Disambiguation of newly derived nominalizations in context: A Distributional Semantics approach.

Gabriella Lapesa, Lea Kawaletz, Ingo Plag, Marios Andreou, Max Kisselew & Sebastian Pado

September 7, 2017

Abstract

One of the central problems in the semantics of derived words is polysemy (see, for example, the recent contributions by Lieber 2016 and Plag et al. 2017). In this paper, we tackle the problem of disambiguating newly derived words in context by applying Distributional Semantics (Firth 1957) to deverbal *-ment* nominalizations (e.g., *bedragglement*, *emplacement*).

We collected a dataset containing contexts of low frequency deverbal *-ment* nominalizations (55 types, 406 tokens) extracted from large corpora such as the Corpus of Contemporary American English. We chose low frequency derivatives because high frequency formations are often lexicalized and thus tend to not exhibit the kind of polysemous readings we are interested in. Furthermore, disambiguating low-frequency words presents an especially difficult task because there is little to no prior knowledge about these words from which their semantic properties can be extrapolated.

The data was manually annotated according to eventive vs. non-eventive interpretations, allowing also an 'ambiguous' label in those cases where the context did not disambiguate. Our question then was to what extent, and under which conditions, context-derived representations such as those of Distributional Semantics can be successfully employed in the disambiguation of low-frequency derivatives.

Our results show that, first, our models are able to distinguish between eventive and non-eventive readings with some success. Second, very small context windows are sufficient to find the intended interpretation in the majority of cases. Third, ambiguous instances tend to be classified as events. Fourth, the performance of the classifier differed for different subcategories of nouns, with non-eventive derivatives being harder to classify correctly. We present indirect evidence that this is due to the semantic similarity of abstract non-eventive nouns to eventive nouns. Overall, this paper demonstrates that distributional semantic models can be fruitfully employed for the disambiguation of low frequency words in spite of the scarcity of available contextual information.

[Acknowledgements to be added]

Contents

1 Introduction **3**

2 Ambiguity, disambiguation and the construal of meaning **5**

3 Disambiguating *-ment* derivatives in context **7**

4 Data set and coding **9**
4.1 The suffix *-ment* 9
4.2 Sampling newly derived words in *-ment* 10
4.3 Sampling attestations 11
4.4 Coding 12

5 Implementation **13**
5.1 Distributional Semantics 13
5.2 Semantic class disambiguation as supervised classification 17

6 Results **20**
6.1 Main experiment 20
6.2 Analysis 1: Per-class performance 22
6.3 Analysis 2: Lemma vectors vs. instance Vectors 26

7 Discussion and conclusion **28**

References **28**

A Implementation details **33**
A.1 Step 0: Collecting training nouns and training sentences 33
A.2 Step 1: Obtaining distributional representations for training nouns and training sentences 34
A.3 Step 2: Learning to distinguish eventive and non-eventive nouns 35
A.4 Step 3: disambiguating -ment derivatives 35

1 Introduction

In many languages polysemy in word-formation is all-pervasive (e.g. Rainer 2014). Bauer et al. (2013), Kawaletz and Plag (2015, 291), Lieber (2016, 18f), for example, come up with lists of readings available to English deverbal nominalizations involving the suffixes *-ing, -ation, -ment, -ance/-ence, -th* and conversion, similar as the one given in table 1. Other languages show similar patterns of polysemy, for example French (Uth 2011; Fradin 2011, 2012a,b), or German (Roßdeutscher and Kamp 2010; Roßdeutscher 2010).

Table 1: Readings of English nominalizations

Semantic category	**paraphrase**	**examples**
Event	'the event of V-ing'	*production, training*
Result	'the outcome of V-ing'	*acceptance, alteration*
Product	'the thing that is created by V-ing'	*pavement, growth*
Instrument	'the thing that V-s'	*seasoning, advertisement*
Location	'the place of V-ing'	*dump, residence*
Agent	'people or person who V-s'	*administration, cook*
Measure	'how much is V-ed'	*pinch, deceleration*
Path	'the direction of V-ing'	*decline, direction*
Patient	'the thing affected or moved by V-ing'	*catch, acquisition*
State	'the state of V-ing or being V-ed'	*alienation, disappointment*
Instance	'an instance of V-ing'	*belch, cuddle*

Given the variety of interpretations that derivatives of a given affix can give rise to, how do listeners find the right reading? This is not a trivial task, especially if we consider derivatives that are newly coined, or so infrequent that it is highly unlikely that listeners have stored available readings for these forms. And even if they do, a given form still can have different stored readings. Thus it seems that the context in which a given form occurs must provide the necessary cues that help listeners to come up with the reading that is actually intended by their interlocutor who uses the word.

In her recent book on nominalizations Lieber (2016, 8) formalizes the work of the context in deriving a particular reading as 'Contextual Coercion', "by which specific readings of those words are realized in specific syntactic contexts". The notion of Contextual Coercion is very similar to Pustejovsky's 'accomodation subtyping' (Pustejovsky 2011, 1411). In this view an inherently underspecified or unspecified semantic feature (for example [material]) is fixed to one value (for example [+material]) on the basis of other information available in the syntactic context. Let us use an example from Lieber (2016, 122ff) for illustration. The example is taken from the Corpus of Contemporary American English (COCA, Davies 2008), emphasis ours.

(1) The Guggenheim, considered the generator of 80 percent of Bilbao's 1 million annual tourists (Calvo 2001b), provided President Fraga with an example of a successful **construction** that helped to remake a city's image. (COCA ACAD 2012)

In this case, encyclopedic knowledge and the definite article tells the reader that *The Guggenheim* refers to a museum, which, via Contextual Coercion, leads to the inference that *construction* must be concrete and inanimate, which fixes the value of the feature [material] to [+material]. It should be noted, however, that an alternative interpretation (not mentioned by Lieber) still seems possible. Thus a city's image can not only be remade by having a nice new building, but also by successfully building one. For example, the city of Berlin's image has seriously suffered from the problems surrounding the construction (event reading!) of the new airport BER, and a successful construction (event reading!) of some other building may boost Berlin's image in this domain.

The example nicely shows that sometimes larger contexts and very specific shared knowledge may need to be fed into the construal or disambiguation of the meaning of polysemous derivatives. If such knowledge is not available or not clearly alluded to, listeners and readers may be stuck with ambiguous or underspecified meanings. The literature (e.g. Bauer et al. 2013; Lieber 2016; Andreou 2017) often discusses individual cases of disambiguation, but it is currently unclear to what extent, across many derivatives and contexts, the syntactic context of derived words actually helps to disambiguate a given derivative.

There is also a more general theoretical problem involved in approaching the interpretation of complex words that we have glossed over so far, i.e. what we mean more specifically by 'disambiguation'. Do we mean the selection of a particular meaning among some lexically listed meanings? Or the construal of one particular reading on the basis of different sources of information? Especially with regard to morphologically derived words, this is a vexed issue, since the interplay between affix and base may sometimes lead to lexically fixed possibilities of meaning (especially with established derivatives), and may sometimes necessarily involve the construal of meanings based on the semantic representations of base and affix (especially with novel forms). This problem will be discussed in more detail in section 2. For the purposes of our study a broad notion of disambiguation as 'finding the right reading' is used which allows for both possibilities mentioned.

In this paper, we empirically explore, for words derived with the suffix *-ment* in English, to what extent syntactic contexts provide the necessary cues for disambiguation. We will do so for newly derived and rare derivatives, for which readers and listeners do not have fixed meanings available. These cases are especially interesting because they necessarily involve the construal of meaning instead of retrieving attested interpretations for whole words from the mental lexicon. At the same time, this makes disambiguation an even harder task.

For our exploration, we adopt a corpus-based methodology developed to deal with meaning in context, Distributional Semantics (Firth 1957; Miller and Charles 1991; Schütze 1998). In this approach, the meaning of a word is represented as a vector which mathematically expresses the co-occurrences of this word with very many other words in a given corpus (see Section 5.1 for details and explanation). Distributional Semantics has been used successfully to model various aspects of lexical semantics. However, in the domain

of morphological derivation there is only little work using distributional semantic tools. For example, Cotterell and Schütze (2017) is a general attempt to use vector semantics for morphological segmentation, Kisselew et al. (2016) employ Distributional Semantics to determine the directionality of conversion, some scholars (e.g. Marelli 2015) investigated the transparency of morphological categories, or the semantic differences between certain suffixes (Varvara, 2017). To our knowledge, the present paper is the first distributional semantic approach to tackle the problem of disambiguation of newly derived words.

Focusing on only two readings, EVENTIVE and NON-EVENTIVE, we used distributional semantic models to predict the interpretations of a sample of newly derived *-ment* derivatives in their sentential context (55 types, 406 tokens from COCA). We then compared the predicted interpretations with the interpretations that we had annotated manually.

The paper is structured as follows. In the next section we will clarify the notion of disambiguation to set the theoretical scene for our investigation. Section 3 lays out our task, and section 4 introduces the data set and coding. Section 5 introduces our distributional semantic tools and their implementation, followed in section 6 by the presentation of the results. The final section summarizes and discusses our results.

2 Ambiguity, disambiguation and the construal of meaning

One of the central problems in the study of the semantics of derivation is ambiguity. That is, derived formations can be interpreted in more than one way (Rainer, 2014; Szymanek, 2013; Lieber, 2016; Plag and Balling, 2017; Andreou, 2017). There are several kinds of ambiguity that relate to the internal structure of derived words, to lexicalization, or to contextual factors.

One kind of ambiguity is structural in nature and arises through the possibility of assigning two different internal hierarchical structures to the same string. This is called 'structural ambiguity' (e.g. Szymanek 2013) or 'compositional ambiguity' (e.g. Löbner 2013, 48). For instance, the derived word *unlockable* may be interpreted in two ways depending on the internal hierarchical structure one ascribes to it. In particular, the bracketing in (2-a) derives the reading 'which cannot be locked', whereas the bracketing in (2-b) gives rise to the reading 'which can be unlocked'. (2-c) gives a syntactic example (in which it is not clear who has the telescope).

(2) a. [un-[[lock]-able]]
b. [un-[lock]]-able]
c. I saw the man with a telescope.

In this paper we do not address structural ambiguity.

Another kind of ambiguity relates to the semantic representation of lexical items and is often called 'lexical ambiguity. This type of ambiguity comprises two major sub-types, 'homonymy' and 'polysemy'. In cases of homonymy, two lexemes share all properties (e.g.

sound form and grammatical category) but have unrelated meanings. Thus, cases such as *bank*$_1$ 'a financial establishment' and *bank*$_2$ 'edge of a river' (paraphrases taken from the *Oxford English Dictionary*) are considered as distinct entries in the lexicon. In contrast to homonymy, in which different readings of the same form are attributed to different lexemes, in polysemy, the various readings are considered as variants of the same lexeme. Consider, for example, the derived *government*. It has a number of readings including 'the continuous exercise of authority over a person, group, etc.', 'a period of rule', and 'the body of people charged with the duty of governing' (paraphrases taken from the *OED*). Crucially, all readings are interrelated and considered as variants of the same lexeme, namely *government*.

Finally, ambiguity arises with the use of lexemes in context (Asher, 2011; Lieber, 2016; Andreou, 2017). The context may eliminate particular readings of a lexeme, or modify its meaning. In the case of meaning modification, the context enriches the meaning of a lexeme and triggers meaning shifts. This is often referred to as 'coercion'. The context, for example, may cause a lexeme to be interpreted metonymically or metaphorically.

Verbs may determine the meaning of the nouns that head their complements through selectional preference and vice versa ('co-composition', Pustejovsky 1995). For instance, in *bake the cake* the verb *bake* is interpreted as a verb of creation, whereas in *bake the potato* as a change of state verb. In other cases, referred to as 'co-predication' (Asher, 2011), not just one but both senses of a polysemous word are selected in the same utterance. Consider the sentence *Lunch was delicious but it took forever*: both the object and the eventive reading of *lunch* are selected in the same sentence.

Such examples of coercion or co-composition bring a further element into our disambiguation picture, which will be crucial throughout our study, namely the need of a pre-established set of senses to which the disambiguation target can be assigned. Whether one defines such senses as 'types' as in Pustejovsky (1995), or as 'aspects' (see Asher 2011), the introduction of a sense inventory allows the researcher to model the contextual dynamics of disambiguation in a more constrained way. Additionally, taking the perspective of disambiguation within a sense inventory uncovers a fundamental feature of polysemy: its systematicity. Some nouns (defined 'dot-types'or 'dual aspect'), come with a specified pairing of meanings (e.g., *book* as both object and piece of information) and in some contexts, crucially, both these meanings are selected at the same time. In *John read the book* the action of reading applies to both the object and the communication, as opposed to *The book fell*, in which only the object type is selected.

A basic difference between contextual ambiguity and polysemy is that in contextual ambiguity the various readings are not necessarily lexically stored but may be construed online. The distinction between contextual ambiguity and polysemy is fuzzy and, more often than not, a metonymical or metaphorical reading of a lexeme may be stored in the lexicon. This is, for example, the case with *government* in which various readings that are linked by meaning shift mechanisms such as metonymy were lexicalized and, thus, stored as meaning variants of the lexeme in question.

The problem of disambiguation is particularly salient in derived words since the increased number of meaningful elements and their interaction opens up even more possi-

bilities of interpretation than it is the case with monomorphemic words. For illustration consider the interpretation of prefixal negation with *un-*. A pertinent form, *un-diva*, is given with its context in (3).

(3) Dawn Upshaw has been called the "un-diva" of the opera world, often preferring to perform innovative, relatively obscure works that emphasize words over music in an informal style, often - imagine this - even chatting with an audience at recitals. (COCA SPOK 1994)

Dawn Upshaw, "the un-diva of the opera world", is actually a diva who breaks down the stereotype for the category DIVA. Thus, she is not a stereotypical member of the category DIVA. In (3), the use of the prefix *un-* with *diva* informs us that the derived word has a flavor of negativity, but it is not until the derived *un-diva* is embedded into the particular context in (3) that its reading is fixed. In particular, contextual cues such as "preferring to perform innovative, relatively obscure works that emphasize words over music in an informal style", and "even chatting with an audience" modify and enrich the meaning of the derived word *un-diva*.

This means that in (3), two kinds of ambiguity co-exist. First, there is the ambiguity of the prefix *un-*. This prefix has a number of readings. It may give rise to contrary and contradictory readings on adjectives (e.g. *unfriendly*, *undeniable*) and reversative readings on verbs (e.g. *unlock*). On nouns it derives stereotype negation and general negative readings. If one opts for a polysemous analysis of all these readings, then the various readings are treated as variants of the same prefix. If one opts for an analysis based on homonymy, then the various readings are attributed to more than one *un-* prefix. In any case, the interaction of the prefix with its base narrows down the range of possible interpretations.

The second kind of ambiguity is contextual in nature. It is the context that enriches the meaning of *un-diva* and guides us towards the right interpretation, i.e. stereotype negation instead of general negation (see Andreou 2017 for a discussion of stereotype negation).

Given these problems in interpreting derived words, one would like to know how exactly the context feeds into the interpretation of newly derived words. The literature on derivational semantics often discusses individual cases (see Bauer et al. 2013; Lieber 2016 for many examples), but there is no study available that investigates the problem on a more systematic empirical basis. This is the aim of the present paper.

3 Disambiguating *-ment* derivatives in context

The crucial assumption behind our approach is that the context reduces the number of possible readings by eliminating certain readings and, quite often but not necessarily, forces one reading for a given lexeme. This means that the various readings of an ambiguous lexeme appear in different types of context. Let us illustrate this with the monomorphemic lexeme *window* in (4), for which the *OED* lists more than 20 different interrelated meaning variants.

(4) a. They can say all they want, that this shows a lot of people are interested. But they've got a very short **window** here, a couple of months. And if they don't get the 7 million people, including a large number of healthy people, this system will not work. (COCA SPOK 2013)

b. I clicked the SUBMIT button. Nothing happened. Not at first. The program opened a new **window** and a message appeared across the top. (COCA FIC 2015)

c. He has no dog in the fight and can't understand why the city would tear it down. "They are harassing poor people about putting on a new **window** - well a new window is a big project," he said. Blanchard knows people who were forced to sell because they couldn't keep up with Pagedale's code. (COCA NEWS 2015)

As evident in the examples in (4), the meaning variants of *window* may appear in different types of contexts. Thus, based on the contextual information "a couple of months", we infer that the meaning variant of *window* that appears in (4-a) relates to time, and can be paraphrased as 'an interval of time which affords an opportunity to perform a particular action, or within which a particular action must be performed'. In (4-b), several contextual cues such as "clicked the SUBMIT button" and "the programm opened" relate to computing, and force the reading 'a rectangular, typically framed area of the display screen that is produced by a graphical user interface in order to display information, an image, or an interface for an application.' Finally, the contextual information "tear it down" and "putting on" in (4-c) guide us towards the reading 'an opening in the wall or roof of a building, for admitting light or air and allowing people to see out; ... the glazed frame intended to fit such an opening'.

As pointed out in section 2, the issue of ambiguity is particularly relevant for words that result from the application of a derivational process to a base. The role of the base in creating multiple meanings is, however, not restricted to potential consequences of base polysemy. The semantic representation of the base itself may also be the source of competing interpretations. This can be nicely explained in terms of referential shifts (Löbner 2013). Derivational semantics can be conceptualized as meaning shifts in semantic representations (e.g. Kawaletz and Plag 2015). In particular, reference is shifted from the original referent of the base to a new referent of the derivative. For example, in the case of agentive *-er* formations reference can be shifted from the event (expressed by the verb) to the subject argument of the verb, or to other entities involved in the event (e.g. its location, as in *diner*, or an instrument, as in *printer*). For *-ment*, it has been shown that such shifts away from the verbal eventive interpretation can target, for example, the stimulus argument or the result state in the case of psych verbs (Kawaletz and Plag 2015). Other types of verb may have other kinds of semantic entities to which reference can shift, e.g. the patient in the result state (Plag et al. 2017). In this way, the semantics of the verbal base has an impact on which readings are possible or likely with *-ment* forms and which ones are not.

In this paper we will empirically explore the extent to which the context surrounding newly derived words with the suffix *-ment* provides the necessary cues for disambiguation.

The task at hand is particularly difficult because newly derived words are often ambiguous but listeners do not have stored meaning variants for them.

We focus on two readings, EVENTIVE and NON-EVENTIVE. In the examples below, the newly derived *emplacement* in (4-a) has an eventive reading, and *bedragglement* in (4-b) has a non-eventive reading.

(5) a. **Eventive reading**
In many places, **emplacement** of granite plutons is synchronous to volcanic eruptions. (Google Website 1995)

b. **Non-eventive reading**
I set down the scrap of dolls dress, a **bedragglement** of loose lace hem (COCA FIC 1999)

As the example of *The Guggenheim* and *construction* discussed in (1) has shown, derived words may not always be disambiguated. Thus, they may remain ambiguous between various readings. In (6), for example, the context surrounding *worsenments* does not disambiguate between an eventive and a non-eventive reading.

(6) **Ambiguous reading**
Yes, in an ideal world the 40D would have improvements over the 30D and no *worsenments*. (Google FORUM 2007)

Derivatives in *-ment* are thus an ideal testing ground for an investigation of the problem of how derived words can be disambiguated.

4 Data set and coding

4.1 The suffix *-ment*

The suffix *-ment* is frequent in contemporary English. This is because it was very productive in the 15th, 16th and 17th centuries (see Marchand 1969; Lindsay and Aronoff 2013). Today, many researchers hold it to be unproductive (e.g. Bauer 1983, 2001; Schmid 2011), but recent corpus studies have shown that numerous novel words can indeed be identified in large corpora such as COCA (The Corpus of Contemporary American English, Davies 2008; see Bauer et al. 2013; Kawaletz and Plag 2015). This finding suggests that still today speakers utilize this suffix in the creation of new words.

The suffix *-ment* prefers Romance verbal bases (*abandonment*) but also attaches to Germanic verbs (*amazement*) as well as other categories such as adjectives (*foolishment*), nouns (*illusionment*), and bound roots (*compartment*, see Bauer et al. 2013, 198). Derivatives in *-ment* exhibit a large range of readings (see Bauer et al. 2013; Kawaletz and Plag 2015), as shown in (7):

(7)

events	*assessment*
results	*improvement*

states	*contentment*
products	*pavement*
instruments	*refreshment*
locations	*embankment*
patient/theme	*investment*

Derivatives which are long since established are often highly lexicalized and may show all kinds of idiosyncrasies. We are, however, interested in the productive derivational process, not in opaque forms. That is, we want to know how today's speakers use the suffix when they form or try to understand new words. Therefore the focus of this study is on neologisms and very rare forms. These are usually transparent in order to enable successful communication (e.g. Plag 1999). The dataset for the present study is taken from Kawaletz (2018). It was built using various corpora and data bases, as described below.

4.2 Sampling newly derived words in *-ment*

A major and established source in the search for neologisms is the OED. With 600,000 words and 3 million quotations, the OED is an exceptionally detailed and comprehensive dictionary of the English language. It is continuously updated with new words and usages of existing entries, giving dates of first citation for every sense in which a lemma is attested. It is therefore a convenient tool for the identification of neologisms.

A list of entries containing possible neologisms was retrieved using the interface provided by the OED. Using string search we looked for words with first citations dating from 1900 to today (see, for example, Plag 1999 for a similar procedure). The resulting list of 134 types of raw data was then subjected to a standard revision procedure, weeding out non-pertinent data. 18 derivatives remained after cleaning.

In addition to the OED neologisms, the data set was substantially extended by extracting hapax legomena and other very rare forms from COCA. Hapax legomena (or 'hapaxes', for short) are words which occur only once in a given context, such as, in the present study, a corpus. Hapax legomena are central in the study of productive derivational processes. It can be shown that the majority of neologisms in any given corpus is contained precisely in this group of hapaxes (see Plag 2003, 68). This means that hapaxes are a valid source of neologisms.[1]

Three tools were employed in this step of the process: COCA (Davies 2008), VerbNet (Kipper et al. 2008) and Coquery (Kunter 2015). With more than 450 million words written and spoken between 1990 and 2012, COCA is an appropriately large corpus for the identification of hapaxes as potential neologisms. VerbNet is a hierarchical verb lexicon of – at the time that this is written – 6088 English verbs. It is based on the verb classification developed in Levin (1993) and includes syntactic and semantic information. The corpus query

[1]Note that it is not claimed that every hapax is indeed a neologism. In fact, a large number of hapaxes are actually very rare or specific technical terms, archaisms, non-transparent ad hoc inventions, typing errors or other kinds of errors. The size of the corpus is also a decisive factor: The larger the corpus, the higher the proportion of neologisms among the hapaxes (see Baayen 1996, Baayen 2009).

tool Coquery was used to conduct an automatized search of the DVD-version of COCA (Davies 2008) for all verbs listed in VerbNet, plus the orthographic string <ment*>. Those with a frequency of 1, 2 or 3 were considered for further investigation.[2] A total of 158 types of raw data could be identified, which was reduced to 117 types after cleaning.

A second COCA search was conducted in order to identify *-ment* derivatives which are not formed on the basis of verbs already listed in VerbNet. Using the web interface provided by Brigham Young University[3], COCA was searched for all words with a frequency of 1 or 2 ending in either <ment> or <ments>. The resulting list of 5142 types had to be weeded heavily. Of the remaining 109 types, 60 were based on verbs not listed in VerbNet and could thus be added to the dataset.

Merging the COCA dataset with the types from the OED led to a data set of 244 derivatives (i.e. types) based on 36 different verb classes. This dataset was reduced to a more manageable number, containing only types based on the four biggest represented base verb classes: Psych verbs, change of state verbs, putting verbs and force verbs. The final data set contains 55 types. The words from COCA were all produced between 1990 and 2012 (the complete range of the corpus); the oldest attestation from the OED dates back to 1900, while the most recent one was produced in 1961.

4.3 Sampling attestations

The ambiguity of derived words in context is of course also pertinent when investigating hapaxes, which are by definition attested only once in a given corpus. In any such unique attestation, one of two things may happen. If the hapax is unambiguous in the given context, it is impossible to know which further readings are conceivable. If the hapax is ambiguous in its context, it cannot be determined which meaning was intended by the speaker. A related problem occurs in the dictionary data. Although the OED aims at wide coverage, for obvious reasons it does not include every meaning variant ever attested. Since, however, it is exactly these innovative, spontaneous, and fully transparent formations that are of interest for this study, it seemed necessary to enhance the data set with further attestations.

To this end, further corpora such as WebCorp (Renouf et al. 2006), GloWbE (Davies 2013), or Google were searched for additional attestations of the types sampled so far. The sampling of attestations was not random, as we specifically looked for attestations that reflected the polysemous nature of *-ment* derivatives. For example, the first attestations sampled for a given word might have been all eventive. We then searched until we found words with other possible interpretations. This biased sampling procedure was necessary

[2]The reason why not only hapaxes, but also dis and tris legomena were included is that the search results may be corrupted in various ways, "hiding" actual hapaxes. Take, for instance, the case of *musement*: The noun is listed with a frequency of 2 in COCA, but one of the attestations is actually *bemusement* with a wrongly placed space ("be musement"). Also, it occasionally happens that the very same context is listed twice. By including dis and tris legomena, the chances of avoiding these problems and thus finding a larger number of pertinent forms can be increased.

[3]http://corpus.byu.edu/coca/

to ensure that we have enough eventive and non-eventive, abstract and concrete readings in our data set.[4] The final data set comprised 406 attestations.[5]

4.4 Coding

All tokens were manually categorized according the binary distinction between EVENTIVE vs. NON-EVENTIVE readings, corresponding to our distinction of interest. In addition, we annotated a binary distinction between ABSTRACT vs. CONCRETE readings. This additional coding was performed to enable more detailed investigations in subsequent analyses (cf. Section 6.2).

The EVENTIVE category is taken here to cover all kinds of events, processes and states. NON-EVENTIVE readings comprise both concrete entities such as objects, animals or persons, and abstract entities such as quantities or means of communication. Underspecified contexts were classified as AMBIGUOUS. There were also a few cases where, regardless of context, it was impossible to assign the word to any of the three categories because it was conceptually unclear to which category (EVENTIVE or NON-EVENTIVE) it should belong in the first place.

Tokens denoting an idea, quality, state or event were coded as abstract. All eventive nouns therefore automatically fell into the ABSTRACT category. Concrete nouns, on the other hand, are those denoting entities existing in a physical form.

The resulting distribution of classes across the data set is given in table 2.

Table 2: Distribution of semantic categories in the dataset, cross-classified

	total	ABSTRACT	CONCRETE	AMBIGUOUS
EVENTIVE	275	275	0	0
NON-EVENTIVE	70	15	49	6
AMBIGUOUS	55	1	0	54
UNCLEAR	6	6	0	0
total	406	297	49	60

We also coded *-ment* derivatives for the type of base verb they feature. This was done according to the verb classes proposed by Levin (1993) (and extended in the VerbNet

[4]The distribution of readings in our data set can therefore not be taken as being representative of the distribution of these readings in the language at large.

[5]Web-search tools such as Google exhibit certain shortcomings in the context of serious linguistic investigation (e.g. almost unlimited corpus size, no data organization, no annotation, no control about the origin of the data). However, it has also been shown that they can be a convenient indicator for innovative language use (see Diemer 2011, and the papers in Hundt et al. 2006). In order to meet the requirements of academic research as well as possible, any indication that the author of a given text might not be a native speaker of English was taken as a reason to exclude this attestation. For this, the wider context was scanned for grammatical errors, awkward formulations or straight-forward indicators of the country of origin.

project Kipper et al. 2008). The base verbs in the final data set come from four verb classes: psych verbs, verbs of change of state, putting verbs, and force verbs. All four verb classes feature eventive, non-eventive and ambiguous readings.

Let us now turn to our disambiguation task. In order to explore the extent to which the context can resolve the meaning ambiguity exhibited by newly derived words, our analysis and implementation proceeded as follows. First we obtained distributional semantic vectors for typical event nouns and typical entity nouns (our 'training set') based on a source corpus (a concatenation of the British National Corpus and UkWaC, 2.6 billion tokens in total). We then trained a classifier algorithm to detect the class of a given word (i.e. eventive or non-eventive) in the training set based on the vector of this word. In the second step ('disambiguating'), we tested how well our trained classifier can disambiguate the *-ment* derivatives in our data set, using the vectors we obtained for these derivatives from their rare attestations in our sources.

5 Implementation

In this section we first prepare the ground for our empirical investigation by introducing Distributional Semantics. This is followed by a detailed description of our work-flow (i.e. training and testing).

5.1 Distributional Semantics

Distributional Semantics is the name of a family of approaches to semantics that rest on the assumption that the meaning of words is correlated with the ranges of contexts in which it can appear (e.g. Wittgenstein and Schulte 2001; Harris 1954; Firth 1957). As a computational approach it seeks to formalize this idea in such a way that it becomes mathematically tractable in the form of vector representations. In the simplest type of distributional model, the meaning of a word (referred to as 'target') is represented as a list of numbers, i.e. a numerical vector, which represents the word's frequency of occurrence with all other words (the 'contexts') in a given corpus within a specified window to the target word's left and right. This type of distributional model (both targets and contexts are words) is defined as 'bag-of-words' model (see Turney and Pantel (2010) for an overview of other types of model). The vectors are collected in a matrix, with each row of the matrix corresponding to the distributional vector for a target word. Table 3 gives a toy example of such a matrix. It lists the frequency of co-occurrence of four target words (*t-shirt, tie, lawyer* and *judge*) with two context words (*wear* and *law*).

Context \ Target	*t-shirt*	*tie*	*lawyer*	*judge*
wear	9	7	2	4
law	1	3	7	9

Table 3: Matrix with frequencies of co-occurrence for four target words

The matrix of table 3 can also be represented graphically as in Figure 1.

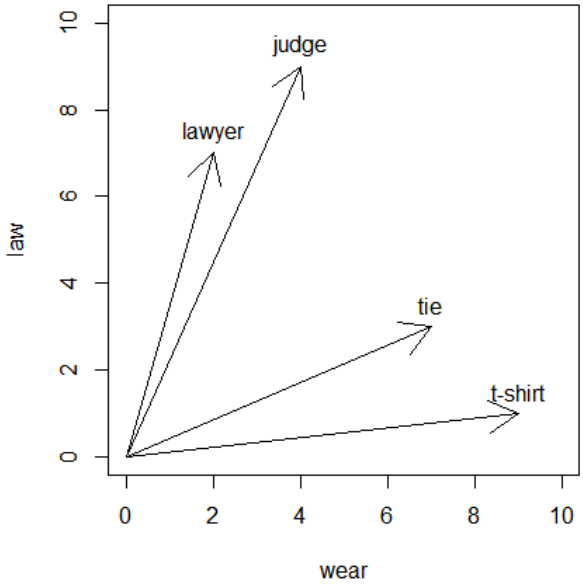

Figure 1: Lemma vectors for eight word pairs

In figure 1 one can easily see that the vectors for the words *lawyer* and *judge* are much closer to each other than to the words *tie* and *t-shirt*. One can also see that *tie* and *t-shirt* are very close to each other, and that *tie* is closer to *lawyer* and *judge* than *t-shirt* is. The distance between the vector of two target words can be interpreted as the empirical correlate of the amount of meaning they share, that is to say, their similarity. The closer two vectors are to each other, the more similar the words are semantically. This is in accordance with the traditional insight that words that are similar in meaning are likely to occur in similar contexts. The vector-based quantification of similarity can be implemented mathematically in different ways, for example by using the cosine and the Euclidean distance, or by other measures used in mathematics for the purposes of comparing vectors.

While computation of pairwise similarities is one of the typical applications of Distributional Semantics, this is not the approach employed in this study, because our aim is

to learn typical contextual properties of semantically motivated groups of lexical items (i.e., eventive vs. non eventive). In the what follows, we provide a high-level sketch of our approach to the disambiguation of *-ment* derivatives.

Lemma and instance vectors In a distributional semantic model as described above, distributional vectors are typically collected for word types, that is at the lemma level: we therefore in what follows will refer to such vectors as 'lemma vectors'. Lemma vectors are aggregates over all instances of a given lemma in a given corpus in a pre-defined contextual window (for example five words before and after the word in question). Lemma vectors thus represent the usage of a word (i.e. a lemma) across a potentially large number of occurrences. This is the main strength of distributional semantic models, but it also brings in some limitations. First, as each lemma word is represented by a unique aggregated distributional vector extracted from all its contexts, the senses of polysemous words are conflated into the same representation. Second, if a lemma is not frequent enough, its vector representation will be not reliable, since most potential context words will simply not be attested; from this perspective, unattested words represent an extreme case of low-frequency: in this case, it is just impossible to collect a lemma vector.

Our study faces both problems, as the *-ment* derivatives in our dataset are both infrequent and prone to different readings. Our approach builds on a very common solution proposed in the distributional semantic literature for the polysemy issue, which, crucially, turns out to be a good solution for the low-frequency issue as well. A way out of the polysemy problem is to compute so-called 'instance vectors', that is, vector representations for individual instances of words, i.e. tokens, rather than lemmas, i.e. types. This approach not only addresses the sparsity problem for the rare words that we consider in this article, but also enables us to model context-based disambiguation phenomena for individual occurrences of such words.

To compute instance vectors, we adapt the proposal by Schütze (1998). We consider a window of context words around our target words. We are only using content words for the instance vectors (nouns, verbs, adjectives, adverbs). This means that we remove non-content words when we calculate the window. The words from the context window form a vector. Each word in this vector is then replaced by its distributional semantic lemma vector. The resulting set of vectors is then transformed into a new vector by averaging all vectors. We thus obtain a vector for each instance in its specific context, i.e. our instance vector. With these instance vectors, further computations become possible, for example the comparison with other instance vectors of the same lemma, or classification with respect to more abstract semantic categories. Note that this computation, crucially for our purposes, does not require a vector for the target word itself – the model can be thought of performing a kind of cloze test (Taylor, 1953).

Consider Figure 2. It gives the fictitious instance vectors in red color for two tokens of *suit* (*suit1* and *suit2*) as they occur in the sentences given in (8).

(8) a. The suit is next to the tie and the t-shirt.
    b. The lawyer filed a suit to the judge.

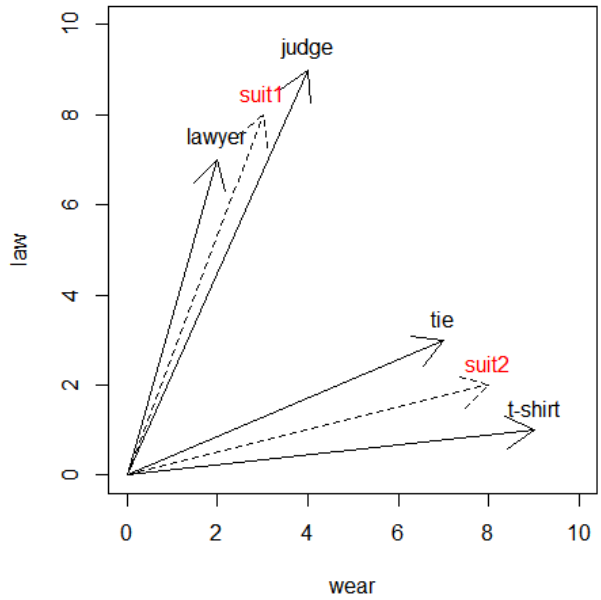

Figure 2: Sentence vectors for two tokens of *suit*

In this toy example, the two instance vectors are able to discriminate between the two senses of *suit*. In a similar fashion, we hypothesized, instance vectors can be used to discriminate between potential eventive vs. non-eventive readings of a given word in its context. For example, one can assume that the verb *happen* is likely to occur in the context of the eventive noun *accident* but not so likely to occur in the context of an entity noun such as *desk*. Thus, the presence of *happen* in a given context can be used to predict the semantic class of nouns in its context.

The notion of context Before going into more details of our approach to disambiguation, let us provide a few clarifications concerning the notion of context our study relies on, from a distributional semantic perspective. In our theoretical discussion, we have used this term mostly in the sense of 'syntactic context'.[6]

[6] A more general note of caution is in order concerning the notion of 'context'. The examples in (1) and (6) have shown that such a restricted notion of context is probably too limited. The fact that the semantics of some derivatives may not be fully determined even when they are embedded in a particular syntactic context does not mean that the ambiguity of *construction* in (1) and *embrittlement* in (8) is not resolved at all. In particular, there are factors other than the immediate sentential context that may work to resolve ambiguity. Such factors involve more general discourse contexts and common ground between the speaker and the addressee. These factors are very hard, if not impossible to investigate systematically on a larger scale. In the present study, we restrict ourselves to the immediate sentential context as described above, but will keep track of how the algorithm treats readings that have been identified as ambiguous in our manual coding.

Our bag-of-words approach, however, collects all the contexts within a context window (e.g., five content words to the left and right of the target) irrespective of syntactic structure proper. While contexts which are considered as prototypically syntactic are very likely to occur in that window (for instance verbs of which the target noun is the head of an object or subject noun phrase), this is not guaranteed (for example due to long distance dependencies). A whole set of contexts which are not part of the argument structure of the target will also contribute to its distributional representation, for example coordinates (as in *dogs and cats*) and topically related words as in *the bread in this bakery*).

5.2 Semantic class disambiguation as supervised classification

In this study, we implement semantic class disambiguation as a supervised classification task. Supervised classification is a machine learning technique which assigns each item (in our case, the *-ment* instances in the dataset) to one class (in our case either EVENTIVE or NON-EVENTIVE), based on a generalization built from a series of training examples. Training examples are pairs consisting of an item with its corresponding class which are considered to be a reliable representation of the generalization we are trying to build. Once the generalization has been built, it can be applied to predict a class for new items.

Applied to our concrete task, recall that distributional representations (i.e. our vectors) can be interpreted as coordinates in a multidimensional space. In this geometric conceptualization, what the classifier learns is a boundary in the space (cf. Figure 1) that separates the instances of the two classes (EVENTIVE VS. NON-EVENTIVE) from one another as well as possible, and can make predictions for novel items by checking on which side of the boundary the item lies.

Let us now turn to the different steps we took in our study. Figure 3 provides a visual representation of the workflow. In order not to clutter the presentation, technical details concerning the computational implementation are not discussed in the following passages, but are spelled out in Appendix A.

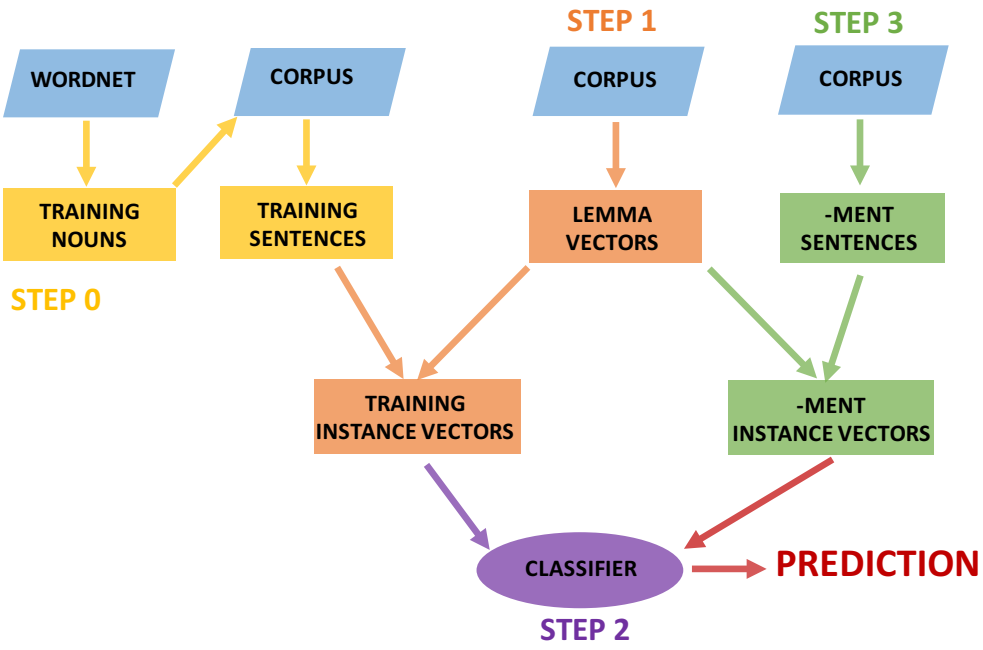

Figure 3: Workflow of the implementation. Blue boxes represent external resources. The different steps are represented by color-coded boxes and arrows.

The following paragraphs will take the reader through the steps shown in Figure 3. While the tools we employ are known and well-established, our study integrates them for a new enterprise, the exploration of the semantics of derivational morphology. In doing so, we provide a contribution to the exploration of the semantic representation of the long tail of the frequency distribution, populated by large amounts of low-frequency items.

Step 0: Collecting training nouns and training sentences As discussed above, supervised classification requires examples that can be used for training. These examples serve as the source of the generalization we are going to apply to the *-ment* derivatives.

The typical procedure in classification is to use a part of the dataset for training and another for testing. However, since our focus is on meaning construal for rare words, we use a different route: We first train a classification model on a set of nouns that are independently classified for their senses, and then apply this model to novel, rare derivations.

There is a rich body of literature on lexical disambiguation (Word Sense Disambiguation) with distributional semantic methods, see e.g. Navigli (2009) for an overview. These studies typically use the sense inventory provided by a lexical resource such as WordNet (Fellbaum, 1998), a large-coverage lexical data base for English. Even though we do not use classical Word Sense Disambiguation methods in this study, we will also make use of WordNet as a source of semantic class information. With that decision we follow studies that have used distributional methods in conjunction with top-level WordNet classes to model systematic polysemy (Boleda et al., 2012a), class-based (i.e. abstract) Word Sense Disambiguation (Izquierdo et al., 2009), or lexical acquisition (Joanis et al., 2006).

From WordNet we collected frequent, unambiguous lemmas as training nouns for WordNet's semantic classes EVENTIVE, NON-EVENTIVE, STRICT OBJECT, LIVING THING, and LAX OBJECT. We chose the rather broad categories of EVENTIVE and NON-EVENTIVE as well as more fine-grained subcategories of NON-EVENTIVE (i.e. STRICT OBJECT, LIVING THING, and LAX OBJECT) to be able to try out different dichotomies for the classification task (see Sections 6.1 and 6.3 for the different outcomes). For these nouns we then extracted sentences containing nouns from a large reference corpus.

Step 1: Obtaining distributional representations for training nouns and training sentences In this step, distributional representations for the training nouns (i.e. lemma vectors) are built, followed by the computation of representations for individual occurrences of these nouns (i.e. instance vectors), all in accordance with the standard procedures in distributional semantic modeling outlined in Section 5.1.

Step 2: Learning to distinguish eventive and non-eventive nouns The instance vectors are employed to train a classifier with supervised learning methods. The input of a classifier in the training stage is a set of distributional vectors with associated classification labels. We draw our training examples for the eventive class from the set of occurrences of the non-ambiguous event nouns in the training set. Correspondingly, the training examples for the non-eventive class come from the set of non-ambiguous non-event nouns.

Step 3: Disambiguating *-ment* derivatives Now we can apply the classifier to disambiguate *-ment* derivatives. We first compute an instance vector for each *-ment* derivative in our data set. The classifier is then asked to predict a class label for each instance vector. The predicted class is then compared to the manual annotation described in section 4. Note that, for the quantitative evaluation of the classifier, we focus on the predictions for the *-ment* derivatives which have been classified as eventive or non-eventive, because this is the distinction we trained our classifier on. We can, however, use the classifier to assign the *-ment* derivatives classified as ambiguous by the human annotators to either an eventive or non-eventive reading, and interpret this prediction qualitatively (see section 6.2).

A commonly used measure to assess the performance of a classifier is the F-score. F-scores combine two different measures of performance, recall and precision. Recall is the ratio of the number of items for which the model correctly predicts a given category, say EVENTIVE, and the number of items which are indeed EVENTIVE. Precision is the ratio of the number of items for which the model *correctly* predicts a given category, say EVENTIVE, divided by the number of items for which the model predicts this category. The F-score is the harmonic mean of precision and recall. Since we have two categories that are predicted, two F-scores are computed in the way just described, one for each category. To compute an F-score for the whole classification, we average the two F-scores. This procedure, called 'macro-averaging', is the most conservative manner of computing overall F-scores.

F-scores range between 0 and 1. In a balanced data set with an euqal distribution of two categories, the baseline is 0.5 and any higher F-score value indicates some improvement

over a random decision. In unequally balanced data sets the baseline is set by the more frequent category. For example (and abstracting away from F-scores), in a data set with a 60% vs. 40% distribution of items across two classes one would have a 60% chance of getting it right if one simply assigned all items to the majority category. The baseline for any improvement introduced by a classifier would therefore be 60%.

Summing up, the evaluation of the classifier on the manual annotation allows us to ascertain how successful the algorithm is in finding the right interpretation.

6 Results

This section reports the results of our classification experiments with rare words in respect to the eventive/non-eventive distinction. We first report the outcome of the main experiment in which we vary the semantic classes that we use for training (Section 6.1). We then present more detailed analyses in which we address two more specific questions:

- Which classes are modelled well or badly, and why? (Section 6.2)
- Would lemma vectors theoretically be able to do a better job? (Section 6.3).

6.1 Main experiment

As described in Section 5.2, we trained our classifiers to distinguish the instance vectors of the nouns for the EVENTIVE and NON-EVENTIVE classes, respectively, and then tested the classifiers on the instance vectors of unambiguous *-ment* derivatives.[7]

Initial experiments with this classifier showed reasonable, but not overwhelmingly good results. Our working hypothesis was that this was a natural consequence of both the inherent difficulty of our data and the very heterogeneous nature of the NON-EVENTIVE class. This class comprises as ontologically diverse subcategories as abstract nouns (*liberty*), artifacts (*coffeemaker*), and animals (*dog*). Arguably, it is difficult for distributional models to learn context cues that are equally valid for all of these nouns.

We therefore decided to assess to what extent this heterogeneity problem can be addressed by a more focussed training of the classifier, using more specific classes. Specifically, we trained three additional classifiers for the more specific (and therefore, hopefully simpler) distinctions between (a) EVENTIVE and LAX OBJECT, (b) EVENTIVE and STRICT OBJECT, and (c) EVENTIVE and LIVING THING. The three categories of LAX OBJECT, STRICT OBJECT, and LIVING THING as used by WordNet can be illustrated with abstract nouns such as *liberty* for the LAX OBJECT category, concrete nouns such as *stick* for the STRICT OBJECT category, and animate objects such as *dog* for WordNet's LIVING THING category. The results are shown in Figure 4.

[7]We exclude the ambiguous *-ment* derivatives from the numerical evaluation, since our two-class model is not able to make correct predictions for them. We include the ambiguous derivatives in our subsequent analysis, however. It is generally far from trivial to devise a meaningful model for properly ambiguous lexical items, cf. the studies by Boleda et al. (2012b) for polysemous adjectives.

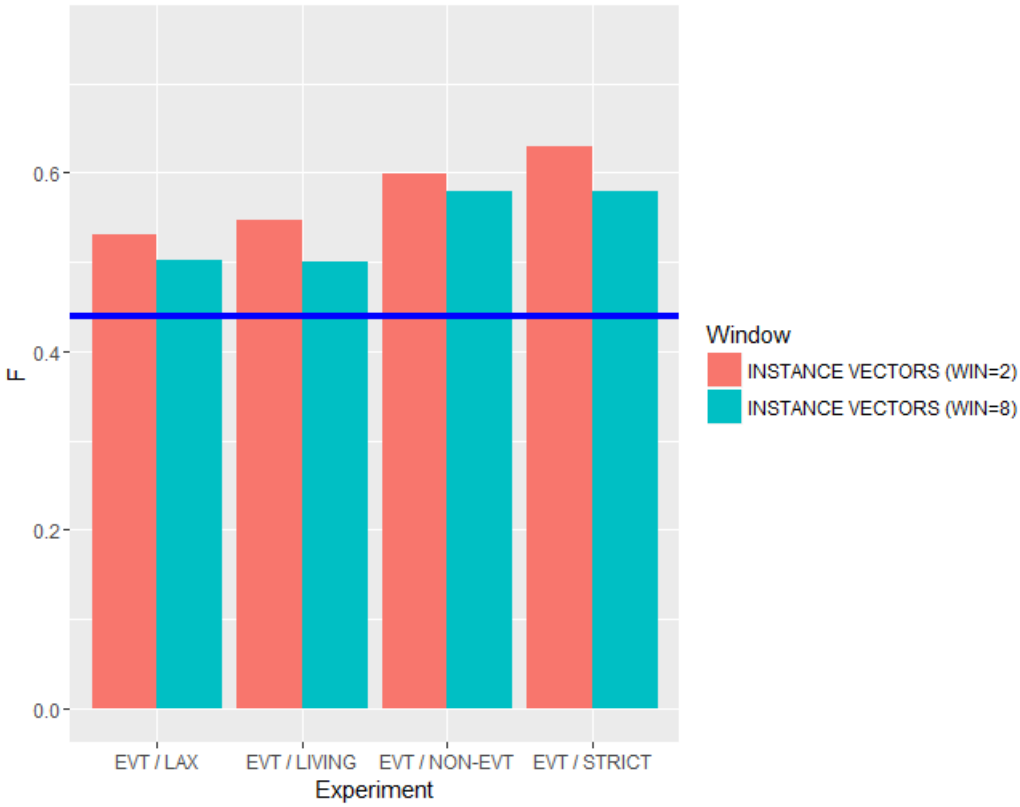

Figure 4: Performance of the classifier in disambiguating *-ment* derivatives, with the classifier having been trained on different NON-EVENTIVE subclasses.

The majority class baseline (always assigning EVENTIVE) comes in at an F-Score of 0.44 (horizontal blue bar). All four classifiers perform well above the baseline, establishing that our model is learning something meaningful about the EVENTIVE–NON-EVENTIVE distinction. The most general model, which is trained on all kinds of NON-EVENTIVE, is shown in the second pair of bars from the right. This model performs at 0.6 F-Score (using a 2-word window). The settings using only LAX and LIVING non-eventive nouns do somewhat worse, presumably because these subclasses are too specific to generalize well to the test set. We find the best overall performance for the classifier trained on EVENTIVE / STRICT OBJECT data, with an F-score of 0.63. Evidently, this setup strikes the best balance between learning a simplified distinction that does not generalize well, and learning a distinction that is very difficult.

As for the size of the context, we see only small differences betweena very narrow context (window size 2) and a relatively broad context (window size 8). Across all classifiers, there is a small advantage for the narrow context. How can we understand this, perhaps surprising, result? Previous studies targeting evaluation of bag-of-words distributional semantic models have explored the modulation of the performance in connection with the manipulation of the window size (Sahlgren, 2006; Lapesa et al., 2014). Generally speaking, their take home message is that smaller windows bias distributional representations

towards true similarity (assigning high similarities to synonyms), while larger windows bring topical relatedness into the picture (assigning high similarities to syntagmatically related words). In this connection, the slightly detrimental effect of larger windows seems to suggest that broader topical relatedness does not support *-ment* disambiguation.

6.2 Analysis 1: Per-class performance

Let us look at the results in more detail to learn more about how context predicts the different interpretations for all items, including the ambiguous ones. To simplify the exposition, we will from now on focus on the best classifier (i.e. the one trained on EVENT / STRICT OBJECT wiht a 2-word window) and see how it performs for the three classes (EVENTIVE, NON-EVENTIVE, AMBIGUOUS). We omit the six cases of nouns annotated as UNCLEAR since this class is too small to allow for any generalization. Percentages are given in Table 4, and Figure 5 plots the distribution of the manual annotations against the predicted interpretations for easier reference. The mosaic plot shows the areas in proportion to the number of observations in the data set.

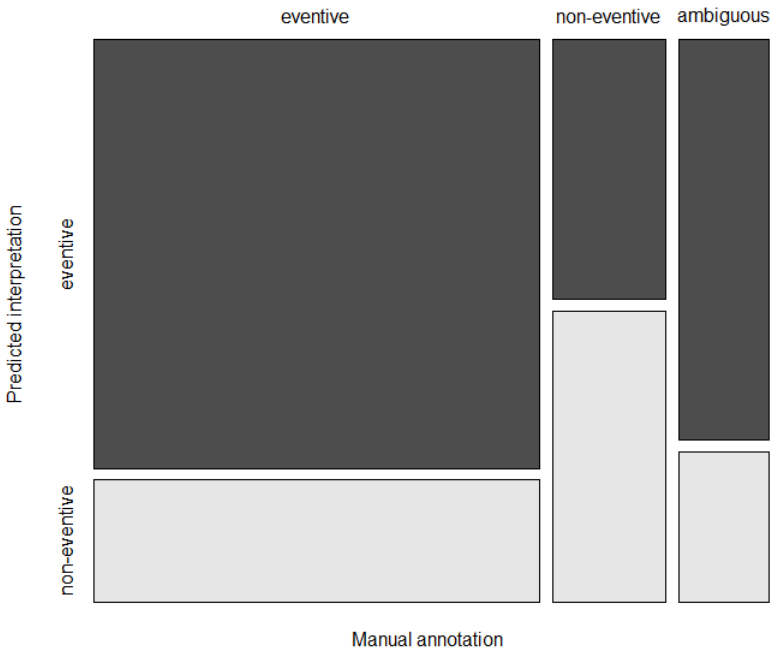

Figure 5: Predictions vs. annotations

Table 4: Proportions of predicted interpretations vs. manual annotations (figures are rounded)

Predictions \ Annotations	eventive	non-eventive	ambiguous
eventive	0.78	0.47	0.73
non-eventive	0.22	0.53	0.27

We can see that 78 percent of the manually annotated events are also classified as events by the classifier. This means that the context, even if very small, provides rather robust cues for finding the right interpretation with eventive nouns.

For a more detailed analysis, we can ask the classifier to provide a probability for each prediction, which enables us to zoom in on cases that it is more certain or less certain about (see Appendix A for details). A look at some examples with a high probability for an eventive reading supports our ideas about the context (emphasis ours, probabilities given in parentheses):

(9) a. I got over that *initial moment* of **dumbfoundment** (I'm making up my own words today) (0.84)
b. We can learn to let go of the agitated states of mind, such as *anger*, **worriment**, *resentment* and *fear*, that produce unhappiness (0.94)
c. As a population that had immediate and intimate access to aboriginal peoples, half-breeds stimulated and intensified anxieties regarding the deleterious effects of alcohol on Indians, and how drunkenness might trouble their moral **upliftment** and *eventual* assimilation into white society. (0.96)
d. Hydrogen, especially atomic hydrogen, is particularly dangerous because it tends to *cause rapid* **embrittlement** even at low temperatures. (0.96)

One can see in each example that the context provides strong cues towards eventive readings. In all examples one can either find other event nouns (e.g. states as in (9-b)) or temporal expressions that point towards events (*initial moment* in (9-a), *eventual* in (9-c), and *cause rapid* in (9-d)).

With non-eventive readings, the prediction is much harder. Only 53 percent of the manually annotated non-eventive readings are successfully categorized by the classifier. In order to understand the discrepancy in success between the classification of eventive nouns and that of non-eventive nouns it is instructive to look at the distributions of the probabilities predicted by the probabilistic classifier (instead of the categorical decisions of the binary classifier).

Figure 6 shows the three distributions of probabilities.

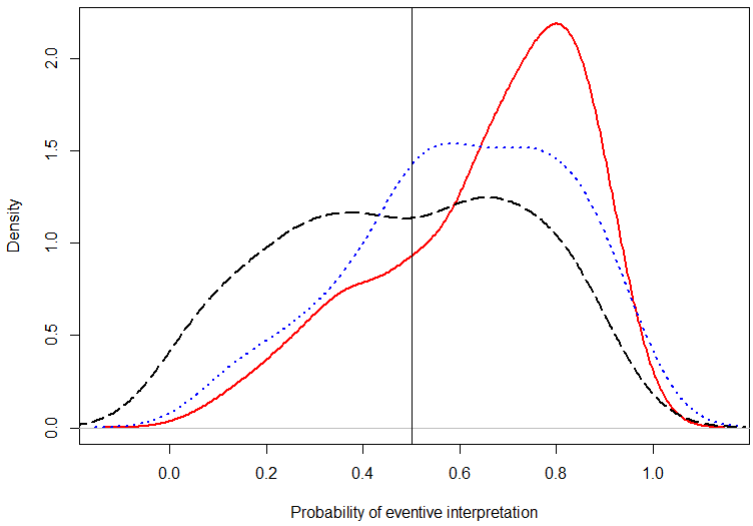

Figure 6: Distribution of predicted probabilities for *-ment* derivatives that are manually annotated as EVENTIVE (unbroken red line) vs. NON-EVENTIVE (dashed black line) vs. AMBIGUOUS (dotted blue line).

The solid red curve (nouns annotated as eventive) is highly right-skewed and has a very clear maximum at about 0.8 and only a little bump at about 0.4. Its shape, location and skewedness indicates that eventive nouns are classified with considerable success as eventive. The dotted blue curve (nouns annotated as amabiguous) has two very slight bumps close to each other at probabilities at around 0.6 and 0.8. This distribution is still normal, as shown by a Shapiro-Wilk normality test (W = 0.96, p-value = 0.08). The black dashed curve, in contrast to the other two curves, shows a clearly bimodal distribution, with two peaks at about 0.38 and 0.65, respectively. Such bimodal distributions are often indicative of an underlying dichotomous distinction, with each sub-class patterning differently. For the right-hand sub-class of nouns, that is the class with the peak at a probability of 0.65, we get a tendency towards an eventive reading. For the left-hand class there is tendency of nouns to be classified as non-eventive.

Which sub-categories of non-eventive nouns might be underlying this bimodal distribution? We hypothesize that the distinction between abstract and concrete objects (cf. Section 4) might be responsible for the bimodal distribution, with the concrete nouns being more prone to correct object classification, and thus on the left-hand side of the plot. To test this assumption, we checked the probabilities for the nouns of these two subclasses. The distributions are shown in figure 7. The boxes contain the central 50 percent of the data, the dots indicate the medians. The whiskers go up and down 1.5 times the interquartile range plus the respective quartile.

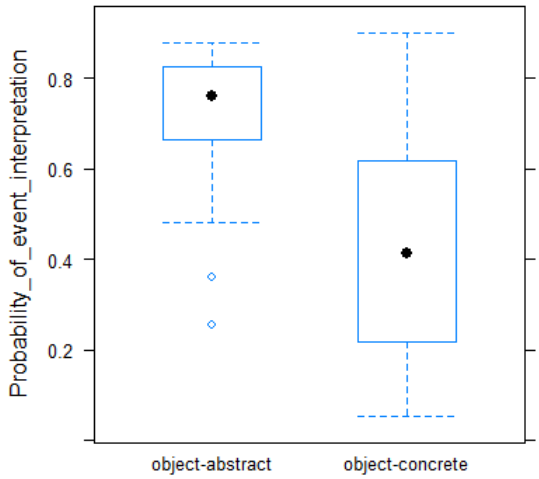

Figure 7: Distribution of predicted probabilities for concrete vs. abstract non-eventive -*ment* derivatives

The two distributions of the probabilities are indeed significantly different ($D = 0.64$, $p < 0.001$, Komolgorov-Smirnoff test). In accordance with our hypothesis, the abstract nouns have a mean probability of 0.70, which means that they have a strong tendency to be classifed as events, while the concrete nouns have a mean probability of 0.41, which means that they tend to be classified as objects. The difference in the means is statistically significant ($W = 609$, $p < 0.001$, Wilcox test).

For illustration consider the following examples of concrete nouns in (10) (probabilities of eventive classification are given in parentheses). The examples show that the context has no clear cues for an eventive interpretation but some cues for an object interpretation. In (10-a) the noun *concrete* is used in the two word window, in (10-b) *cardboard* and *held* provide cues towards a concrete noun interpretation.

(10) a. The "U" shaped cap (app. 1"x1"x1") will cover the frame and hide the old exterior putty or *concrete* **embedment**. Once caulked, the exterior will be neat, finished, and weather tight (0.05)

b. I developed a custom fit *cardboard* **fitment** that *held* the USB all nice and tight with the bonus of a business card next to it for company. (0.05)

It is also instructive to look at wrongly classified concrete nouns, as illustrated in (11-a). Although it is impossible to verify this statistically, the two highly event-related words *ongoing* and *fight* must have been influential in predicting an eventive interpretation, maybe also the word *see* (*during*, which may also be indicative of events, is a preposition and therefore not part of the content word window). The reverse case, an abstract non-eventive noun that is wrongly classified as eventive, is given in (11-b). Two of the four words in the context window of the target noun *debauchment*, are *yoga*, which denotes a kind of activity,

and thus is an eventive noun itself.

(11) a. Militias are *seen* in an **emplacement** during an *ongoing fight* against Syrian Regime Forces in Tal Mayyasat town of Aleppo, Syria on February 03, 2015 (0.90)

b. Gerson refers to most of the newer *yoga* classes as "**debauchment**". *Yoga* purists such as Gerson are calling for a return to teaching yoga in its original form, a program aimed at seeking self-enlightenment by training the mind. (0.88).

Let us finally turn to the ambiguous words. They tend to be classified as eventive. (12) gives two examples of such nouns for illustration. In these examples there is admittedly very little that could guide the listener's search for the right reading.

(12) a. Dont you see afk people trying to get the trinket as an **annoyment**? (0.22)

b. Client site just got a whole lot more awesome :) a way to go, work to be done, **convincements** still needed but a whole lot better in a flick (0.80)

If we compare the distribution of the ambiguous nouns with that of the abstract non-eventive nouns we see that they are very similar. The right maximum of the black dashed line (abstract non-eventive) is very close to the maximum of the blue dotted line (ambiguous). The means of these two sets of nouns are actually not significantly different from each other ($W = 527$, $p = 0.10$). This suggests that the semantic space of the ambiguous nouns is very similar to the semantic space of the abstract object nouns. And this space, as shown above, is similar to the space of eventive nouns.

6.3 Analysis 2: Lemma vectors vs. instance Vectors

Due to our focus on rare derivatives in this study, we had to use instance vectors, since no lemma vectors can be computed for unseen (and very rarely seen) words. Nevertheless, we would like to know how much noise the use of instance vectors introduces into the semantic class assignment task, as compared to the use of lemma vectors as is standard for frequent words. In other words, we would like to assess the difference in success when predicting the reading of a frequent word as against that of a word that is attested only once.

To make this assessment, we perform an experiment on the set of unambiguous, mid-frequent nouns from WordNet that we used before to train the eventive vs. non-eventive classifier. We use a technique called cross-validation (cf. Appendix A), dividing these nouns up into a training portion (with a 50-50 split between eventive and non-eventive nouns) and a test portion (with the same distribution). Thus, the baseline for classification in this setup is F=0.50.

In this setup, we can now either train a classifier directly on the lemma vectors of these nouns, or on the instance vectors for their individual occurrences. The results for both alternatives are shown in Figure 8, which gives the F-scores for our training nouns (on the y-axis). As in Analysis 2, we experimented with different subtypes of eventive nouns for

training. The green bars indicate the performance of the lemma vectors, the red and blue bars indicate the performance of the instance vectors for the two different window sizes. The baseline is indicated by the horizontal dark blue bar.

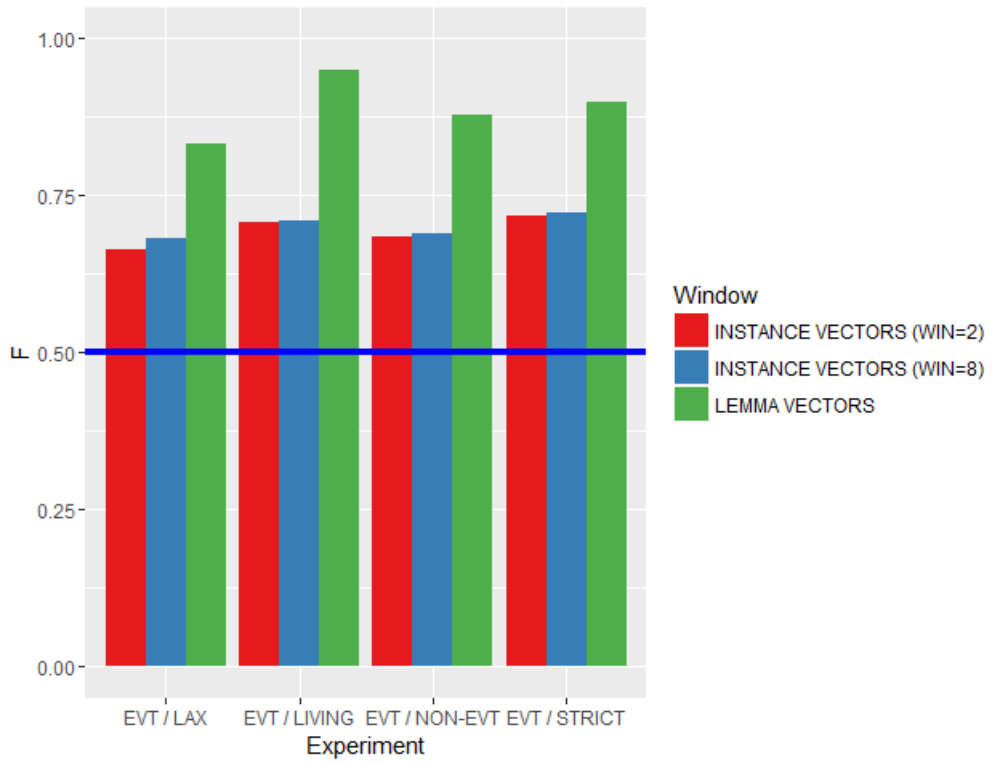

Figure 8: Distinguishing eventive and non-eventive nouns: Results for mid-frequency nouns, comparing lemma vector-based and instance vector-based models

We see that the lemma vectors yield good predictions (all four are around 0.8). The upper bound is a perfect score of 100% which the classifier can in principle achieve, but which it clearly misses.

The instance vectors work somewhat worse, with F-scores around 0.7. However, relatively speaking, this performance is rather comparable to the instance vectors in our main experiment (see 6.1). There, the improvement for the two most successful categories were 16 and 19 percent (2-word window) over the baseline of 0.44, whereas these categories show an improvement of 19 and 22 percent over the baseline for the training nouns. Taking into account that the training nouns are substantially more frequent than the rare derivatives we classified in our main experiment, and unambiguous to begin with, we can conclude that the instance vectors do a good job modeling the meaning of the rare derivatives.

The two different window sizes for the instance vectors result in very similar F-scores.

7 Discussion and conclusion

Our results show that it is possible to use Distributional Semantics to disambiguate the meaning of newly formed words. Although the predictions derived from the vector space still leave room for improvement, they are way above the baseline probabilities and thus show the usefulness of the context in disambiguation.

It also turned out that very small context windows are sufficient to find the intended interpretation in the majority of cases. Quite surprisingly a large window of eight words did not generally improve the predictions. A window with two content words on each side suffices to make predictions that are just as good, or even better than those derived on the basis of a larger window. Generalizing from the model to humans, this result may indicate that, as a general tendency, speakers or writers introduce and use new words in a way that helps listeners (or readers) to find the right interpretation without being forced to mobilize too many resources. The immediate linguistic context carries the semantics a long way.

The performance of the classifier differed for different subcategories of nouns. Non-eventive derivatives are hard to classify as such. Closer inspection revealed that the difficulty lies in the semantic similarity of abstract non-eventive nouns to eventive nouns, which led to many misclassifications in favor of eventive interpretations. The relationship of abstract nouns and eventive expressions is hardly discussed in the literature, which either focuses on different kinds of event nouns and event structure, or discusses the nature of abstract nouns vs. concrete nouns. An exception is Melloni (2011, 47), who mentions the problem in her discussion of Bierwisch's (1991/1992) treatment of German nominalizations in *-ung*. Bierwisch analyzes the nominalization *Ordnung* 'arrangement' in *Die Ordnung der Bücher war schwer wiederherzustellen* ('The arrangement of the books was hard to restore') as denoting a result state, i.e. as an event noun. Melloni points out that the noun can equally well be analyzed as denoting an abstract object. Incidentally, one and the same context can accomodate both types of noun, as in (13-a) and (13-b).

(13) a. The arrangement of the book is misleading. (result state)
b. The translation of the book is misleading. (abstract object)

In sum, this means that non-eventive abstract nouns and eventive nouns are not only similar in their semantic properties, they may also occur in the same contexts. Both facts make disambiguation of such nouns a hard task.

Perhaps not surprisingly, it also became clear that there are quite a few cases in which the interpretation of new words remains unclear, even if all resources are taken into account. In a sample of 406 tokens representing 55 types the manual annotation still had to live with 55, i.e. 14 percent, ambiguous forms. That means that there is a non-negligible proportion of forms that could not be classified convincingly by a human annotator, even if the wider textual context and world knowledge was taken into account.

References

Alpaydin, E. (2010). *Introduction to Machine Learning*. MIT Press, Cambridge (MA), 2nd edition.

Andreou, M. (2017). Stereotype negation in Frame Semantics. *Glossa*.

Asher, N. (2011). *Lexical Meaning in Context: A Web of Words*. Cambridge University Press, Cambridge.

Baayen, R. H. (1996). The effect of lexical specialization on the growth curve of the vocabulary. *Computational Linguistics*, 22:455–480.

Baayen, R. H. (2009). Corpus linguistics in morphology: morphological productivity. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics*, pages 900–919. Mouton de Gruyter, Berlin.

Baroni, M. and Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):1–49.

Bauer, L. (1983). *English word-formation*. Cambridge University Press, Cambridge.

Bauer, L. (2001). *Morphological productivity*. Cambridge University Press, Cambridge.

Bauer, L., Lieber, R., and Plag, I. (2013). *The Oxford reference guide to English morphology*. Oxford University Press, Oxford.

Bierwisch, M. (1990/1991). Event nominalizations: Proposals and problems. *Linguistica Hungarica*, 40(1-2):19–84.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc, Secaucus, NJ, USA.

Boleda, G., Padó, S., and Utt, J. (2012a). Regular polysemy: A distributional model. In *The First Joint Conference on Lexical and Computational Semantics*, pages 151–160, Montréal, Canada.

Boleda, G., Schulte, S., and Badia, T. (2012b). Modeling Regular Polysemy: A Study on the Semantic Classification of Catalan Adjectives. (July 2011).

Cotterell, R. and Schütze, H. (2017). Joint Semantic Synthesis and Morphological Analysis of the Derived Word. *Transactions of the Association of Computational Linguistics*.

Davies, M. (2008). The Corpus of Contemporary American English: 400+ million words, 1990-present.

Davies, M. (2013). Corpus of Global Web-Based English (GloWbE).

Diemer, S. (2011). Corpus linguistics with Google? In *Proceedings of ISLE 2 Boston 2008*.

Fellbaum, C., editor (1998). *WordNet: An electronic lexical database*. MIT Press.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, pages 1–32.

Fradin, B. (2011). Remarks on state denoting nominalizations. *Recherches Linguistiques de Vincennes*, 40:73–99.

Fradin, B. (2012a). Les nominalisations et la lecture 'moyen'. *Lexique*, 20:125–152.

Fradin, B. (September 7, 2012b). Sur la corrélation existant entre les suffixes -age et -ment et les distinctions sémantiques observables dans les nominalisations du français.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Hundt, M., Nesselhauf, N., and Biewer, C., editors (2006). *Corpus linguistics and the web*. Language and Computers. Rodopi, Amsterdam.

Izquierdo, R., Suárez, A., and Rigau, G. (2009). An empirical study on class-based word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 389–397, Athens, Greece.

Joanis, E., Stevenson, S., and James, D. (2006). A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367.

Kawaletz, L. (2018). *Predicting the semantics of English nominalizations. A frame-semantic analysis of the suffix -ment*. PhD dissertation, Heinrich-Heine-Universität Düsseldorf, Düsseldorf.

Kawaletz, L. and Plag, I. (2015). Predicting the semantics of English nominalizations: a frame-based analysis of -ment suffixation. In Bauer, L., Stekauer, P., and Kortvelyessy, L., editors, *Semantics of Complex Words*, pages 289–319. Springer, Dordrecht.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40.

Kisselew, M., Rimell, L., Palmer, A., and Padó, S. (2016). Predicting the Direction of Derivation in English conversion. In *Proceedings of the ACL SIGMORPHON workshop*, pages 93–98, Berlin, Germany.

Kunter, G. (2015). Coquery: A corpus query tool.

Lapesa, G., Evert, S., and Im Schulte Walde, S. (2014). Contrasting Syntagmatic and Paradigmatic Relations: Insights from Distributional Semantic Models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 160–170, Dublin, Ireland.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation.* The University of Chicago Press, Chicago.

Levy, O. and Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 302–308.

Lieber, R. (2016). *English nouns: The ecology of nominalization.* Cambridge University Press, Cambridge.

Lindsay, M. and Aronoff, M. (2013). Natural selection in self-organizing morphological systems. In Montermini, F., Boyé, G., and Tseng, J., editors, *Morphology in Toulouse*, Germany. Lincom Europa.

Löbner, S. (2013). *Understanding semantics.* Arnold, London, 2nd, revised edition edition.

Marchand, H. (1969). *Categories and types of Present-Day English word-formation.* C.H. Beck, München, second edition edition.

Marelli, Marco—Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, 122(3):485–515.

Melloni, C. (2011). *Event and result nominals: A morpho-semantic approach.* Peter Lang, Bern and New York.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41:1–69.

OED (2013). *The Oxford English dictionary online.* Oxford University Press, Oxford.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Plag, I. (1999). *Morphological productivity: structural constraints in English derivation.* Mouton de Gruyter, Berlin.

Plag, I. (2003). *Word-formation in English.* Cambridge University Press, Cambridge.

Plag, I., Andreou, M., and Kawaletz, L. (2017). A frame-semantic approach to polysemy in affixation. In Bonami, O., Boyé, G., Dal, G., Giraudo, H., and Namer, F., editors, *The lexeme in descriptive and theoretical morphology.* Language Science Press, Berlin.

Plag, I. and Balling, L. W. (2017). Derivational morphology: An integrated perspective. In Pirrelli, V., Dressler, W. U., and Plag, I., editors, *Word knowledge and word usage: A cross-disciplinary guide to the mental lexicon*. de Gruyter Mouton, Berlin.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge University Press, Cambridge.

Pustejovsky, J. (2011). Coercion in a general theory of argument selection. *Linguistics*, 49(6):1401–1431.

Rainer, F. (2014). Polysemy in derivation. In Lieber, R. and Štekauer, P., editors, *The Oxford handbook of derivational morphology*, Oxford Handbooks in linguistics, pages 338–353. Oxford University Press, Oxford.

Renouf, A., Kehoe, A., and Banerjee, J. (2006). WebCorp: an integrated system for web text search. In Hundt, M., Nesselhauf, N., and Biewer, C., editors, *Corpus linguistics and the web*, volume 59 of *Language and Computers*, pages 47–67. Rodopi, Amsterdam.

Roßdeutscher, A. (2010). German -ung-formation: An explanation of formation and interpretation in a root-based account. *Linguistische Berichte*, 17(117):101–132.

Roßdeutscher, A. and Kamp, H. (2010). Syntactic and semantic constraints on the formation and interpretation of-ung nouns. In Rathert, M. and Alexiadou, A., editors, *The semantics of nominalizations across languages and frameworks*, pages 169–214. de Gruyter Mouton, Berlin, New York.

Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, University of Stockolm.

Schmid, H.-J. (2011). *English Morphology and Word-Formation: Second revised and translated edition*. Erich Schmidt Verlag, Berlin.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 27(1):97–123.

Szymanek, B. (2013). Structural ambiguity in English word-formation. In Bondaruk, A. and Malicka-Kleparska, A., editors, *Ambiguity. Multifaceted structures in syntax, morphology and phonology*, pages 299–316. Wydawnictwo KUL, Lublin.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Uth, M. (2011). *Französische Ereignisnominalisierungen: Abstrakte Bedeutung und regelhafte Wortbildung.* De Gruyter, Berlin/New York.

Varvara, R. (2017). *Verbs as nouns: Empirical investigations on event-denoting nominalizations.* PhD dissertation, Università degli Studi di Trento, Trento.

Wittgenstein, L. and Schulte, J. (2001). *Philosophische Untersuchungen: Kritisch-genetische Edition.* Suhrkamp Verlag, Frankfurt am Main, 1. aufl. edition.

A Implementation details

In this appendix, we provide technical details characterizing the implementation of steps 0 to 3 outlined in Section 5.2 in the interest of reproducibility. The data are available as supplementary material at TO BE SPECIFIED UPON PUBLICATION

A.1 Step 0: Collecting training nouns and training sentences

Sampling training nouns We selected our training nouns based on the semantic field annotation in the WordNet 3.0 lexicographer files (Fellbaum, 1998). The semantic field annotation maps each noun in WordNet to 25 abstract semantic classes (called 'top nodes' because they form the roots of the WordNet noun hierarchy). Examples of such high level classes are: **artifact, communication, animal, plant, person**, etc.

To select the seed nouns for our experiments, we manually established the following mapping between our classes of interest and the WordNet semantic fields:

- EVENTIVE: state, feeling, process, phenomenon, event, act
- NON-EVENTIVE:
  - LAX OBJECT (LAX for short): communication, quantity, relation, social relation, possession
  - STRICT OBJECT (STRICT for short): object, substance, food, location, artefact, body
  - LIVING THING (LIVING for short): person, animal, plant

As a first step, we extracted five lists (EVENTIVE, NON-EVENTIVE, LAX, STRICT, LIVING) containing all nouns belonging to one or more of the relevant semantic fields. We did not strive for strictly monosemous training nouns: polysemy was allowed within the WordNet subtrees covered by each high-level class. For example, WordNet lists two readings for *fox*, 'animal' and 'sly person'. Since both are subsumed by LIVING THING, we included *fox* as training noun.

The five lists were further sampled according to frequency criteria. We extracted frequency information from a concatenation of three corpora: British National Corpus[8], WaCkypedia and ukWaC[9]. This combination (with 2.6 billions tokens) is used widely in compuational linguistics as a resource that covers various domains and genres (Baroni and Lenci, 2010). We lemmatized and part-of-speech tagged thecorpus. A first filtering step, we discarded all seeds with frequency below 100, which we consider a minimum prerequisite for reliable distributional lemma vectors. Taken together, the five lists amounted to 15k nouns. To further reduce their size, we concentrated on seed nouns whose frequency lay in the second and third quartile of the seeds' frequency distribution ($f > 298$ and $f < 3903$), following the assumption that mid-frequency lemmas strike a balance between reliability and generalization to rare derivatives. In the resulting lists, the smallest class was LAX, with 784 nouns. Consequently, we randomly selected 784 nouns for each class to obtain same-size training sets of 784 nouns for each of the five classes (EVENTIVE, NON-EVENTIVE, LAX, STRICT, LIVING).

Sampling training sentences For each noun in each of the five lists we extracted (from the same selection of corpora) all sentences containing the nouns with a sentence length between 6 and 150 word tokens[10]. We further randomly selected 10 sentences per noun. The sampling procedure outlined above resulted in five lists (EVENTIVE, NON-EVENTIVE, LAX, STRICT, LIVING) of $784 \cdot 10 = 7840$ training sentences each.

A.2 Step 1: Obtaining distributional representations for training nouns and training sentences

Using publicly available vectors is good practice in computational linguistics for reasons of reproducibility. Therefore, we did not build our own distributional lemma vectors. Instead, we used a set of 183.870 lemma vectors produced by Levy and Goldberg (2014) from the English Wikipedia by lemmatizing and lowercasing the data and applying the SkipGram algorithm with a bag-of-words context window of size 5 to obtain 300-dimensional lemma vectors. This is a state-of-the-art setup.

Based on these lemma vectors, we computed instance vectors for the 7840 training instances and 406 test instances by summing the lemma vectors for all context words within some window around the target noun instance that we wanted to disambiguate (cf. Section 5.1). In order to be able to assess how the size of the context window may influence disambiguation we computed the instance vectors once for a small context window (2 words to the left and to the right of the target noun) and once for a larger one (5 words to the

[8] `http://www.natcorp.ox.ac.uk`

[9] UkWaC and WaCkypedia are available from `http://wacky.sslmit.unibo.it/doku.php\?id=corpora`

[10] The lower bound is due to the fact that we want to be sure to find at least some content words as contexts of our seed nouns; the upper bound is due to the fact that sentences longer than 150 tokens, which come from Web corpora often contain lists of names or web links. (Sentences longer than 150 tokens constitute only the 0.03% of the sentences in BNC)

left and to the right).[11] When we did not find a lemma vector for a context word in the Levy & Goldberg set, we simply omitted the contribution of this context word.

A.3 Step 2: Learning to distinguish eventive and non-eventive nouns

As classifier, we used a support vector machine, which finds the hyperplane which best separates two classes in the feature space. Support vector machines are a frequent choice for classification settings in machine learning. They are in particular robust to overfitting, an important consideration when dealing with rare events. See, e.g., Bishop 2006; Alpaydin 2010 for the formalization. At the implementation level, we used the `svm` function from the `sklearn` Python package, version 0.18.1 (Pedregosa et al. 2011). We used a radial basis function kernel, with default settings. To obtain probabilities for classification decisions, we use Platt scaling (Platt, 1999), a technique that logistically transforms the distances from the decision boundary produced by the SVM into a conditional probability $P(class|instance)$.

A.4 Step 3: disambiguating -ment derivatives

Instance vectors for the sentences in the *-ment* dataset are computed in the way described in Section 5.1, using the contexts described in section 4. We used the same two context window sizes as before, and four different constellations of classes (EVENTIVE VS. NON-EVENTIVE, EVENTIVE VS. LAX, EVENTIVE VS. STRICT, EVENTIVE VS. LIVING). This amounts to the eight supervised classification experiments reported on in Section 6.1.

[11] Note that the choices of context windows for the lemma vectors and the instance vectors are independent.