object_Group]
[Part]:	[Language_Part], [Music_Part], [Physical_Object_Part], [Speech_Act_Part], [Document_Part], [Movie_Part], [Recording_Part]
[Property]:	[Cognitive_State], [Role], [Visible_Feature], [Character_Trait], [Injury], [Institution_Role], [Pace], [Use], [Weight]
[Not_Connected]	

Each semantic type inherits the formal property of the type above it in the hierarchy ((Cinkova and Hanks, 2010)). The CPA semantic types represent cognitive concepts in the context of their use but they are not linked to sets of concrete concepts and their lexical representations – this is achieved through mapping the CPA with WordNet.

2.2. Mapping CPA ontology and WordNet noun hierarchy

The WordNet noun synset hierarchy was mapped onto the semantic type hierarchy in the CPA ontology by matching the CPA semantic types with the WordNet synsets. The matching was done manually – the most probable (according to the definition) and populated (i.e., having the most hyponyms) synset was matched to a CPA semantic type by two independent annotators with a third annotator validating the cases of disagreement; the resulting assignments are inherited along the whole hypernym / hyponym 'tree'. The semantic types borrowed from the CPA ontology were added – as complementary semantic primitives – to the WordNet semantic classes (the process is described in (Koeva et al., 2018)).
As a result, the hyponyms of a synset to which a CPA semantic type is mapped, are labeled by the respective WordNet semantic class and the CPA semantic type as in (2) – the assigned CPA semantic type [Artifact] encodes the information that the New York State Barge Canal is an artificial system:

(2)
{New York State Barge Canal:1} 'a system of canals crossing New York State and connecting the Great Lakes with the Hudson River and Lake Champlain' *noun.location*
[Artifact], [Watercourse], [Waterway]

The 253 CPA semantic types are mapped to the respective WordNet concept (synset) with: (a) 199 semantic types mapped directly to one concept; (b) 39 semantic types mapped to two WordNet concepts ([Route] is mapped to {road:2; route:4} 'an open way (generally public) for travel or transportation', *noun.artifact*, and {path:3; route:5; itinerary:3} 'an established line of travel or access', *noun.location*); (c) 12 semantic types are mapped to three concepts; 2 semantic types are mapped to four concepts; and 1 semantic type is mapped to five concepts. Not each CPA semantic type can be mapped to one synset but the hyponyms of the respective nodes in the WordNet hierarchy inherit the semantic specifications of the specific class.

It is assumed that the concepts in WordNet are divided into {abstract entity:1} and {physical entity:1} [3], thus the CPA types are marked as follows (matching the CPA subtypes in the respective sub-hierarchies with probable noun synset(s), which are linked to either of the two *noun.Tops*; some types – [Energy], [Self], [Event] – involve subtypes that are matched to WordNet concepts that can be traced back to both {abstract entity:1} and {physical entity:1}):

CPA	WordNet
[Entity]	entity:1
[Abstract_Entity]	abstract entity:1
[Energy]	abstract entity:1, physical entity:1
[Physical_Object]	physical entity:1
[Particle]	physical entity:1
[Self]	abstract entity:1, physical entity:1
[Eventuality]	abstract entity:1
[Event]	abstract entity:1, physical entity:1
[State_of_Affairs]	abstract entity:1
[Group]	abstract entity:1
[Part]	abstract entity:1
[Language_Part]	abstract entity:1
[Physical_Object_Part]	physical entity:1
[Property]	abstract entity:1

As a result, the new semantic types borrowed from the CPA ontology were added to the WordNet structure as complementary semantic classes. The semantic types of the hypernym were inherited by its hyponyms if the hyponyms was not assigned other semantic types – for example, {wine:4; vino:1} was assigned the types [Part], [Drug], [Abstract_Entity], [Wine], [Food] because there is the CPA type [Wine] while its co-hyponym {home brew:1; homebrew:1} was assigned [Part], [Drug], [Abstract_Entity], [Alcoholic_Drink], [Food] types that are inherited from the hypernym {alcohol:1; alcoholic drink:1; alcoholic beverage:1}. However, certain errors and mismatches were found in the hypernym / hyponym structure under the top-level concepts as not every of their hyponyms instantiates another hypernym / hyponym tree.

3. Validation

Following the mapping of the CPA semantic types to the Wordnet noun hierarchy with the hyponyms inheriting the CPA type, we checked whether the assigned CPA semantic type was the correct one according to a number of conditions drawing upon already available data in WordNet. Since a synset may be assigned one or more CPA semantic types, an error may arise at each assignment.

3.1. Noun.foods

To illustrate, we will focus our attention on the noun synsets that refer to foods and drinks and are (mostly) labeled *noun.food* with the assigned CPA semantic type of [Food]. Table 1 gives the numbers of synsets assigned the type [Food] in a combination with other CPA semantic types.

[3] The third synset under the hypernym {entity:1} – {thing:4} which is classified as *noun.artifact* – comprises 8 synsets only.

CPA semantic type	**No**	**Examples**
[Stuff], [Food]	978	fast food, meal, wiener roast (*noun.food*)
[Natural_Landscape_Feature], [Stuff], [Food]	2	multivitamin, vitamin pill (*noun.food*)
[Part], [Abstract_Entity], [Stuff], [Food]	3	milk, mother's milk, colostrum (*noun.body*)
[Human_Group], [Abstract_Entity], [Stuff], [Food]	1	power breakfast (*noun.group*)
[Drug], [Stuff], [Food]	1	powdered mustard (*noun.substance*)
[Solid], [Food]	726	takeout, sugarloaf, quiche, cherry tomato (*noun.food*)
[Part], [Abstract_Entity], [Material], [Food]	2	fish meal, pantothenic acid (*noun.substance*)
[Part], [Abstract_Entity], [Beverage], [Food]	97	coffee substitute, chicory, cow's milk (*noun.food*)
[Part], [Drug], [Abstract_Entity], [Beverage], [Food]	2	elixir, elixir of life (*noun.food*)
[Part], [Drug], [Abstract_Entity], [Alcoholic_Drink], [Food]	164	malt liquor, kvass (*noun.food*)
[Part], [Abstract_Entity], [Water], [Food]	12	bottled water, sparkling water (*noun.food*)
[Solid], [Natural_Landscape_Feature], [Tree_Part], [Food]	197	edible fruit, strawberry, apple (*noun.food*)

Table 3: Foods and drinks.

The least populous combinations are the ones whose members are synsets of another semantic class, different from the expected *noun.food*.

In six combinations, there is an [Abstract_Entity] semantic type and in two – [Natural_Landscape_Feature] semantic type. In some cases, the [Abstract_Entity] is admissible as with 'power breakfast' and 'elixir' and 'elixir of life' while the [Natural_Landscape_Feature] can be applied to natural objects which is true in the case of fruit but is not applicable to the vitamil pill.

WordNet is heavily anthropocentric, thus {milk:4} 'produced by mammary glands of female mammals for feeding their young' is *noun.body* and [Part], [Abstract_Entity], [Stuff], [Food] but {milk:5} 'a white nutritious liquid secreted by mammals and used as food by human beings' is *noun.food* and [Part], [Abstract_Entity], [Beverage], [Food] – the same difference is kept between {mother's milk:1} and {cow's milk:1}. Here, the [Abstract_Entity] type is not appropriate, though.

Further, the vegetables are classified as solid foods, while fruit are solid foods which are parts of a tree – [Tree_Part] (including berries). For example, {edible fruit:1}, which is *noun.food* and [Solid], [Natural_Landscape_Feature], [Tree_Part], [Food], inherits its features (semantic types) from its two hypernyms – {produce:8; green goods:1; green groceries:1; garden truck:1} 'fresh fruits and vegetable grown for the market' which is *noun.food* and [Solid], [Food], and {fruit:5} 'the ripened reproductive body of a seed plant' which is *noun.plant* and [Natural_Landscape_Feature], [Tree_Part].

Drinks can be classified as abstract entities (with the semantic type [Abstract_Entity]) as many of them are man-made products. For example, {sparkling water:1} is *noun.food* and [Part], [Abstract_Entity], [Water], [Food] while {tap water:1} is also *noun.food* but [Part], [Abstract_Entity], [Material], [Water].

Further, there is {power breakfast:1} 'a meeting of influential people to conduct business while eating breakfast' which is *noun.group* but has two hypernyms whose semantic types it inherits: {breakfast:3} which is *noun.food* and {meeting:5; get together:5} which is *noun.group*. On the other hand, there is {dinner:2; dinner party:1} 'a party of people assembled to have dinner together' which has only one hypernym {party:3}, therefore the semantic types it inherits are only [Human_Group], [Abstract_Entity]. Such single instances should be taken into consideration.

3.2. Steps for validation

In order to check the correctness of the assigned CPA semantic types, we have to observe several conditions in relation to the data encoding information that is already available in WordNet, in the following order:

1. Check whether the WordNet semantic class is compatible with the assigned CPA semantic type, as in (3) where *noun.food* projects to [Food] but not in (4) where the semantic class is *noun.group*.

(3)
{dish:6} *noun.food* [Stuff], [**Food**] (True)

(4)
{power breakfast:1} *noun.group* [Human_Group], [Abstract_Entity], [**Stuff**] [**Food**] (False)

2. Check whether there are literals in the synset that are compatible with the assigned CPA semantic type, as in (5) where the literal is the same.

(5)
{alcohol:1; alcoholic drink:1; alcoholic beverage:1} *noun.food* [**Alcoholic_Drink**] [Food] (True)

3. Check whether the hypernym synset is assigned a CPA semantic type that is compatible, as these are inherited as in (6) where the *noun.food* {liqueur:1; cordial:1} inherits the type [Alcoholic_Drink] from its hypernym and transfers it to its hyponym {absinth:1; absinthe:1}.

(6)
{absinth:1; absinthe:1} *noun.food* [**Alcoholic_Drink**] [Food]
hypernym: {liqueur:1; cordial:1} *noun.food* [**Alcoholic_Drink**] [Food]
hypernym: {alcohol:1; alcoholic drink:1; alcoholic beverage:1} *noun.food* [**Alcoholic_Drink**] [Food] (True)

4. Check the inheritance along the hypernym / hyponym tree as in (7) where {mother's milk:1} inherits [Abstract_Entity] semantic type from its hypernym {milk:4} which, on its turn, has inherited it from one of its own two hypernyms. However, this semantic type is found much further down the tree and is probably not applicable here – such cases should be studied and eventually corrected.

(7)
{mother's milk:1} noun.food [**Abstract_Entity**] (False), [Stuff], [**Food**] (True)
hypernym: {milk:4} *noun.body* (with two hypernyms)
hypernym: {liquid body substance:1; bodily fluid:1; body fluid:1; humor:4; humour:4} *noun.substance* [**Abstract_Entity**] (False), [**Stuff**] (True)
hypernym: {body substance:1} *noun.body*
hypernym: {substance:1} *noun.substance* (with two hypernyms)
hypernym: {matter:1} *noun.substance*
hypernym: {physical entity:1}
hypernym: {part:18; portion:7; component part:1; component:3} *noun.relation*
hypernym: {relation:1} *noun.relation*
hypernym: {abstract entity:1} [**Abstract_Entity**]
hypernym: {nutriment:1; nourishment:2; nutrition:2; sustenance:2; aliment:2; alimentation:2; victuals:2} *noun.food* [Stuff], [Food]
. . . hypernym: {physical entity:1}

In all the data, there is persistent assignment of the semantic type of [Natural_Landscape_Feature], as in (8):

(8)
{paring:1} *noun.food* [Natural_Landscape_Feature]
{multivitamin:1; multivitamin pill:1} *noun.food* [Natural_Landscape_Feature], [Stuff], [Food]

In addition, there are other types of foods that are not classified with [Food] semantic type but with types referring to their source ([Meat]) or the fact that they are part of some other entity ([Quantity]) – see Table 4.

CPA semantic type	**No**	**Examples**
[Solid], [Meat]	198	sirloin steak, roast, confit (*noun.food*)
[Quantity]	193	wing, turkey wing, oyster, cutlet (*noun.food*)

Table 4: Foods not assigned [Food] type.

Information about the classification of these concepts as foods is already available from the WordNet semantic class. This means that somewhere down the hypernym / hyponym tree there is a synset which contains a literal that repeats the semantic class at hand (e.g., {substance:1} and {substance:2} have a semantic class of *noun.substance*, {person:1} has a semantic class of *noun.person*, {artifact:1; artefact:1} has a semantic class of *noun.artifact*, etc.).
The mapping can be additionally applied to the following pairs of WordNet semantic classes and CPA semantic types: *noun.substance* = [Stuff]; *noun.person* = [Human]; *noun.artifact* = [Artifact]; *noun.plant* = [Plant]; *noun.animal* = [Animal]; *noun.location* = [Location]; *noun.group* = [X_Group] ([Human_Group], [Physical_Object_Group], [Animal_Group], etc.), *noun.time* = [Time_Period], etc.

4. Conclusion

The paper discussed the the results of an effort on enriching the WordNet through expansion of the noun semantic classes by mapping the WordNet data with the semantic types within another corpus-based ontology within the Corpus Pattern Analysis (CPA). The validity of the results was checked on the basis of synsets about food and drinks, a couple of erroneous assignments was discussed along with the conditions behind these assignments based on the data in WordNet.
Although the CPA semantic types may add explicit information to WordNet semantics, this information is already available on different levels in the WordNet structure.

Acknowledgements

The work is funded under the project "Towards a Semantic Network Enriched with a Variety of Relations" (DN 10-3 / 14.12.2016), financed by the Bulgarian National Science Fund (BNSF).

References

Cinkova, S. and Hanks, P. (2010). *Validation of Corpus Pattern Analysis – Assigning Pattern Numbers to Random Verb Samples*. `http://ufal.mff.cuni.cz/project/spr/data/publications/annotation_manual.pdf`.

Dorr, B. J. (1997). Large-scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation*, 12(4):271–322.

Fellbaum, C., Ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Hanks, P. and Pustejovsky, J. (2005). A Pattern Dictionary for Natural Language Processing. *Revue française de linguistique appliquée*, 10 (3):63–82.

Hanks, P. (2004). Corpus Pattern Analysis. In *Proceedings of Euralex*, pages 87–97.

Hanks, P. (2008). Mapping Meaning onto Use: A Pattern Dictionary of English Verbs. In *Proceedings of the AACL 2008, Utah*.

Koeva, S., Dimitrova, T., Stefanova, V., and Hristov, D. (2018). Mapping WordNet concepts with CPA ontology. In *Proceedings of the 9th Global WordNet Conference (GWC'2018)*, pages 70–77.

Koeva, S., Dimitrova, T., Stefanova, V., and Hristov, D. (2019a). Towards Conceptual Frame. *Chuzhdoezikovo obuchenie*, 46(6):551–564.

Koeva, S., Hristov, D., Dimitrova, T., and Stefanova, V. (2019b). Enriching Wordnet with Frame Semantics. In *Dokladi ot Mezhdunarodnata godishna konferentsiya na Instituta za balgarski ezik Prof. Lyubomir Andreychin" (Sofiya, 14 – 15 may 2019 godina)*, pages 300–308.

Korhonen, A. (2002). Assigning Verbs to Semantic Classes via Wordnet. In *Proceedings of the 2002 Workshop on Building and Using Semantic Networks*, volume 11.

Leseva, S., Stoyanova, I., Todorova, M., Dimitrova, Tsvetana, R., and Koeva, S. (2015). Automatic classification of WordNet morphosemantic relations. In *The 5th Workshop on Balto-Slavic Natural Language Processing*, pages 59–64.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K., Eds. (1993). *Introduction to WordNet: an On-line Lexical Database. Five Papers on WordNet*. Princeton, NJ: Princeton University.

Miller, G. (1990). Nouns in WordNet: a lexical inheritance system. *International journal of Lexicography*, 3.4:254–264.

Pazienza, M. T., Pennacchiotti, M., and Zanzotto, F. M. (2006). Mixing wordnet, verbnet and propbank for studying verb relations. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*.

Shi, L. and Mihalcea, R. (2005). Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Computational linguistics and Intelligent Text Processing*, page 100–111.