Expanding and Enhancing Derivational and Morphosemantic Relations in Princeton WordNet

Ivelina Stoyanova
Institute for Bulgarian Language
Bulgarian Academy of Sciences
Sofia, Bulgaria
iva@dcl.bas.bg

Svetlozara Leseva
Institute for Bulgarian Language
Bulgarian Academy of Sciences
Sofia, Bulgaria
zarka@dcl.bas.bg

Gianina Iordăchioaia
University of Graz
Graz, Austria
gianina.iordachioaia@uni-graz.at

Verginica Barbu Mititelu
RACAI
Bucharest, Romania
vergi@racai.ro

Abstract

We propose enhancements of the Princeton WordNet data by expanding and enriching the collection of verb – noun derivational pairs with additional linguistic information. We focus on: (i) assigning an explicit derivational affix to the verb or noun involved in each of the derivational relations defined in Princeton WordNet (both for the pairs in the provided standoff file and for those additionally extracted); (ii) determining the derivational direction, especially for zero-derived pairs; (iii) providing a morphosemantic relation for the non-annotated derivationally-related pairs along the lines already implemented in the annotated collection made available in the standoff file.

The RESULTING DATASET includes 5330 cases of zero derivation (2964 zero_V and 2366 zero_N), 833 cases of direct noun-to-verb suffixal derivation, 15,801 cases of verb-to-noun suffixal derivation, as well as 1454 cases of either indirect derivation or derivation of both the verb and the noun from a third word.

1 Introduction

In this paper, we propose the enrichment of the Princeton WordNet (PWN) with morphological, derivational and semantic information for the verb – noun pairs that have an explicit derivational relation. Our objectives include: (i) enriching the derivational information in WordNet by assigning an explicit derivational affix to the verb and/or noun involved in each of the derivational relations starting with those pairs included in the PWN morphosemantic database made available as a standoff file and continuing with the derivationally related pairs additionally extracted from PWN; (ii) introducing the new feature of derivational direction within the description of the derivational relations, especially for zero-derived pairs; (iii) expanding the semantic relations by assigning morphosemantic relations to the identified derivationally related verb–noun pairs in PWN that have not yet been annotated with such a relation, along the lines adopted in the standoff file.

For the purpose of the work, we rely on several resources: (a) PWN (Miller et al., 1990b; Fellbaum, 1998), version 3.0, and the derivational relations defined in it; (b) the PWN morphosemantic database distributed as a standoff file[1] (Fellbaum et al., 2009), in which pairs of derivationally related nouns and verbs are labeled with one of 14 morphosemantic relations: Agent, Event, By-means-of, etc.; (c) the Oxford English Dictionary (OED) accessed through an API[2] and two lists of verb-to-noun zero-derived pairs and noun-to-verb zero-derived pairs provided by the OED team[3]; (d) lists of verbalising and nominalising suffixes compiled from theoretical literature, with a generalised invariant for each suffix.

Our contribution consists, first of all, in supplementing the derivational relations in WordNet with information about the nominalising and/or verbalising suffixes used in the derivation. The resulting dataset contains both originally provided verb–noun pairs supplied with morphosemantic relations from the standoff files, as well as additionally identified derivationally related pairs of literals. Second, not only pairs with overt suffixes, but also zero-derivation verb–noun pairs are supplied with information about the direction of derivation: verb-

[1] https://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls

[2] https://developer.oxforddictionaries.com/

[3] We thank James McCracken and Emily Hoyland (the OED team) for providing us with the lists of zero nouns and zero verbs and API access.

to-noun or noun-to-verb. In this way, we are able to analyse the zero verb-to-noun suffix separately from the zero noun-to-verb suffix, and study the two comparatively. Moreover, we have attempted to develop methods for automatic detection of the direction of derivation based on lexicographic information from the OED (year of attestation, number of senses of the verb and the noun, frequency of usage). Third, we expand the number of pairs labeled with a morphosemantic relation by assigning them to the derivationally related pairs extracted from PWN.

Various subsets of the dataset can be used for a number of tasks: morphological and derivational analyses, validation of hypotheses on suffixal and zero-derivation, morphosemantic and other verb–noun relations in WordNet. We discuss our considerations on how such a dataset can help in identifying regular polysemy and in polysemy resolution. (Chalub et al., 2016) have shown that adding morphosemantic relations to a wordnet (in their case to the Portuguese Open WordNet (de Paiva et al., 2012)) helps to improve its quality.

Section 2 provides a summary of the recent research in derivational morphology, making a case for the need for large-scale empirical data as a testing ground for linguistic hypotheses. Section 3 describes the challenges and motivates the use of the PWN as a reliable resource in such a task. Section 4 centres on the methodology used in the annotation of the data, starting with how the derivational and semantic information in PWN can be combined towards the enhancement of the initial datasets, as well as outlining the procedures for expanding the data with new levels of description. Section 5 presents the final dataset, while Section 6 discusses the possibility to deepen the analysis by exploring some features of the dataset and PWN structural organisation. We finally outline the next steps of our research in Section 7.

2 Related work

Although polysemy in derivation has long been assumed to follow the general patterns of polysemy with lexical words (Rainer, 2004), recent research in lexical semantics recognizes that derivational processes contribute their share to polysemy and that affix polyfunctionality, in particular, poses a real challenge for lexical semantics (Grimshaw, 1990; Plag, 1999; Lieber, 2004, 2016; Bierwisch, 2009; Melloni, 2011; Bauer et al., 2013; Iordăchioaia and Melloni, 2023; Salvadori and Huyghe, 2022; Kawaletz, 2023; Valera, 2023).

Among the early attempts to model polysemy in derivation and affix polyfunctionality we find Plag (1999), who investigates the productivity of the different affixes involved in verb derivation in English, by looking at neologisms and using the OED for their meanings. Plag uses relations such as locative, ornative, causative, resultative, inchoative, performative and similative to describe derived verbs and argues that *-ify* and *-ate* are phonologically conditioned allomorphs of *-ise*, the most representative suffix, which systematically incorporates and expresses all the relations above. One important claim that Plag makes is that the variety of meanings that conversion/zero may express is so large that there cannot be any specific meaning attached to it (unlike for *-ise*; see also Clark and Clark (1979); Lieber (2004)). He argues that at least relations such as instrumental, privative and stative need to be added to those above to describe zero-derived verbs (Valera, 2023, for discussion on these relations in verbal derivation). This implies that the meaning of verbalising zero should be less predictable than that of the suffix *-ise* in relation to the derivational base.

The unpredictability of zero for verbal derivation, especially from noun bases, has been further confirmed by a large corpus-based study with distributional semantic methods by Kisselew et al. (2016), who predict the derivational direction of zero noun-verb pairs from semantic specificity (building on the assumption that derived words are more specific than their bases) and apply measures of information content (Entropy and Kullback-Leibler Divergence) to distributional representations of verbs and nouns in English. Their results show that information content is a good predictor for zero nouns derived from verbs but not for zero verbs derived from nouns. This means that the zero verbalising suffix is not predictable from the meaning of the base, while the zero nominalising suffix is. This further entails that zero is not necessarily less predictable than overt suffixes, but the derivational direction is crucial. In addition, the study presented by Kisselew et al. (2016) also makes the case for the need for both large-scale corpora for testing the linguistic hypotheses and gold-standard datasets against which the testing may be carried out.

Other quantitative studies with a computational approach such as Varvara (2017); Varvara et al. (2021, 2022) investigate nominalising suffixes in

languages like Italian, German and French. Varvara (2017) focuses on event nominalisations in Italian and German with the aim to model the competition between different nominalising suffixes. The conclusion Varvara (2017) reaches is that the main difference between the competing nominalising suffixes in Italian is semantic, to the extent that they refer to different senses of their base verbs and disambiguate the vagueness of the base. For German, Varvara (2017); Varvara et al. (2021) argue that infinitive-based nominalisations are semantically more predictable from the base verb than those formed with the derivational suffix *-ung*, and that they are closer to inflectional processes like participial formation in this respect, while *-ung* is closer to proper derivation by means of the agentive suffix *-er*.

Varvara et al. (2022) develop and test an annotation scheme for derived nouns in context, by which they annotate the ontological type of their meaning and their relationship with the base (see this work also for further similar efforts on other languages). They annotate 4,500 corpus occurrences of 90 deverbal nouns ending with 6 different suffixes in French. For this, they develop an annotation sample that includes 23 ontological types (e.g. animate, artefact, event, etc.), 21 relational types (agent, beneficiary, result, etc.) and 62 complete types, the last of which is an attested combination of the other two. As a follow-up, these authors have enriched their work to include 42 derivational suffixes and 4 conversion forms in French (Varvara et al., 2024).

Lapesa et al. (2018) investigate the polysemy of deverbal nominalisations with the suffix *-ment* in English, by analysing a set of 55 types and 406 tokens extracted from the Corpus of Contemporary American English (Davies, 2010) and annotating them as eventive vs. non-eventive vs. ambiguous. To avoid possibly lexicalized readings and to extract productive polysemy patterns, they consider only low frequency neologisms of this kind. They build a distributional semantic model which can distinguish between eventive and non-eventive readings to a good extent and mostly needs only small context windows to identify the intended meaning. They find out that eventive readings are easier to classify than the non-eventive ones, which is usually due to the semantic similarity between abstract non-eventive and eventive nouns. While the study of Lapesa et al. (2018) does not make such fine distinctions as the one by Varvara et al. (2022), a question that arises for the latter is how viable the semantic distinctions they propose may be for other languages and how successfully they could be used for an automatic annotation.

Current research in derivational morphology, involving studies in both theoretical and computational or corpus linguistics, makes a strong claim for the need for constructing large-scale databases comprising rich derivational information that would enable reliable and possibly bias-free observations and conclusions.

The effort described in this paper and the resulting resource aim exactly at providing a database of derivationally and semantically related verb – noun pairs by building on an existing resource, increasing the amount of the data and enriching it with explicit suffixes and the direction of derivation.

3 Motivation and challenges

As previous literature shows (Plag, 1999; Fernández Alcaina, 2021; Lapesa et al., 2018; Varvara et al., 2022; Kawaletz, 2023; Lara Clares, 2023; Valera, 2023), work on the meaning and polyfunctionality of affixes involves large human annotation efforts. Unfortunately, such efforts often remain only locally exploited and are not further refined by other researchers because the methodology underlying the annotation is too specific to the purposes of the individual task. Indeed, these annotations are not free of possible bias due to the theoretical framework or practical assumptions adopted in the annotation guidelines. Ideally, human annotation efforts should target a more general lexical semantic description of a language, so that lexical semanticists would be able to employ such independently created and unbiased lexical resources to more objectively test their theoretical hypotheses.

This is precisely what the PWN project provides for English derivational morphology. WordNet is a rich lexical semantic resource that can be exploited to obtain new insights into affix polyfunctionality and develop automatic tools for disambiguation and polysemy resolution. Moreover, the PWN database was created on the basis of principles independent of the purpose of affix disambiguation, which makes it all the more reliable for such a task, unlike in the case of most databases created and annotated for this precise purpose.

Another advantage of the PWN is the organisation of the data along the concept of sense (represented by a synonym set) and not word. This allows a precise description and better un-

derstanding of the meaning relations that occur between derivationally-related words (in their different senses). In particular, the same or different morphosemantic relations may be found to hold between different senses of two derivationally-related words, a fine distinction that is not always captured by dictionaries. For instance, the verb *articulate* and the noun *articulation* occur as a pair several times and, depending on the meanings they enter the pairs, they are labeled with a different morphosemantic relation: Uses (for the meanings 'provide with a joint' and '(anatomy) the point of connection between two bones or elements of a skeleton (especially if it allows motion)' and for the meanings 'provide with a joint' and 'the shape or manner in which things come together and a connection is made'), Event (for the meanings 'speak, pronounce, or utter in a certain way' and 'expressing in coherent verbal form', for the meanings 'express or state clearly' and 'expressing in coherent verbal form', for the meanings 'provide with a joint' and 'the act of joining things in such a way that motion is possible' and for the meanings 'put into words or an expression' and 'expressing in coherent verbal form'), Property (for the meanings 'speak, pronounce, or utter in a certain way' and 'the aspect of pronunciation that involves bringing articulatory organs together so as to shape the sounds of speech').

Last but not least, the 14 morphosemantic relations in PWN are formulated to capture both derivational directions: from noun to verb and verb to noun. This leads to a more coherent and systematic ontology of the relations than in the previous literature, where, for instance, the diverse morphosemantic relations identified in verb formation (Plag, 1999; Valera, 2023) are entirely different from those in noun formation (Lieber, 2016; Varvara et al., 2022), without any evidence of the need to distinguish them. The PWN actually offers evidence for quite the opposite.

4 Methodology

The section presents the procedures for enhancing the original data supplied by the PWN project in terms of derivational and morphosemantic relations, by expanding their coverage, on the one hand, and adding new levels of description, on the other. We start with outlining some prerequisites that we rely on in the annotation procedures.

4.1 Prerequisites

We employ data derived from the PWN, and supplement it with information compiled from other resources.

Morphosemantic Database from PWN (MSD DATASET). This dataset is distributed by the PWN project[4] as a standoff file consisting of 17,739 derivationally related noun-verb pairs, which are labeled with a morphosemantic relation (such as Agent, Instrument, Event, etc.) (Fellbaum et al., 2009). Table 3 shows the set of morphosemantic relations with short definitions and examples.

Dataset of derivationally related verb–noun pairs in PWN (DERIV DATASET). This dataset covers 18,344 synset pairs in total, between whose members a derivational relation exists in PWN. Additionally, we identify automatically 23,418 pairs of literals within each pair of synsets which share a common base and use verbalising and nominalising suffixes.

In the DERIV DATASET there are 4,520 noun–verb pairs of literals that are not in the MSD DATASET and are linked by a derivational relation alone (i.e., with no morphosemantic relation assigned), e.g. *carbon* and *carbonate*.

Dataset from OED of zero derivation verb–noun pairs with the direction of derivation (OED ZERO DATASET). The dataset contains two lists derived from OED[5]:

- a list of 2,830 pairs of verb and noun OED lexical entries labeled as verb-to-noun zero derivation;
- a list of 5,921 pairs of verb and noun OED lexical entries labeled as noun-to-verb zero derivation.

Set of nominalising, verbalising and other suffixes (SUFFIX DATASET). This dataset comprises 7 verbalising and 26 nominalising suffixes in total, which are reduced to 5 and 11 invariant verbalising and nominalising suffixes, respectively. These represent direct noun-to-verb or verb-to-noun derivations, including also zero$_N$ and zero$_V$ suffixes.

There are further 36 suffixes which take part in more complex derivational patterns, such that both

[4] https://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls

[5] https://www.oed.com

the noun and the verb are derived from a third word (e.g., *rigidify – rigidness* both derived from the adjective *rigid*; *criminalise – criminal* both derived from the adjective *criminal*) or they are derived in several steps (e.g., *argue – argumentation* with an intermediary step *argument*; *attract – attractiveness* with an intermediary *attractive*).

4.2 Identifying the suffixes in derivationally related pairs of literals

For each derivational pair in the combined data from MSD DATASET and DERIV DATASET, we identify the following categories: (a) zero derivational pairs both for the noun to verb (*bottle* > *to bottle*) and the verb to noun (*to walk* > *the walk*) direction; (b) cases of overt verbalising suffixes such as *-ise*, *-ify*, *-ate*, etc. involved in the derivation of verbs from nouns; (c) cases of overt nominalising suffixes such as *-ion*, *-ing*, *-ment*, etc. involved in the derivation of nouns from verbs; (d) cases of other derivational models.

Suffix identification was performed automatically by finding the common base of the noun and the verb and the endings, matching them to known suffixes. The allomorphs of suffixes are mapped to the canonical form of each suffix. For instance, *-ion*, *-tion* and *-ation* are treated as variants of the same suffix *-ion*. In this way, we provide a unified and more complete representation in terms of the derivational pairs and the morphosemantic relations each (abstract) suffix is involved in.

In such a way, the productivity and the polyfunctionality of each suffix is made explicit along with the relative probability for a suffix to be the exponent of a given relation.

4.3 Direction of conversion and affixal derivation

The overt suffixes are clearly associated with a particular direction of derivation since they are either verbalising (e.g., *-ise, -ify*, etc.) or nominalising (e.g., *-ment, -ion, -ing*, etc.). In our data, 833 pairs of verb–noun literals are identified with a verbalising suffix, thus assigned a noun-to-verb direction; and 15,801 pairs of verb–noun literals are identified with a nominalising suffix, thus assigned a verb-to-noun direction.

The dataset contains 5,330 cases of zero derivation. For these the OED ZERO DATASET and the OED API Service are used to determine the derivational direction. Via the OED API we obtain the definitions of the verb and noun zero derivatives and match them semi-automatically (automatic matching with manual validation) to PWN senses based on lexical and semantic similarity between OED sense definitions and PWN glosses. The identification of the closest match of OED entries to PWN synsets allows us to transfer the information about the direction of derivation from the OED ZERO DATASET to the zero derivational verb–noun pairs in PWN.

Incorporating information on the direction in zero derivation verb–noun pairs allows us to distinguish between verbalising and nominalising zero affixes, given the important differences between them (Kisselew et al., 2016), and to treat them uniformly with overt suffixes. Directionality in zero derivation is a long-known non-trivial task, especially for languages such as English with barely any morphological evidence (Marchand, 1964; Kiparsky, 1982; Plank, 2010; Bram, 2011). To ensure a high reliability of the derivational direction, we coded it on the basis of the OED ZERO DATASET lists.[6] Building further on our dataset, future research on directionality in conversion will be able to employ the morphosemantic relations specific to a derivational relation as a further directionality criterion (Barbu Mititelu et al., 2023).

There are further 1454 verb–noun derivationally related pairs of literals exhibiting more complex derivational patterns. These cases have not been assigned a direction of derivation.

4.4 Newly assigned morphosemantic relations

We develop procedures to expand the morphosemantic relations annotation in PWN. For this purpose, we explore several semantic features in the PWN description.

We analyse the synset gloss of the derived word and search for certain triggers which signal a particular morphosemantic relation. For instance, a noun synset gloss beginning with "the act of..." or "an event of..." indicates the existence of an Event relation between the noun and the verb in the pair under observation. Triggers such as "an instrument for...", "a device for...", "an implement for...", etc. correlate with an Instrument relation, and so forth.

[6] To determine directionality, OED lexicographers have reportedly considered the full history of each word, including date of attestation, early frequency of use, but also linguistic and etymological factors such as the behavior of cognate words, the donor in case of loanwords, and semantic properties to the extent that the more basic meaning would be associated with the base word (Philip Durkin, p. c.); see also Plag (2003) for directionality tests.

Another feature used in a complementary manner with the gloss, especially when the latter does not provide sufficient information, is the semantic class (defined as a semantic primitive or prime) assigned to each noun or verb synset in PWN. Semantic primes as presented in Miller et al. (1990a) define language-independent semantic classes, in particular 25 noun classes, e.g. noun.person, noun.artifact, noun.act, and 15 verb classes, e.g. verb.emotion, verb.motion, verb.communication. For instance, when considering the morphosemantic relation to be assigned to the pair *sink* "go under" (e.g., *The raft sank and its occupants drowned*) – *sinking* "a descent as through liquid (especially through water)", the annotator should be prompted by the prime of the nominal synset (noun.event) and the association between the Event relation and semantic class noun.event.

A third feature employed in the assignment of the morphosemantic relations is the membership of the related synsets in certain subtrees in which a morphosemantic relation occurs (frequently) between other pairs. Consider the case of *nickel*:1 – *nickel*:1, *silver*:1 – *silver*:1, *copper*:1 – *copper*:1 and *chrome*:1 – *chrome*:1. The verb synsets are hyponyms of *cover*:1, while the nouns are hyponyms of *metal*:1, and all the pairs are assigned the morphosemantic relation Uses, which in this case may be interpreted as a relation between (i) a noun denoting a metal, which is used to cover a surface or a thing so that it acquires some quality of the metal, and (ii) a verb denoting the action of covering with the metal. There are other pairs in PWN, e.g.: *aluminium*:1 – *aluminize*:1, where the verb is a hyponym of *cover*:1 and the noun is a hyponym of *metal*:1, but their derivational relation is not labeled morphosemantically. The semantics of the relationship between the members of the unlabeled pair is very likely the same as for the labeled ones above. After inspecting the glosses, one can confirm with certainty that the pair *aluminium*:1 – *aluminize*:1 is an instance of the Uses relation.

The judgments were made by trained annotators according to a methodology which is based on the linguistic features outlined above.

These criteria are applied on their own or in combination, as well as in consideration of the suffix and its distribution across relations in the available data. For instance, with nouns derived with the suffix *-er* and having the prime noun.person, the relation is unambiguously determined as Agent, while for those with the semantic prime noun.artifact, the relation is most likely Instrument. This conclusion becomes self-evident if one considers the very strong correlation between these primes and the respective relations established for the suffix.

Suffix	Base	Derivation	#
zero$_V$	*fake*	*to fake*	2964
-ise	*agony*	*to agonise*	463
-ate	*acetyl*	*to acetylate*	219
-ify	*city*	*to citify*	151
-en	*threat*	*to threaten*	16
-ion	*to admit*	*admission*	4330
-er	*to adjust*	*adjuster*	3442
zero$_N$	*to glide*	*the glide*	2366
-ing	*to play*	*the playing*	1987
-ment	*to replace*	*replacement*	699
-ance	*to occur*	*occurrence*	367
-ant	*to pollute*	*pollutant*	159
-age	*to parent*	*parentage*	145
-al	*to dispose*	*disposal*	135
-ure	*to press*	*pressure*	108
-ee	*to train*	*trainee*	83
Other	*rigid*.adj	*to rigidify – rigidness*	1454

Table 1: Distribution of verbalising and nominalising suffixes in the RESULTING DATASET with examples

5 Results

The resulting resource, RESULTING DATASET, presents a comprehensive description of derivationally related pairs of verb–noun literals including the suffixes, the direction of derivation (whenever available), as well as semantic information in the form of an assigned morphosemantic relation between the noun and the verb. While based on the derivational relation, in essence, the morphosemantic relation is semantic and thus extends beyond the particular pair of literals, and holds between the corresponding verb and noun synsets.

Table 1 shows the distribution of the data for each derivational direction and each suffix.

Less productive and unproductive suffixes such as *-th* in *grow - growth*, have also been included in the RESULTING DATASET, as they provide valuable data about the frequency of the respective derivational processes, their representation in comparison with derivations with other suffixes and the roots involved in them.

Along with derivational pairs obtained in a single derivational step, we also preserve the ones that

involve a more complex process, such as a 2-step derivation or derivation from another base word. Although these pairs have no clear direction of derivation, a particular morphosemantic relation can be identified, and for this the direction is not mandatory.

The resulting resource also includes British/American spelling doublets such as *acclimatize – acclimatization* and *acclimatise – acclimatisation*, thus providing a fuller picture of the variants of English. In certain cases, e.g. for the purposes of machine learning and automatic identification of morphosemantic relations, these can be considered as duplicates and removed in order to avoid skew in the data.

Table 2 shows the association of each canonical suffix with different subsets of the 14 morphosemantic relations. While there is substantial ambiguity among suffixes, there are additional semantic features in PWN which can help to reduce ambiguity. For example, the suffix *-er* is associated with 12 out of the 14 morphosemantic relations, but the prevalent ones are Agent (in 76.8% of cases) and Instrument (in 11.7% of cases), which explains why *-er* is usually considered an agentive suffix and gives a strong indication as to the most likely relation. Moreover, when the noun semantic class is noun.person, it is associated with the agentive morphosemantic relation, while noun.artefact is associated with the instrument relation.

Table 3 presents the results of the expansion of the morphosemantic relations providing the initial number of the relations in MSD DATASET compared to the final number in the RESULTING DATASET. No relation is assigned in the cases where no suitable relation is identified among the set of 14 morphosemantic relations. The initial number of assigned relations has been increased by 20% up to 21,251.

The RESULTING DATASET (version 1.0) is distributed as a stand-alone resource that can be linked to PWN 3.0 or any other wordnet. The dataset is released under the Creative Commons Attribution-NonCommercial 4.0 International license.[7]

Suffix	# rel.	Morphosemantic relations
zero$_V$	12	Event (35.2%), Result (8.3%), By-means-of (8.3%)
-ise	9	Result (31.1%), By-means-of (15.2%), Event (10.5%)
-ate	11	Result (27.3%), Event (21.7%), By-means-of (16.8%)
-ify	8	Result (53.4%), By-means-of (18.4%), Uses (12.6%)
-en	4	Event (61.5%), Result (15.4%), Property (15.4%)
-ion	14	Event (71.3%), Result (9.7%)
-er	12	Agent (76.8%), Instrument (11.7%)
zero$_N$	12	Event (71.4%), Result (8.1%), By-means-of (8.1%)
-ment	12	Event (65.4%), State (11.2%), By-means-of (8.4%)
-ance	7	Event (67.3%), By-means-of (12.7%)
-ant	7	Agent (53.5%), By-means-of (14.9%), Material (8.9%), Uses (8.9%)
-age	9	Event (56.4%), Result (10.3%), Location (10.3%)
-al	6	Event (78.7%), Result (10.3%)
-ure	9	Event (51.0%), Result (14.3%), State (12.2%)
-ee	8	Undergoer (59.3%), Agent (19.8%), Destination (15.1%)

Table 2: Ambiguity of suffixes: number of morphosemantic relations (out of 14) covered by a particular suffix, and the most frequent of them.

6 Discussion

Regular polysemy is reflected in morphosemantic relations, especially since from a contemporary point of view a verb's sense may be considered related to more than one (closely) related noun senses or vice versa. Such an example is found with nouns of the class noun.artifact (mostly containers) and nouns denoting the quantity that the respective container holds, e.g. *barrel*:2, *cask*:2 ('a cylindrical container that holds liquids') and *barrel*:4, *barrelful*:1 ('the quantity that a barrel (of any size) will hold'). Each of the two synsets is related to *barrel*:1 ('put in barrels') by means of the relations Location and Undergoer, respectively. Regular polysemy reveals how regularities between related meanings in the nominal or the verbal domain are reflected in the semantics of the relation in verb-noun pairs.

Observations on structured parts of the lexicon, such as the ones discussed above, also enable us to predict missing relations, both morphoseman-

[7] https://github.com/WordNetMorhosemantics/SuffixalDerivation

Relation	Description	Example	Initial #	Result #
Agent	an entity that acts volitionally so as to bring about a result	*colonize – colonizer*	3,043	3,356
Body-part	a part of the body (e.g. of an Agent) involved in the situation	*extend – extensor*	43	46
By-means-of	something that causes, facilitates, enables the occurrence of	*decree – decree*	1,235	1,480
Destination	a recipient, an addressee or a goal	*patent – patentee*	17	19
Event	something that happens at a given place and time	*wash – washing*	8,158	10,015
Instrument	an object (rarely abstract) acting under the control of an Agent	*browse – browser*	813	875
Location	a concrete or an abstract place involved in the situation	*hospitalize – hospital*	288	394
Material	a substance or material used to obtain a certain effect or result	*polymerize – polymer*	114	116
Property	an attribute or a quality	*beautify – beauty*	318	486
Result	the outcome of the situation described by the verb	*syllabify – syllable*	1,439	1,784
State	an abstract entity, such as a feeling, a cognitive state, etc.	*anger – anger*	528	677
Undergoer	an entity affected by the situation described by the verb	*invite – invitee*	878	1006
Uses	a function an entity has or a purpose it serves	*cement – cement*	740	896
Vehicle	an artifact serving as a means of transportation	*fight – fighter*	87	101

Table 3: The set of morphosemantic relations and initial number of occurrences in the MSD DATASET compared to the final number in the RESULTING DATASET.

tic and derivational. Consider *jar*:5 ('place in a cylindrical vessel') and the noun synsets *jar*:1 ('a vessel (usually cylindrical') and *jar*:2, *jarful*:1 ('the quantity contained in a jar'). Although only the Undergoer relation is encoded, the Location relation is easily predictable on the basis of the *barrel* example above. Exploring further the hyponyms of the synset *containerful*:1 ('the quantity that a container will hold'), we discover that 25 out of its 67 hyponyms have corresponding verbs, but only 3 of the verbs are appropriately linked to the noun synsets denoting the respective quantity and artifact (in a like manner to *barrel*) – the remaining verbs lack one or both morphosemantic relations or even the derivational ones (e.g., *bag(ful) - bag*). In such a way, we are able to tackle the inconsistencies in derivational and morphosemantic relations throughout this and other parts of the PWN structure.

These observations can also be implemented into automatic procedures to discover morphosemantic and other semantic relations between single units in PWN, as well as between (sub)trees or semantic (sub)classes.

7 Conclusions and future work

We have presented here the creation of a large set of derivationally related word pairs, by enhancing the PWN morphosemantic database with new pairs of literals linked through derivation with the explicit suffixes involved, as well as with a morphosemantic relation. The RESULTING DATASET includes over 21,000 verb–noun pairs.

Our dataset seems to confirm previous insights according to which verbalising suffixes are more polyfunctional and less predictable than the nominalising suffixes (Kisselew et al., 2016; Barbu Mititelu et al., 2023). However, surprisingly, zero suffixes, although polyfunctional, are not less predictable than the overt ones, as claimed by Plag (1999); Lieber (2004). The zero and overt suffixes

exhibit similar levels of polyfunctionality, and the developed dataset provides material for detailed analysis into these claims.

One of the directions to be further investigated is the development and improvement of various procedures for automatic analysis, identifications of derivational or semantic features (e.g., the suffix, the direction of derivation, the morphosemantic relation, etc.), in order to increase the scope of the data and the depth of the analysis.

A natural extension of the work would be to study the derivational relations for other languages using the corresponding (aligned) wordnets and taking as a point of departure the assumption that the semantic dimension of the morphosemantic relations is transferable across languages.

Acknowledgments

We thank our student assistants Anna-Lena Feichter and Mariya Kavaldzhieva for a first round of manual annotation of the data.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Verginica Barbu Mititelu, Gianina Iordăchioaia, Svetlozara Leseva, and Ivelina Stoyanova. 2023. The meaning of zero nouns and zero verbs. In Sven Kotowski and Ingo Plag, editors, *The semantics of derivational morphology. Theory, methods, evidence*, pages 63–102. Walter de Gruyter, Berlin/Boston.

Laurie Bauer, Rochelle Lieber, and Ingo Plag. 2013. *The Oxford Reference Guide to English Morphology*. Oxford University Press, Oxford, UK.

Manfred Bierwisch. 2009. Nominalization – lexical and syntactic aspects. In Anastasia Giannakidou and Monika Rathert, editors, *Quantification, definiteness, and nominalization*, pages 281–320. Oxford University Press, Oxford, UK.

Barli Bram. 2011. *Major Total Conversion in English*. Ph.D. thesis, Victoria University of Wellington.

Fabricio Chalub, Livy Real, Alexandre Rademaker, and Valeria de Paiva. 2016. Semantic links for Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 885–891, Portorož, Slovenia. European Language Resources Association (ELRA).

Eve V. Clark and Herbert H. Clark. 1979. When nouns surface as verbs. *Language*, 55(4):767–811.

Mark Davies. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, pages 447–464.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India. The COLING 2012 Organizing Committee.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Christiane Fellbaum, A. Osherson, and P. E. Clark. 2009. Putting semantics into WordNet's 'morphosemantic' links. In Z. Vetulani and H. Uszkoreit, editors, *Human Language Technology. Challenges of the Information Society. LTC 2007*, pages 350–358. Springer, Berlin, Heidelberg.

Cristina Fernández Alcaina. 2021. *Competition in the derivational paradigm of English verbs*. Ph.D. thesis, University of Granada.

Jane Grimshaw. 1990. *Argument Structure*. MIT Press, Cambridge, MA.

Gianina Iordăchioaia and Chiara Melloni. 2023. The zero suffix in english and italian deverbal nouns. *Zeitschrift für Sprachwissenschaft*, 42:1:109–132.

Lea Kawaletz. 2023. *The semantics of English -ment nominalizations*. Language Science Press, Berlin.

Paul Kiparsky. 1982. From cyclic phonology to lexical phonology. In Harry van der Hulst and Norval Smith, editors, *The structure of phonological representations*, pages 131–175. Foris, Dordrecht.

Max Kisselew, Laura Rimell, Alexis Palmer, and Sebastian Padó. 2016. Predicting the direction of derivation in English conversion. In *Proceedings of the ACL SIG-MORPHON workshop*, pages 93–98, Berlin.

Gabriella Lapesa, Lea Kawaletz, Ingo Plag, Marios Andreou, Max Kisselew, and Sebastian Padó. 2018. Disambiguation of newly derived nominalizations in context: A distributional semantics approach. *Word Structure*, 11:277–312.

Cristina Lara Clares. 2023. *Morphological competition in present-day English nominalisation*. Ph.D. thesis, University of Granada.

Rochelle Lieber. 2004. *Morphology and Lexical Semantics*. Cambridge University Press, Cambridge.

Rochelle Lieber. 2016. *English Nouns. The Ecology of Nominalization*. Cambridge University Press, Cambridge.

Hans Marchand. 1964. A set of criteria for the establishing of derivational relationship between words unmarked by derivational morphemes. *Indogermanische Forschungen [Indo-Germanic research]*, 69:10–19.

Chiara Melloni. 2011. *Event and result nominals*. Peter Lang, Bern.

George A. Miller, Richard Beckwith, Christiane. Fellbaum, Derek Gross, and Katherine Miller. 1990a. Introduction to Wordnet: an on-line lexical database. *International journal of lexicography*, 3(4):235–244.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990b. Introduction to WordNet: An online lexical database. *International Journal of Lexicography*, 3(4):235–244.

Ingo Plag. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. De Gruyter, Berlin/New York.

Ingo Plag. 2003. *Word-formation in English*. Cambridge University Press, Cambridge.

Frans Plank. 2010. Variable direction in zero-derivation and the unity of polysemous lexical items. *Word Structure*, 3.1:82–97.

Franz Rainer. 2004. Polysemy in derivation. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, pages 338–353. Oxford University Press, Oxford.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Justine Salvadori and Richard Huyghe. 2022. Affix polyfunctionality in french deverbal nominalizations. *Morphology*, 33:1–39.

Salvador Valera. 2023. The semantics of noun-to-verb zero-derivation in english and spanish. *Zeitschrift für Sprachwissenschaft*, 42:1:153–180.

Rossella Varvara. 2017. *Verbs as nouns: empirical investigations on event-denoting nominalizations*. Ph.D. thesis, University of Trento.

Rossella Varvara, Gabriella Lapesa, and Sebastian Padó. 2021. Grounding semantic transparency in context a distributional semantic study on german event nominalizations. *Morphology*, 31:409–446.

Rossella Varvara, Justine Salvadori, and Richard Huyghe. 2022. Annotating complex words to investigate the semantics of derivational processes. In Harry Bunt, editor, *Proceedings of the 18th Joint ACL - ISO workshop on Interoperable Semantic Annotation (ISA-18)*, pages 133–141. European Language Resources Association (ELRA), Paris.

Rossella Varvara, Justine Salvadori, and Richard Huyghe. 2024. Creating and exploiting a lexical database of deverbal nouns in French. Talk given in the workshop *Data-based research in word formation*, within *The Biennial of Czech Linguistics*, Charles University, Prague.