Towards the Detection and Formal Representation of Semantic Shifts in Inflectional Morphology

Dagmar Gromann[1]
University of Vienna, Vienna, Austria
https://transvienna.univie.ac.at/en/
dagmar.gromann@gmail.com

Thierry Declerck
DFKI GmbH, Saarbrücken, Germany
ACDH-OEAW, Vienna, Austria
https://www.dfki.de/~declerck/
declerck@dfki.de

Abstract

Semantic shifts caused by derivational morphemes is a common subject of investigation in language modeling, while inflectional morphemes are frequently portrayed as semantically more stable. This study is motivated by the previously established observation that inflectional morphemes can be just as variable as derivational ones. For instance, the English plural "-s" can turn the fabric *silk* into the garments of a jockey, *silks*. While humans know that silk in this sense has no plural, it takes more for machines to arrive at this conclusion. Frequently utilized computational language resources, such as WordNet, or models for representing computational lexicons, like OntoLex-Lemon, have no descriptive mechanism to represent such inflectional semantic shifts. To investigate this phenomenon, we extract word pairs of different grammatical number from WordNet that feature additional senses in the plural and evaluate their distribution in vector space, i.e., pre-trained word2vec and fastText embeddings. We then propose an extension of OntoLex-Lemon to accommodate this phenomenon that we call inflectional morpho-semantic variation to provide a formal representation accessible to algorithms, neural networks, and agents. While the exact scope of the problem is yet to be determined, this first dataset shows that it is not negligible.

2012 ACM Subject Classification Information systems

Keywords and phrases Inflectional morphology, semantic shift, embeddings, formal lexical modeling

Digital Object Identifier 10.4230/OASIcs.LDK.2019.21

Funding Contributions by Thierry Declerck have been supported in part by the H2020 project "ELEXIS" with Grant Agreement number 731015 and by the H2020 project "Prêt-à-LLOD" with Grant Agreement number 825182.

Acknowledgements We would like to thank the anonymous reviewers for their very helpful comments on the original submission of this paper.

1 Introduction

Inflectional morphemes, such as plural *-s* for English nouns, are considered to cause changes in grammatical category without affecting a word's semantics [6]. Semantic shifts are commonly investigated for derivational morphemes, such as *-ment*, that form new lexical items [10], but less so for inflectional morphemes. This study is motivated by the observation that irregularities in inflectional morphemes affect semantic change, a phenomenon that is quite common as we try to show by generating a dataset for the English plural. A monomorphemic example for this phenomenon is the shift from *people* as a common plural of *person* to *peoples*, which refers to a body of persons united by race, ethnicity, and community rather than the

[1] corresponding author

2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 21; pp. 21:1–21:15

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

plural of people. More complex examples include multimorphemic words, e.g. *pyrotechnics*, as much as phrases, such as *blue devil* the weed as opposed to *blue devils*, which refers to depression. While our analysis focuses on English plurality, other examples such as gender, e.g. *la cabeza* for the physical head in Spanish as opposed to *el cabeza* for a male in charge in Spanish, or tense, e.g. *live* as opposed to the augmented senses of making a *living*, support the argument that this phenomenon generalizes across languages and inflectional morphemes.

Assuming regularities in inflectional morphology may lead to a restrictive view in creating language resources and models. In terms of models, Avrahaman and Goldberg [1], for instance, notice a drop in semantic performance when focusing exclusively on base forms without the words' inflections in training morphological embeddings and attribute this to potential inflectional irregularities (i.e. gender) without further investigating the phenomenon. Their overall recommendation, nevertheless, is to train morphological embeddings on lemmas rather than surface forms. When it comes to popular language resources, WordNet [8] implicitly acknowledges this phenomenon by attributing separate definitions to plural forms of nouns where the meaning changes with the grammatical number. We say "implicit" because when searching for the singular form, e.g. *silk* in the sense of the fabric, the plural and its separate meaning(s) are not available. One has to explicitly search for and be aware of a separate entry for the plural form *silks* to find out that it may refer to garments of a jockey.

To systematically investigate this phenomenon of regular inflectional morphemes that cause semantic shifts, which we call inflectional morpho-semantic variation, we limit our analysis for this paper to grammatical number in English nouns. The proposed method consists in detecting, analyzing, and representing such variants. First, we detect morpho-semantic variants in WordNet based on an augmented number of synsets when querying the plural form of a word. Senses specific to the plural are considered to indicate semantic shift. Second, the context of singular-plural pairs with different senses is evaluated by exploring the distribution of their representation in vector space. To this end, we utilized two pretrained embedding repositories: word2vec [15] and fastText [4]. Results thereof confirm a general intuition of two main types of variants: (i) semantic shifts where plural meanings entirely drift from the singular and lose all connections to its senses, and (ii) those with a clear connection to the singular but also additional meanings. To facilitate an improved representation in computational resources, we propose an extension of the OntoLex-Lemon computational resource[2] that represents linguistic and lexical knowledge in relation to a formal representation in order to accommodate inflectional morpho-semantic variation.

We see our main contributions to the broader topic of language, data and its representation as knowledge as providing:

- a dataset of inflectional morpho-semantic variants of grammatical number,
- a theoretical analysis of different types of such variants,
- an analysis of their representation in conventional vector space, and
- a formal representation method to differentiate inflectional morphemes with and without semantic shift.

To detail these contributions, we first discuss differences in inflection and derivation as well as types of inflectional morpho-semantic variants detected. Section 3 then describes our analysis of such variants in WordNet and our method for extracting a dataset from it as well as the results of that extraction process. Section 4 details the analysis of extracted variants in vector space, followed by a proposal to formally represent them. Prior to some concluding remarks, we discuss approaches related to the analysis of morphology in vector space.

[2] OntoLex-Lemon is the result of a W3C Community Group on representing rich linguistic grounding for ontologies. The final specifications of this model are available at `https://www.w3.org/2016/05/ontolex/`. See also [13] or [14].

2 Inflectional Morpho-Semantic Variants

Morphology investigates the structure of words by analyzing the smallest meaning bearing unit in language, called *morphemes* and their contribution to establishing relations between different words. Morphology most commonly differentiates inflection and derivation. In this paper, we are particularly interested in variants and irregularities of the former type.

2.1 Inflection and Derivation

Inflectional morphology is a set of processes that outwardly change the syntactic information of the word typically without changing its semantics, such as verb tenses. In contrast, in derivational morphology the word form change causes a semantic shift in meaning, such as the affix *un-* in English. Thus, affix patterns attributed to the former category are considered semantically regular with the base form of the respective morphological variants, whereas patterns of the latter category are considered semantically irregular leading to changes in meaning. In other words, while the boundary between these two tends to be a continuum rather than a divide, a generally accepted understanding is that derivation in contrast to inflection changes meaning [6].

2.2 Inflectional Variants

Inflectional morphemes have been studied extensively, however, questions regarding their universality across languages and the exact nature of their semantics remain open [10]. Our interest in this paper focuses on semantics of nominal expressions of grammatical number. In general, it can be stated that singulars denote atomic entities, dual numbers denote pairs, and plurals refer to groups of two or more entities. However, plurals with an associative meaning – denoting one person explicitly and a contextually relevant entity or group such as *los reyes* in Spanish that can denote king and queen – already require a different semantic [10]. In associative meanings and other exceptional cases of grammatical number discussed in Kiparsky and Tonhauser's extensive analysis of inflectional semantics [10], a tight semantic coupling between singular and plural is maintained. In this paper, we are interested in cases were this relation is entirely broken apart and entirely different semantics are assigned to the plural. This has to be differentiated from phenomena such as suppletion [5], where inflection causes drastic changes to the surface form, such as *person* being changed to *people*.

Analyzing grammatical number of English nouns, we noted that in some exceptional cases addition of a plural suffix entirely changes the semantics of the word. For instance, *bloomer* refers to a flower that blooms in a certain way or a loaf of white bread, whereas *bloomers* informally and historically refers to a woman's underpants. Singular and plural share no present-day semantics, even though they might be etymologically related[3]. In such cases the plural exclusively refers to this one meaning without any relation to senses of the singular counterpart of the same word. In other words, semantics of the inflected form have no relation to the semantics of the non-inflected, lemmatized form. As a second major group, plural examples with an idiosyncratic meaning might still simultaneously represent two or more specimen of a specific thing or living being. For instance, *names* refers to name calling in the sense of verbal abuse as much as to two or more designations of things or beings. At

[3] Several sources attribute the plural to the person Amelia Bloomer, a women's right advocate who came to be associated with the clothing reform. However, this does not entirely exclude the possibility of a relation to the singular form of the word.

the same time, the singular *name* might refer to a person's reputation, in which case it might only be used in singular. As an example of living beings in this category, *clams* denote bits of sweet chocolate used as ice cream topping as much as several marine mollusks. In some cases, senses of singular and plural forms might not be identical but share some common characteristics. For instance, *antipode* refers to direct opposites whereas *antipodes* refers to places or regions on diametrically opposite sides of the Earth. From this example we can see that even though the latter meaning is considerably more specific than the former, there are some common traits. We call all of these cases of regular inflectional morphemes that cause semantic shifts morpho-semantic variants and in this paper focus on English nominal constructs of grammatical number, but are confident that other cases, such as gender, would be worth investigating in future.

With the proposed dataset of examples, analysis of word embeddings, and theoretical discussion of the problem, this paper contributes to the investigation of the semantics of grammatical numbers. While the exact scope of the problem of inflectional morpho-semantic variants has not yet been fully identified, an initial dataset extracted from a general language lexicon with an exclusive focus on grammatical number already yielded a significant number of examples (see below). Number and quality of obtained examples suggest that inflectional semantics can be as variable as derivational semantics and are presented in the following.

3 Inflectional Suffixes of Number in WordNet

WordNet [8] has been a very influential lexical-semantic resource over the last decades which is being used in a variety of language technology tasks (e.g. [2]). Its popularity can partially be attributed to its considerable coverage and the quality of mainly manually curated entries. In this section, we investigate the representation of nominal inflectional suffixes of grammatical number in English. Building on this analysis, we propose an approach for the automated extraction of irregular cases for the evaluation of their representation in vector space that is proposed in the next section. To the best of our knowledge, existing datasets of semantic relations in inflectional morphology focus on regular cases. Thus, we think that this dataset can also provide a more encompassing and powerful test bed for models of morphological semantics and its variants. For now, it only looks at English nouns, however, considerable extensions in grammatical category and languages covered are ongoing.

3.1 Representation in WordNet

We observe that WordNet's representation of semantics of words does not inherently support the representation of inflected forms. Semantically similar words are grouped into sets of synonymous words, so-called synsets. When searching for a word in a WordNet interface[4], all potential synsets for this word are returned, including its gloss (a short definition) and all synonyms of the word pertaining to the same synset. While it is possible to actively search for plural forms of a noun, in a vast majority of cases the interface returns results for its uninflected counterpart because it lemmatizes the input. In cases of complementary plural entries, WordNet displays augmented lists of synsets: those associated with the singular, e.g. *people*, and those associated with the plural, e.g. *peoples*. All senses for this example are displayed in Listing 1.

[4] See `http://wordnetweb.princeton.edu/perl/webwn` for a Web interace and `http://www.nltk.org/howto/wordnet.html` for a Python interface integrated in the Natural Language Toolkit (NLTK) [3]

Listing 1 The Synsets for "people" vs. "peoples".

```
people.n.01      ((plural) any group of human beings ... collectively)
citizenry.n.01   (the body of citizens of a state or country)
people.n.03      (members of a family line)
multitude.n.03   (the common people generally)
peoples.n.01     (the human beings of a particular nation or community
                 or ethnic group)
```

This differentiation of grammatical number in the representation of synsets and associated meanings intuitively suggests that plural and singular forms do not share all meanings. Regular cases, such as *car* returns no additional synsets and senses for its inflected form *cars*. Thus, it can be assumed that the change of grammatical number does not cause any semantic shift in those cases. This means, in turn, that it can be assumed that the availability of additional senses indicates such semantic shifts and therefore irregular inflectional forms. When we follow this line of thought for the above example and consult an additional resources[5], we find a clear distinction in meaning between *people*, which itself has to be treated as plural in most senses, and *peoples* with an identical meaning as the corresponding synset in Listing 1. We also find indications that *people* by suppletion is the predominantly used plural of *person* (rather than *persons*). To systematically analyze these irregularities we generate a dataset from WordNet described in Section 3.2.

3.2 Dataset Creation Method

Building on this analysis of joint representations of grammatical number and senses in WordNet, we extract a full list of all available English lemmas in WordNet. Each entry in this list is automatically inflected to its potential plural form, which, if available in the noun list of lemmas, is used to query for its senses. If a query for an inflected form returns senses different from the ones obtained by querying the singular, we consider it an indication of semantic shift.

For a focused analysis of one specific phenomenon of inflectional irregularities, we limit the number of investigated cases to nouns and grammatical number. We analyze English inflectional suffixes for plural nouns, which in our dataset turn out to be: addition of *-s*, addition of *-es*, replacement of *-y* by -ies, replacement of *-us* by *-i*. All potential English nouns are inflected using the *inflect*[6] package in Python and then used to query for WordNet senses in its NLTK corpus.

We implement some restrictions to enhance the quality of the obtained variants. First, we limit the part of speech tag to nouns in order to ensure that all returned senses relate to singular and plural nouns only. Second, only words with at least one overlapping sense in singular and plural are considered. This is due to the fact that WordNet automatically lemmatizes the query word and returns singular and plural senses, the latter only where available, the former even if unrelated to the plural. For instance, *silk* and *silks* share no meanings but the query for the latter still returns all senses of the former. In cases where no singular senses are included due to the lemmatized plural, the words are of a different lemma, such as the personal pronouns *us* and *I* or *faro* the card game and *Faroes* referring to the

[5] We find this same distinction in Merriam Webster's online dictionary `https://www.merriam-webster.com/dictionary/people`

[6] `https://pypi.org/project/inflect/`

island. This is also the reason why we have to subtract all singular senses from the plural to ensure we are left with senses unique to the plural version of the noun. Finally, we remove all senses with lemma names that start with a capital letter to avoid including proper names as plurals, such as *sills* referring to the US operatic soprano Beverly Sills as a plural of *sill*.

In terms of evaluation, all authors of this paper manually checked each resulting singular-plural pair and their senses. Even though we used a morphological analysis tool, several basic grammar rules were violated, such as adding an *-s* to words ending in *-s*, which, for instance, turns the *bos*, a cattle, into the *boss*, the leader. In the formation of *-s* plurals, abbreviations (*aids* related to *aid*), Greek letters (*mu* related to *mus*), chemical elements (*co* related to *coes*), and currencies (*lats* related to *lat*) marked the majority of unreasonable pairings. All of these cases were removed from the final dataset, which lead to 23 removals for the cases of *-s* plural endings and 9 removals for *-es* additions and none for the other two types of endings. We publish a file with all removed entries alongside the actual dataset. The following section represents the resulting dataset that does not take these removed elements into consideration.

3.3 Dataset Results

Applying the described method leads to a dataset of inflectional suffixes for grammatical number that cause semantic shifts. The dataset is published[7] in two versions: (i) one with a header line indicating the type of suffix and a word pair per line of format "singular plural", and a (ii) second version with the same as in (i) and additionally all definitions for the singular and plural from WordNet alongside the synset identifier. Version (i) allows for faster parsing of the variants while version (ii) allows for a detailed tracking and (manual) evaluation of the results. We additionally add evaluations of cosine similarities in vector space to the data repository of this paper.

Table 1 Statistics on Final Dataset.

Suffix type	Number of Examples	Example
-s	419	*silk silks*
-es	11	*rich riches*
-y to *-ies*	24	*fifty fifties*
-us to *-i*	1	*fungus fungi*
All	455	–

Quantified results of the dataset are presented in Table 1 as well as examples for each type. The majority of examples could be detected for the most common additive suffix, while the other suffixes where less common. The table presents one example for each category of morpho-semantic variant. As defined before *silk* denotes the fabric and *silks* a jockey's garments. Second, *rich* conventionally refers to people in possession of wealth, whereas *riches* is commonly used to denote wealth as such. Third, the number *fifty* has to be differentiated from the historical decade *fifties* as well as the time in life between the age of 50 and 60. Finally, *fungus* may refer to an organism, while *fungi* refers to the taxonomic kingdom. We were interested to see in how far word embeddings purely trained on contexts are able to capture these semantic irregularities in inflectional morphemes.

[7] `https://github.com/dgromann/imsev`

4 Inflectional Suffixes of Number in Vector Space

Distributed semantic models capture word meaning purely based on its contexts. Real-valued vector representations of words are obtained by analyzing words occurring in the same sentences, where the window size determines the number of words to the left and to the right that are considered during training with a neural network. Resulting word embeddings have turned out to be highly powerful representations of different semantic aspects of individual words. In our case they are utilized to test whether a purely context-based approach is capable of capturing morpho-semantic variation in grammatical number.

To this end, we utilized two different pre-trained embedding repositories for English: word2vec [15] trained on the Google news corpus and fastText [4] trained on the English web-crawl and Wikipedia corpus[8]. In training, word2vec represents a feedforward neural network with a softmax output layer that trains embeddings with negative sampling, predicting the context for a given center word in its widely used skipgram version. This training model is adapted by fastText to encode character n-grams, where word vectors represent compositions of character n-gram vectors. This has the advantage of a reduced out of vocabulary rate due to the flexible composition of new words based on their n-grams. We decided against the utilization of morphological embeddings such as the ones proposed by Avraham and Goldberg [1] who adapt fastText to combine lemma, surface form, and morphological tag. Both lemma and morphological tag could bias the learned vector space towards ignoring irregularities and thus are counter-intuitive training methods for our purposes.

In order to test the location of a vector we need to navigate through vector space created by the embeddings and analyze the environment of a desired target vector. This can be achieved by querying nearest neighbors of the singular and plural of each word in our dataset (if represented in the vocabulary) and then estimate the overlap of neighbors in the top ten returned closest vectors. Apart from the fact that people seem to frequently misspell *people*, Listing 2 shows that *peoples* is not in the immediate neighborhood.

Listing 2 Top six nearest neighbors of "people" in word2vec.

```
peole: 0.6058608293533325
poeple: 0.59071284532547
individuals: 0.5827618837356567
folks: 0.5794459581375122
peple: 0.578874409198761
peo_ple: 0.5768002271652222
```

Listing 3 displays the same query for words having similar contexts as the word *people* in fastText. We can observe that we get a different list of words, but in both cases the word *peoples* is not included. This is an indication that both words do not share a meaning.

Listing 3 Top six nearest neighbors of "people" in fastText.

```
...people: 0.7241666316986084
'people: 0.6962485313415527
people's: 0.6582629680633545
,people: 0.6566357612609863
''people: 0.6466725468635559
@people: 0.6439790725708008
```

[8] https://fasttext.cc/docs/en/english-vectors.html

Table 2 Evaluation of top ten neighbors of singulars for relation to semantically shifted plurals.

	in top ten neighbors	shared meaning	not in neighbors	plural only	OOV
word2vec	258	237	145	16	53
fastText	303	288	114	19	39

In contrast to Listing 2, Listing 4 shows that *peoples* seems easier to spell and very clearly refers to a very different sense. The singular version with a significantly different semantics does not occur in the list of neighbors. We display here only the word2vec based listing.

Listing 4 Top six nearest neighbors of "peoples" in word2vec.

```
Indigenous_peoples, Similarity: 0.54
Diasporas, Similarity: 0.54
indigenous_peoples, Similarity: 0.53
human_being, Similarity: 0.53
pluralistic_societies, Similarity: 0.53
humankind, Similarity: 0.52
```

For the above case and the represented six senses, an overlap of zero neighbors would be the result, showing a strong indication for a semantic shift that is also captured by the distributed semantic model. We repeated the above experiment with all examples in our dataset and found that embeddings can be utilized in order to neatly separate inflectional morpho-semantic variants that share a meaning with the singular from those without a shared meaning. This can be achieved by evaluating whether the inflected plural form from our dataset is part of the list of ten nearest neighbors.

Querying a plural in WordNet always results in the listing of all singular senses of a word and, where available, senses specific to the plural. However, this rigorous listing of singular senses also applies to plural nouns that share no sense with their singular counterpart. For instance, querying the pants *khakis* would result in a listing of all senses related to *khaki* and that of the plural. Thus, we had to turn to a different resource to obtain the differentiation for plural forms that share senses with the singular and those that do not share any senses, i.e. exist only in the plural version. To this end, Wiktionary usefully differentiates between "plural of" a certain singular word and "plural only". All plural instances in the latter category are considered not to share a meaning with their singular counterparts. As a result, we obtain 412 singular-plural pairs that share meanings and 43 plural words that have no Wiktionary link to a potential singular form. This split helped us evaluate whether word embeddings captured this information since their creation is purely based on context.

As represented in in Table 2, in word2vec 258 plurals are found in the top ten nearest neighbors of their singular of which 237 share senses according to our Wiktionary statistics. For fastext, out of 303 plurals in the top ten neighbors, 288 also share senses according to Wiktionary. However, little overlap could be observed between plurals that are not in the vector neighborhood and have exclusively "plural only" meanings in Wiktionary. We believe that this discrepancy can be attributed to words that are predominantly used in their "plural only" meaning but also share a sense with their singular counterpart. Out of vocabulary (OOV) examples are lower with fastText due its character n-gram encoding method.

While this behaviour is also reflected in the similarity measure between singular and plural, there is no exact division line. Around the values of 0.40 to 0.48 cosine distance pairs in word2vec can belong to either category. Lower values clearly separate input pairs

Table 3 Number of examples per cosine similarity range times 100.

	0-20	20-40	40-60	60-80	80+
word2vec	13	66	129	168	14
fastText	3	27	98	231	54

by meaning, while higher values are good indicators of shared as well as separate meanings. For an overview, we provide ranges of cosine similarity in Table 3, which clearly shows a predominant accumulation of pairs in the range of 0.6 to 0.8 cosine similarity. This preliminary evaluation of the representation of inflectional morpho-semantic variants in vector space needs to be grounded in a more substantial and formal evaluation with several annotators and several evaluation metrics, which we intend to do as part of our future work. Furthermore, we intend to extend the created dataset from other sources with more substantial annotations of inflectional behaviors. Nevertheless, this initial method provides an estimation of the magnitude of each subtypes of inflectional morpho-semantic variants, one where the plural sense is entirely shifted to have nothing in common with the singular and one with partial shifts.

5 Representation in OntoLex-Lemon

The OntoLex-Lemon model was originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the description of ontology elements are equipped with an extensive linguistic description. This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or to specialized vocabularies. The main organizing unit for those linguistic descriptions is the lexical entry, which enables the representation of morphological patterns for each word and/or affix. The connection of a lexical entry to an ontological entity is marked mainly by the `denotes` property or is mediated by the `LexicalSense` or the `LexicalConcept` properties, as this is represented in Figure 1, which displays the core module of the model.

The OntoLex-Lemon model describes at its core an entry-sense relation. Form variants of an entry are encoded as instances of the class `Form` and none of this form variants can be linked directly to a lexical sense, which would be a direct way to represent morpho-semantic phenomena. Therefore, in OntoLex-Lemon morpho-semantic variants can only be represented via their linking to distinct lexical entries.

Our OntoLex-Lemon compliant approach consists in creating a new lexical entry for the plural form that has a specific meaning. We showcase this approach with the word pairs *letter-letters*. While several senses can be associated with both the singular and the plural form of the lexical entry *letter*, the literary culture sense can be associated with the plural form. On the other hand, the sense of literal interpretation (e.g. in the case of law texts that are interpreted by the *letter*) is generally assigned to the singular form. In the following listings, we show, in a simplified manner, the way this complex information can be encoded in OntoLex-Lemon.

Listing 5 displays the lexical entry for *letter*. It is stated that two forms are associated with this noun: a singular (the `canonicalForm`) and a plural (the `otherForm`) form. In this simplified entry, we link only to one sense: the one of an exchange between two parties (see listing 8).

Listing 5 The lexical entry for *letter*.

```
:letter
  rdf:type ontolex:Word ;
  lexinfo:partOfSpeech lexinfo:noun ;
  ontolex:canonicalForm :Form_letter ;
  ontolex:otherForm :Form_letters ;
  ontolex:sense :LexicalSense_letter_1 ;
.
```

Listings 6 and 7 display the basic encoding for the two possible word forms for the entry *letter*.

Listing 6 The form for *letter* in singular.

```
:Form_letter
  rdf:type ontolex:Form ;
  lexinfo:number lexinfo:singular ;
  ontolex:writtenRep "letter"@en ;
.
```

Listing 7 The form for *letters* in plural.

```
:Form_letters
  rdf:type ontolex:Form ;
  lexinfo:number lexinfo:plural ;
  ontolex:writtenRep "letters"@en ;
.
```

The next listing is about the shared sense associated with the lexical entry. As there is a Wikidata entry for the type of entity this sense can refer to, we make use of the `ontolex:reference` property in order to link to this data source.

Listing 8 The lexical sense for the entry *letter* (which can have singular and plural forms).

```
:LexicalSense_letter_1
  rdf:type ontolex:LexicalSense ;
  rdfs:comment "letter as a missive from one party to another (taken
      from Wikidata)" ;
  ontolex:isSenseOf :letter ;
  ontolex:reference <https://www.wikidata.org/wiki/Q133492> ;
.
```

Listing 9 is introducing the additional lexical entry for the plural form of *letter* that has a specific meaning that can not be associated to its singular form. Therefore we link this entry only to the plural instance of the class `Form` and to the specific sense encoded in listing 10, where we additionally formulate the constraint that the usage of this sense is restricted to the plural form *letters*.

Listing 9 The special lexical entry for *letters*.

```
:letters
  rdf:type ontolex:Word ;
  lexinfo:partOfSpeech lexinfo:noun ;
  rdfs:comment "encoding singular and plural entries" ;
  ontolex:canonicalForm :Form_letters ;
  ontolex:sense :LexicalSense_letters_1 ;
.
```

Listing 10 The sense for *letters* in plural.

```
:LexicalSense_letters_1
  rdf:type ontolex:LexicalSense ;
  rdfs:comment "\"letters\" as \"literary culture\"" ;
  ontolex:usage :Form_letters ;
.
```

In fact the use of the `ontolex:usage` property could suffice in order to mark that a sense is restricted to a particular inflectional form of an entry, as exemplified below in Listing 11 for the sense of the literal interpretation, without the need to introduce a new lexical entry.

Listing 11 The literal interpretation sense for *letter* in singular.

```
:LexicalSense_letters_1
  rdf:type ontolex:LexicalSense ;
  rdfs:comment "\"letters\" as \"literary culture\"" ;
  ontolex:usage :Form_letters ;
.
```

An alternative approach could be to allow a sense to be (only) expressed by an instance of the class `Form` that denotes a grammatical number of the associated headword. To this end, the current state of the OntoLex-Lemon model would need to be extended in order to allow a relation (or a property) between an instance of the class `ontolex:Forms` and instances of the class `ontolex:LexicalSense`. This would allow for the direct representation of morpho-semantic variants of the type discussed in this paper. But this second approach would signify a departure from the core module of OntoLex-Lemon, which stipulates that only a lexical entry can be linked to a sense, a concept or an ontological reference.

In both cases, we are able to model together both the information obtained from WordNet and insights derived from the word embeddings. This could lead to a mapping of word embeddings to a computational lexicon. This mapping could be utilized to validate WordNet entries and dynamically create new ones.

6 Discussion

In terms of the creation method for the dataset, we opted for the utilization of simple grammatical rules of inflectional morphology in form of an existing Python package. While this method could be improved on several levels – as for instance utilizing several resources as sources of information, applying more complex morphological components, or analyzing a larger variety of morphosemantic variants – it still returned a significant number of examples

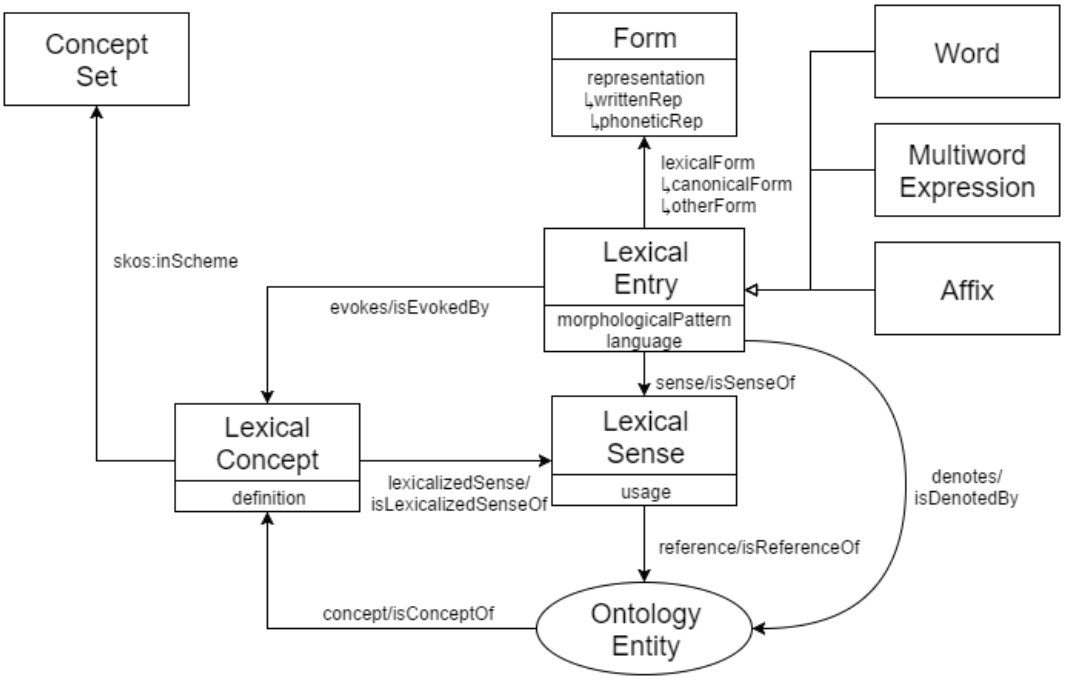

Figure 1 The core module of OntoLex-Lemon: Ontology Lexicon Interface. Graphic taken from `https://www.w3.org/2016/05/ontolex/`.

for the phenomenon under investigation. One central issue of the dataset is the duplication of entries due to more than one plural version of a word, which for now occurred only once with "dominos" and "dominoes" as valid plural versions, but could be aggravated with a larger and more complex dataset. This is one more argument in favor of a more complex morphological analysis tool in the dataset creation process.

For the dataset creation method, WordNet is a very useful resource to identify regular inflectional plurals with additional senses. Applying similar techniques with extended rule sets to other lexical and terminological resources promises to result in a larger and more heterogeneous dataset. For now Wiktionary was utilized to check the separation of plurals that share a meaning with singular and those that have no sense in common with the singular and compare this separation to plurals in the ten nearest neighbors in vector space of two word embeddings. One issue that has to be mentioned here is that modeling of plural-singular connections turns out to be inconsistent in Wiktionary. We consulted plural pages only and categorized plural words into "plural only" senses if no reference "plural of" could be found on the page. However, at times, the reference is missing from the plural page, e.g. "graphics (uncountable)", but a reference to the plural can still be found on the singular page, e.g. "graphic (plural graphics)". Such inconsistent modeling complicate any automated information extraction process.

Thus, in the long run, this separation of plural only and shared meanings of our English noun pairs should be improved upon. One option is the costly manual annotation that might suffer from the complexity of the task, since users may not be familiar with all senses of a word. On the other hand, a corpus-based statistic on the frequency of different plural meanings might provide a more principled analysis of which words are predominantly used in their plural only meaning rather than as a plural of the singular with a shared meaning. In this regard, it would also be interesting to see how to automatically integrate the semantic shift as well as corpus- and vector-based information into OntoLex-Lemon.

In terms of exploring vector space, it would be interesting to repeat our experiments with embedding repositories other than word2vec and fastText. However, to some extend the choice here is limited, since more powerful recent embedding libraries, such as the Bidirectional Encoder Representations from Transformers (BERT) [7], are directed towards words in context and/or sentential embeddings. Querying such contextualized word embeddings with individual words devoid of any context somewhat defeats their purpose.

7 Related Work

While we have presented some studies on inflectional morphology in Section 2.1, this section focuses on computational morphological approaches using embeddings. Analyses of the relation between semantics and morphology have particularly lately been done based on word embeddings. Extensive analogy-based evaluations of morphological and semantic relations in word embedding models across more than 40 categories have shown that inflectional relations are among the best performing ones [9]. Nevertheless, none of them reached an 80% accuracy mark. On the one hand, this could be attributed to the nature of the analogy task and it has been attempted to better adapt the nature of the task to morphological variations [11]. On the other hand, embeddings can be improved by making them morphologically aware, that is, learn embeddings for morphological components (e.g. lemmas, affixes), morphological categories, and word surface forms.

Recent approaches have focused on composing morphologically aware embeddings to improve on the semantic performance of embedding models. Avrahaman and Goldberg [1] adapt fastText [4] to train embeddings for all possible combinations of surface form, lemma, and morphological tags of a word and test on common and rare words. They attribute semantic information to the lemma and morphological information to the affix. In conclusion, they explicitly recommend using lemmas only as for most tasks morphological affixes an be dropped. Nevertheless, their analyses of common words reveals a drop if excluding surface forms (limiting vectors to lemmas and morphological tags), which they attribute to semantic shifts in morphological templates without further investigating the phenomenon, which is exactly where we take over in this paper.

In general, it has been shown that complex composition models tend to outperform simple vector addition or composition methods. Malouf [12] propose a Recurrent Neural Network (RNN) model that predicts complex inflectional classes, which takes a lexeme, set of morpho-syntactic features, and a partial word form as input and outputs a probability distribution for the next segment in the word form in seven morphologically complex languages. Cotterell and Schütze [6] find that approximating a vector with a trained Recurrent Neural Network (RNN)-based model outperforms additive vector composition. However, this approach focuses on derivational morphology, which by definition investigates semantic shifts induced by morphological changes to a word.

In derivational morphology, several linguistic factors have been analyzed in connection with word embeddings. Pado et al. [16] analyze linguistic factors in the ability of Compositional Distributional Semantic Models (CDSMs) to predict distributional vectors for derived word forms given the vectors for their base forms, which they test on 74 derivation patterns in German. Most difficult derivational patterns to predict were found to be those modifying argument structure, semantic irregularities, and within-POS derivation. We believe that more studies in this directions might be in order for the semantic behaviour of inflectional morphology, since those causing semantic shift currently have not been considered in modern approaches.

8 Conclusion and Future Work

We present ongoing work on detecting and formally representing inflectional morpho-semantic variants. While their impact on morphological embeddings has been noted, to the best of our knowledge no comprehensive study has been provided. Our contributions are a dataset of English nominal inflectional morpho-semantic variants of grammatical number and an analysis of their representation in vector space models. One major outcome of this work is the realization that the problem of semantic shifts in inflectional variants of regular morphemes is a significant phenomenon and that it seems that inflectional semantics can be as variable as derivational semantics.

As a second major contribution of this work, we propose a method for representing such variants in a machine readable and formal model called OntoLex-Lemon. To this end, the current version of the model needs to be slightly adapted to account for morpho-semantic variants of grammatical number. This extension can serve as a basis for its potential use for latest neural network based approaches on morphological modeling, such as the potential to include morphological information in training knowledge graph embeddings. Testing these potentials is future work.

For the time being, our approach focuses on English nouns and grammatical number. However, we have good reason to believe that discussed phenomena can be observed in different inflectional morpheme types as well as natural languages, both of which are left as future directions. We also intend to extend our study o different pre-trained embeddings.

References

1 Oded Avraham and Yoav Goldberg. The Interplay of Semantics and Morphology in Word Embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 422–426. Association for Computational Linguistics, 2017.

2 Jiang Bian, Bin Gao, and Tie-Yan Liu. Knowledge-Powered Deep Learning for Word Embedding. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 132–148. Springer, 2014.

3 Steven Bird. NLTK: The Natural Language Toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Association for Computational Linguistics, 2002.

4 Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

5 Greville G Corbett. Canonical Typolgy, Suppletion, and Possible Words. *Language*, pages 8–42, 2007.

6 Ryan Cotterell and Hinrich Schütze. Joint Semantic Synthesis and Morphological Analysis of the Derived Word. *Transactions of the Association of Computational Linguistics*, 6:33–48, 2018.

7 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805, 2018. `arXiv:1810.04805`.

8 Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May 1998.

9 Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. Analogy-based Detection of Morphological and Semantic Relations with Word Embeddings: What Works and What doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, 2016.

10 Paul Kiparsky and Judith Tonhauser. Semantics of Inflection. *Handbook of Semantics*, 3:2070–2097, 2012.

11 Tal Linzen. Issues in Evaluating Semantic Spaces Using Word Analogies. In *In Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, 2016.

12 Robert Malouf. Abstractive Morphological Learning with a Recurrent Neural Network. *Morphology*, 27(4):431–458, 2017.

13 John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asuncion Gomez-Perez, Jorge Garcia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. Interchanging Lexical Resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719, 2012.

14 John P. McCrae, Paul Buitelaar, and Philipp Cimiano. The OntoLex-Lemon Model: Development and Applications. In Iztok Kosem, Jelena Kallas, Carole Tiberius, Simon Krek, Miloš Jakubíček, and Vít Baisa, editors, *Proceedings of eLex 2017*, pages 587–597. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., September 2017.

15 Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781, 2013. `arXiv:1301.3781`.

16 Sebastian Padó, Aurélie Herbelot, Max Kisselew, and Jan Šnajder. Predictability of Distributional Semantics in Derivational Word Formation. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 1285–1296, 2016.