Varvara, Rossella; Salvadori, Justine & Huyghe, Richard. 2024. Lexical ambiguity in contextualized word embeddings: A case study of nominalizations. *Lingue e Linguaggio* 1/2024, 141-182.

Lexical ambiguity in contextualized word embeddings:

A case study of nominalizations

Rossella Varvara, Justine Salvadori, Richard Huyghe
Université de Fribourg

Abstract

In this paper we investigate the extent to which contextualized word embeddings can encode lexical ambiguity. Specifically, we focus on nominalizations in French, which constitute an interesting case for the study of ambiguity because of their frequent polysemy and their relationship with polyfunctional morphological processes. Given a random sample of occurrences of 90 nouns, we compute for each word the pairwise cosine similarity (SelfSim) among their token embeddings extracted from the pre-trained model FlauBERT and we test it as a predictor of the degree of ambiguity of nominalizations. For the evaluation we make use of a manual annotation of lexical ambiguity, testing different annotation strategies: defining word senses with different semantic classifications and granularities; annotating lexemes in isolation or based on a sample of tokens. Our findings contribute to the understanding of (i) the lexical semantic component of contextual embeddings, enhancing their interpretability, (ii) aspects of lexical ambiguity related to derivational semantics and to the contextual variation of meaning.

1. Introduction[1]

Lexical ambiguity has been largely studied in linguistics, and nominalizations have been a fertile field of research on the topic, especially because they frequently show regular polysemy. Previous work on the ambiguity of nominalizations has mainly focused on describing some specific alternations of meanings or on compiling lists of polysemy patterns attested in a language, but further issues should be considered for a comprehensive study on the topic. For example, the importance and the diffusion of this phenomenon among nominalizations have not been frequently assessed, especially with a quantitative approach. It is indeed not clear what is the proportion of ambiguous nouns in the lexicon and if nominalizations do present a higher than average rate of ambiguity. Moreover, whether ambiguous words have a more dominant

[1] This work was supported by the Swiss National Science Foundation under grant 100012_188782 (“The semantics of deverbal nouns in French”). We thank the anonymous reviewers for their helpful comments and suggestions.

sense and how senses are distributed among occurrences is usually not investigated. Nominalizations being complex words, the contribution of the morphological process involved in the derivation should also be considered when analyzing their semantics, since some features may be morphologically conditioned. The lexical ambiguity of complex words may be indeed at least partially related to the semantic polyfunctionality of a word-formation process. We believe that, to reach a more complete understanding on these topics, it is necessary to consider a large representative sample of nominalizations and to analyze real data in a corpus-based approach.

In this paper, we take the first steps in this direction. We test whether recent computational large language models (namely, contextualized word embeddings) can provide an automatic assessment of ambiguity degrees based on their pre-training on corpus data, and how they compare with human annotation. Our aim is two-fold: first, we want to contribute to the development of an automatic quantitative instrument of analysis that could be used in theoretical research on lexical ambiguity, limiting time-consuming manual annotation; second, we make use of the specific case of nominalizations to challenge these NLP systems, deepening our understanding of their inner working, their interpretability and evaluation. Moreover, we provide insights into the semantic characterization of derived nouns, discussing different strategies to analyze their complex semantics.

Research in natural language processing has largely investigated lexical ambiguity, trying most notably to solve the problem of word sense disambiguation, a task necessary for the understanding and modeling of semantics in automatic systems. Recently, a new generation of computational language models, called contextualized word embeddings, have been developed and settled as the state-of-the-art model in many NLP tasks. In previous work (e.g., Garí Soler & Apidianaki 2021; Haber & Poesio 2021), these models have shown to be promising in the automatic assessment of lexical ambiguity. They provide a vector representation for each occurrence of a word, contrary to previous word embeddings that only offered a unique representation for the whole word, conflating multiple senses.

We put to the test a French version of the well-known BERT model (Devlin et al. 2018), namely FlauBERT (Le et al. 2020), testing if its pre-trained representations are able to predict the degree of ambiguity of 90 French nominalizations. Specifically, we hypothesize that monosemous nouns will present corpus tokens with more similar embeddings, whereas ambiguous nouns will have more sparse distributional representations, since their occurrences will be related to different senses. We compute the average pairwise similarity among token embeddings of a word (a measure previously called SelfSim) and we expect this measure to be inversely correlated with ambiguity. The evaluation made by the computational system is compared with a manual annotation of degrees of ambiguity obtained in two ways. In the first case, we annotate each noun at the lexical level, trying to define an exhaustive list of its possible semantic types, relying on lexical resources and dictionaries. In the second case, we annotate a sample of 50 occurrences of each noun randomly extracted from a corpus.[2] In both cases, the degree of ambiguity corresponds to the number of different types annotated, based on the same semantic classification. These two annotation types allow us (i) to better evaluate the behavior of the token embeddings, and (ii) to delve into methodological questions for the annotation of lexical ambiguity, highlighting differences between the two methodologies.

The article is organized as follows. In Section 2, we provide an overview of previous work on the ambiguity of nominalizations and an introduction to the computational model used in our

[2] In what follows, we will refer to the first case as lexical annotation (or lexically annotated gold standard) and contextual annotation (or contextually annotated gold standard) for the second one.

study. In Section 3, we present the sample of nominalizations under scrutiny and the annotation scheme. We describe the results of the annotation in Section 4, with particular attention to the differences found between the two types of annotation (lexical and contextual), and between the different nominalizing suffixes (in terms of semantic polyfunctionality and ambiguity of derivatives). In Section 5, we present the main experiment of the study, in which we consider the distributional measure SelfSim and test its ability to predict degrees of lexical ambiguity. To further analyze the results of this experiment, in Section 6 we explore how contextualized embeddings can encode different semantic types, training a lightGBM classifier to distinguish each type from the others. Lastly, in Section 7, we discuss the overall results and their implications.

2. Background

Lexical ambiguity can be defined as the property of a word form to be associated with more than one meaning. It covers both homonymy, i.e., when the different meanings are not related (e.g., bank 'institution'/ 'area of land'), and polysemy, i.e., when the different meanings are related, mostly through metaphor (e.g., *heart* 'organ'/ 'central part') or metonymy (e.g., *crown* 'decoration'/ 'person who rules'). It is well known that many if not most nouns derived from verbs are ambiguous (e.g., Rainer 1996, 2014; Melloni 2007; Lieber 2018), which has prompted numerous studies over the last decades. Regardless of differences in theoretical frameworks, two closely related lines of research stand out.

On the one hand, recurrent associations of senses in ambiguous derivatives are investigated. They are essentially cases of regular polysemy as defined by Apresjan (1974: 16), according to whom "polysemy of the word A with the meanings ai and aj is called regular if, in the given language, there exists at least one other word B with the meanings bi and bj, which are semantically distinguished from each other in exactly the same way as ai and aj and if ai and bi, aj and bj are nonsynonymous". The most discussed semantic alternation in nominalization is of the EVENT/REFERENTIAL type,[3] where one meaning corresponds to an event and the other can refer to any abstract or concrete entity. In this case, senses are generally considered to be metonymically related. The entity designated by the referential meaning can often be thought as a potential participant (e.g., agent, instrument, result) in the action associated with the event meaning, which is a transposition of the meaning of the base verb. While some studies address this alternation as a whole (e.g., Melloni 2011; Barque *et al.* 2014), most work focuses on associations comprising a subtype of referential meaning (see e.g., Bisetto & Melloni 2007 for EVENT/RESULT; Jacquey 2006 for EVENT/ARTEFACT; Ferret & Villoing 2015 for EVENT/INSTRUMENT; Fradin 2012 for EVENT/MEANS). Alternations involving only referential meanings (see e.g., Booij 1986 for AGENT/INSTRUMENT) or only eventualities (see e.g., Montermini 2015 for EVENT/PROPERTY) are explored as well, although more marginally. Incidentally, it is worth noting that these studies are also complemented by semasiological efforts that address the ambiguity of individual nominalizations or that of derivatives formed with specific suffixes. Examining *-er* nominalizations, Panther & Thornburg (2002), for instance, provide a case study of the polysemy of *sleeper*, whereas Kawaletz (2021) reports that attestations of neological nouns derived with *-ment* are generally ambiguous between different semantic types.

[3] The label RESULT (e.g., Grimshaw 1990; Alexiadou 2019) is also frequently used instead of REFERENTIAL (e.g., Melloni 2011; Lieber 2016). To avoid confusion in terminology (see Jezek 2008), we reserve it for the denotation of actual results or by-products of actions.

On the other hand, links between morphological processes and ambiguity are also frequently debated. As in the simplex lexicon, semantic figures such as metaphor and metonymy may apply to existing derivatives (Bauer 2017). This becomes particularly apparent when the metonymic or metaphorical meaning has no connection with the morphological base. In French, for example, only one meaning of the derivative *planteur* 'planter' can be traced back to the base verb *planter* 'plant', the other meaning 'Planter's punch' being a metonymic extension of the first one. Other cases are not necessarily so clear cut. The morphologically complex nature of deverbal nouns additionally raises questions about the role of bases, suffixes and their interaction in the generation of ambiguity in nominalization (e.g., Ferret & Villoing 2015; Kawaletz & Plag 2015). Of particular interest to this study is the relationship between affix polyfunctionality – or ambiguity at the affix level – and lexical ambiguity in derivatives, which cannot be reduced to one another. That a given nominalizing suffix can serve more than one semantic function (e.g., forming eventive, agentive or instrumental nouns) does not automatically entail that all its functions are instantiated by all its derivatives. Some affixes may have a stronger propensity to form ambiguous words than others, independently of the number of functions they serve.

It appears that describing ambiguity in nominalization is theoretically challenging and raises several methodological issues. First is that different forms of ambiguity should potentially be distinguished in classifications. Like the different meanings of ambiguous simplex nouns, those of ambiguous derivatives are generally mutually exclusive, i.e., they refer to distinct situations or entities, be they related or not. The noun *building*, for instance, denotes either an action or a concrete result but never both at the same time. However, as argued by Jacquey (2006) and Melloni (2011), senses of derivatives can also represent distinct ontological facets of a single referent and thus instantiate what has been called "inherent polysemy" in the literature (e.g., Pustejovsky 1995; Cruse 1995; Godard & Jayez 1993; Kleiber 1999; Asher 2011; Murphy 2021). The most distinctive feature of facets is that, unlike senses of standard polysemes, they typically accept co-predication, i.e., they can be used jointly in context without creating a zeugma effect. In (1), for example, both facets of the French noun *attestation* 'certificate', derived from *attester* 'confirm', are present. The verb *récupérer* 'pick up' selects the one associated with the material object, whereas the verb *indiquer* 'say' calls for the one related to the informational content.

1) *Je suis arrivée ici à 8 heures, ce matin, pour récupérer l'attestation qui indique que je peux reprendre mon poste.* (web)
   'I arrived here at 8 a.m. this morning to pick up the *certificate* that says I can return to my job.'

Another issue pertains to levels of semantic analysis. Ideally, information about the nature of the referent and information about the relation of the derivative with its base should not be conflated, as is the case in many existing classifications. Even though some privileged relations can be observed (e.g., agents performing actions denoted by verbs are generally animate entities), they are not systematic. A derivative that expresses a result with regard to its base may denote an artefact (e.g., *draw* → *drawing*), a state (e.g., *annoy* → *annoyance*) or an animate entity (e.g., *create* → *creature*) ontologically speaking. Conversely, a derivative that denotes an artefact may express various relations with respect to its base: a result (e.g., *draw* → *drawing*), an instrument (e.g., *trim* → *trimmer*) or a location (e.g., *dine* → *diner*), for instance. Finally, methods of semantic analysis should be made as explicit as possible to guarantee the reliability of the annotation scheme. This includes providing detailed definitions or tests in addition to examples for the labels used, as well as calculating agreement scores among annotators. As we shall see in Section 3.2, we have tried to take all of these points into consideration when conducting this study.

2.2 Distributional models, contextualized embeddings and lexical ambiguity

In recent years, distributional models of meaning (also called "vector space models" or "word embeddings") have become the standard tool for representing semantics in natural language processing. All of these models rely on the Distributional Hypothesis (Harris 1954), according to which "difference of meaning correlates with difference of distribution". They make use of the distributional information extracted from very large corpora to generate vectors that approximate the meaning of words. Similarity between pairs of vectors indicates that the two words are similar in meaning, since they occur with the same set of context words. The similarity between vectors is frequently computed by means of the cosine similarity measure, but other mathematical operations are possible as well. Since their first appearance, they have achieved really good results in many NLP tasks. They have been applied in various fields of theoretical linguistics, including semantics and morphology, to investigate basic research questions with a quantitative and empirical approach (see Boleda 2020 for an overview, and Wauquier 2022).

In only a few decades, distributional models have evolved rapidly, and their evolution is continuously underway. The first models (known as "count- based models", Baroni *et al.* 2014) approximated the meaning of a word by a vector of co-occurrences with other words in a corpus. In the last decade, neural models (or "predictive models") have been introduced and have rapidly become the most widely used, starting from the success of the work by Mikolov and colleagues (2013a; 2013b).[4] These models, instead of counting word co- occurrences, are trained on large corpora to predict a word given its context (or vice versa, to predict the context of a given word). The implicit representation built by the model to solve this language prediction task corresponds to a vector (an embedding), used as a representation for the target word. Popular neural models include Word2Vec (Mikolov *et al.* 2013a), GloVe (Pennington *et al.* 2014), or FastText (Bojanowski *et al.* 2017), just to name a few.

Despite this difference, both count and predictive models produce distributional representations for word types, thus conflating the different senses of a word into a unique representation. These static models are not able to account *per se* for the variety of senses of ambiguous words. Both monosemous and polysemous nouns have indeed only one representation. This problem pushed the NLP community to investigate methods to account for polysemy, for example by deriving an embedding for each sense of a word listed in a lexicographic resource (e.g., Schütze 1998; Pantel & Lin 2002; Iacobacci *et al.* 2015; Pilehvar & Collier 2016). Some approaches used semantic composition, combining together all the vectors of words in a sentence to represent the meaning of a target word in that context (Landauer & Dumais 1997; Mitchell & Lapata 2008). Other solutions have also been developed (e.g., multi-prototype embeddings, Pelevina *et al.* 2016; see Camacho-Collados & Pilehvar 2018, for an overview of different word sense modeling approaches).

More recently, contextualized word embeddings, a new family of neural language models (e.g., BERT, Devlin *et al.* 2018; ELMo, Peters *et al.* 2018), have been conceived to generate a representation for each instance of a word, i.e., for word tokens instead of word types. BERT's family of models are multi- layer bidirectional transformer encoder (Vaswani *et al.* 2017) that are trained on two tasks: a Masked Language Modeling task, that consists in predicting a random word that has been masked in the sentence; a Next Sentence Prediction task, in which it has to predict if two sentences are actually adjacent sentences in the training corpus or if they

[4] See Bengio *et al.* (2000) or Collobert & Weston (2008) for earlier accounts of neural models.

are unrelated. The model obtained from these pre- training tasks can be used to derive representations of tokens in new contexts or can be fine-tuned for further NLP tasks using labeled data. The model architecture is composed of multiple layers, i.e., transformer blocks, in which the output of the previous layer is used as input for the next one. BERT was released with two main configurations: BERT base, which comprises 12 layers (12 attention heads, 110 million parameters); BERT large, with 24 layers (16 attention heads, 340 million parameters). It follows that in a 12-layer BERT model a token will have 12 representations, one for each layer.

Since they produce token vectors, these models seem well suited to approach lexical ambiguity and to solve the task of word sense disambiguation (WSD). Different studies have tested them on such a task (see e.g., Loureiro *et al.* 2021), whereas others have tried to probe their knowledge about lexical ambiguity (see Apidianaki 2023 for an overview). For example, Reif *et al.* (2019) and Wiedemann *et al.* (2019) observe that representations of word tokens provided by BERT are organized in the semantic space across senses, with tokens related to the same sense closer to each other. Moreover, they note that BERT performs well in the WSD task, achieving state-of-the-art results. Haber & Poesio (2021) show that the contextualized representations capture differences between homonymous and polysemous words, with token embeddings of polysemous words being closer to each other in the semantic space than those of homonymous words. On the other hand, Garí Soler & Apidianaki (2021) observe that BERT is sensitive to the difference between monosemous and polysemous words, with token embeddings of monosemous words being more similar to each other than those of polysemous words. This difference is found even if the occurrences of polysemous words are sampled to represent only one sense, thus suggesting that BERT encodes information about the polysemous nature of a word regardless of the distribution of senses among the sentences used to extract the contextualized representations. It appears that BERT relies on the knowledge acquired from the pre-training corpus to represent new occurrences.

A large body of research investigates the knowledge the model encodes in the token representations and the differences among layers, frequently reaching conflicting conclusions (see Rogers *et al.* 2021 for an overview of the first interpretability studies). It has been observed that the lower layers encode mostly information about linear word order (Liu *et al.* 2019), whereas the middle layers more prominently encode syntactic information (Hewitt & Manning 2019; Goldberg 2019; Jawahar *et al.* 2019). High-level semantic features seem to appear in the higher layers (Tenney *et al.* 2019), but another study has shown that lexical knowledge is spread throughout multiple layers, especially in the lower ones (Vulic´ *et al.* 2020). Ethayarajh (2019) observes that the higher layers produce more context-specific representations, since embeddings for specific words are more similar to each other in the other layers. According to Liu *et al.* (2019), the final layers are the most task-specific – and it follows that in the pre-trained model these are specific to the MLM task – whereas the middle layers are overall better and more transferable across tasks. Even if a unique shared interpretation has not been reached yet, the differences among layers' representations are largely investigated. In the present study, we will consider this issue while testing the ability of BERT to evaluate lexical ambiguity degrees. We make use of a French version of the BERT architecture, namely FlauBERT (Le *et al.* 2020), since BERT has been shown to perform better than other contextual models in distinguishing monosemous and polysemous words (Garí Soler & Apidianaki 2021).

3. ANNOTATING LEXICAL AMBIGUITY THROUGH SEMANTIC TYPES

3.1 Dataset

Our case study employs data taken from the FRCOW16A corpus (Schäfer & Bildhauer 2012; Schäfer 2015), both for lexical sampling and for the random selection of corpus occurrences. We consider a sample of 90 French deverbal nouns formed with 6 suffixes, viz. *-ade*, *-aire*, *-is*, *-ment*, *-oir* and *-ure*, as illustrated in (2).

2)

a) *dégoulinade* 'drip', *fusillade* 'shooting', *glissade* 'slip'
b) *bénéficiaire* 'beneficiary', *commentaire* 'comment', *dispensaire* 'dispensary'
c) *croquis* 'sketch', *logis* 'dwelling', *roulis* 'rolling'
d) *gouvernement* 'government', *isolement* 'isolation', *miaulement* 'meow'
e) *mouchoir* 'handkerchief', *promenoir* 'walkway', *tiroir* 'drawer'
f) *moulure* 'moulding', *reliure* 'binding', *souillure* 'stain'

These 6 suffixes were selected from an initial list of 46 suffixes used to derive nouns from verbs in French (see Dubois 1962; Thiele 1987; Apothéloz 2002 among others). The 46 suffixes were divided into 3 groups based on the number of semantic functions they serve, as described by Salvadori & Huyghe (2023), and 2 suffixes were randomly selected in each group. While *-ment* and *-ure* are among the most polyfunctional nominalizing suffixes, *-aire* and *-oir* have a limited number of semantic functions, and *-is* and *-ade* are in the middle range. For each suffix, 15 nouns were randomly extracted from the frequency list of the corpus according to two criteria: (i) lemmas should have a frequency higher than 50, so that a sufficient number of occurrences could be used to investigate lexical ambiguity; (ii) for each suffix, derivatives were equally selected among three frequency ranges, in order to consider similar variation of frequency across suffixes. Five nouns were randomly selected among those with a token frequency up to the general median value (50-223), 5 nouns with a token frequency from the median value to the third quartile (224-3,799), and 5 nouns with a token frequency higher than the third quartile (3,800-3,966,941). For the contextual annotation, we randomly extracted 50 occurrences of each of the 90 lemmas, which brought the total number of tokens to be examined to 4,500.

3.2 The annotation scheme

In this study, we do not approximate ambiguity degrees through the number of specific senses assigned to each word, listed in a lexicographic fashion. Instead, all analyses are based on the number of different semantic types instantiated by each word. Semantic types allow for generalization across both derivational and regular polysemy patterns and for comparison between words that express the same type of meaning. Regularities among derivatives with the same affix can be explored, and variable semantic granularity can be tested to determine the level of information that best matches distributional data. Two separate classifications – one ontological and one relational (see Section 2.1) – can be used to analyze the semantic properties

of nominalizations. These two classifications can be concatenated to form what we will refer to as "combined types" in the remainder of the article.[5]

The ontological classification presented in Table 1 is employed to describe the nature of the referents denoted by nouns (e.g., artefact, event, animate entity). It includes 14 simple types as well as 7 complex types to account for inherent polysemy (see Section 2.1). For example, the noun *brochure* 'brochure' is assigned the complex type ARTEFACT*COGNITIVE because it denotes a manufactured object with an informational content. Each simple and complex type can in turn be added a COLLECTIVE label in case the noun examined has a plural reference when in the singular form (e.g., Flaux 1999; Lammert 2006; De Vries 2021). The French noun *lotissement* 'housing estate', for instance, does not denote a single house but a group of houses. Accordingly, it is assigned the ontological type ARTEFACT-COLLECTIVE rather than just ARTEFACT. The selection of the appropriate ontological type for a given noun sense is based on linguistic tests taken from studies dedicated to nominal semantics in French (Godard & Jayez 1993; Flaux & Van de Velde 2000; Huyghe 2015; Haas *et al.* 2023).

The relational classification is used to describe the semantic relation of derivatives with their base verb. The 18 types are presented in Table 2. With the exception of the TRANSPOSITION type which serves to indicate that the noun denotes the same eventuality as its base (e.g., *rouler* 'roll' → *roulade 'roll'*), they essentially correspond to semantic roles that can be assigned by verbs to their arguments. The classification we use is adapted from LIRICS (Petukhova & Bunt 2008) and VerbNet (Schuler 2005), and semantic roles are identified based on explicit definitions. The noun *grattoir* 'scraper', for instance, is assigned the relational type INSTRUMENT because it refers to an entity that is manipulated in order to perform the action denoted by *gratter* 'scrape'. An additional label FIGURATIVE can also be used in case a given sense of an ambiguous noun has no semantic relation with the morphological base but constitutes a metaphorical or metonymic extension of another meaning of the same noun. In such a case, the figurative meaning is assigned the same relational type as the meaning from which it is semantically derived, complemented with the additional label FIGURATIVE. For example, the meaning of *bavure* 'blunder' that is not directly derived from the verb *baver* 'drip' but is a metaphor of *bavure* 'drip' is annotated as figurative in the following schema: *baver* 'drip' → *bavure1* 'drip' [RESULT] → *bavure2* 'blunder' [RESULT-FIGURATIVE].

[5] The annotation guide used for the semantic analysis is available at https://github.com/ semantics-deverbal-nouns/annotation-guide.

ONTOLOGICAL TYPES		
Animate	Institution	Artefact*cognitive
Artefact	Natural	Artefact*institution
Cognitive	Phenomenon	Cognitive*event
Disease	Property	Event*financial
Domain	Quantity	Event*natural
Event	State	Event*phenomenon
Financial	Time	Event*state

TABLE 1: ONTOLOGICAL TYPES.

RELATIONAL TYPES		
Agent	Instrument	Result
Beneficiary	Location	Source
Cause	Manner	Stimulus
Destination	Path	Theme
Experiencer	Patient	Topic
Extent	Pivot	Transposition

TABLE 2: RELATIONAL TYPES.

Ontological and relational classifications are complementary and can be combined to provide a full description of the semantic properties of nominalizations. Combined semantic types provide detailed information about deverbal nouns, and one can ask whether such a fine-grained description can be captured by BERT embeddings. To anticipate this possible issue and to evaluate the accuracy of BERT with respect to various degrees of semantic precision, we defined 3 different levels of semantic granularity for both ontological and relational types: (i) fine types, corresponding to the original annotation scheme outlined above; (ii) medium types, merging complex types as well as collective and figurative meanings; (iii) coarse types, reduced to 3 ontological and 4 relational general classes. These three levels can be extended to combined types through the concatenation of ontological and relational types of identical granularity. The correspondence of types and labels across the different granularities is reported in Appendix (Table 9 for ontological types, Table 10 for relational ones).

The annotation scheme described above was applied to the sample of nominalizations presented in Section 3.1. Two annotation campaigns were conducted. The first one consisted in determining the different semantic types of the 90 selected nouns, based on the variety of meanings mentioned in lexicographic resources (*Le Petit Robert* and *Trésor de la Langue Française informatisé*). The different word senses provided by dictionaries for each noun were reevaluated and sometimes (un)grouped to be classified into ontological and relational types. The goal of this operation was to list as exhaustively as possible the different semantic types associated with each noun, regardless of their use frequency. This methodology was successfully tested in a previous study (Huyghe *et al.* 2023) involving two annotators who co-

authored the present paper. Inter-annotator agreement was substantial for both ontological types (Cohen's κ = 0.77) and relational types (Cohen's κ = 0.78). The second annotation campaign was performed by 3 annotators on the 4,500 selected corpus tokens. It was primarily aimed at assessing the contextual distribution of the semantic types associated with each selected noun. For example, a given noun that has three semantic types (e.g., EVENT-TRANSPOSITION, ARTEFACT-RESULT, PHENOMENON-CAUSE) may actually realize only two of them frequently (e.g., EVENT-TRANSPOSITION, ARTEFACT-RESULT). In this second campaign, a double-blind annotation was performed on two samples of 300 corpus tokens, showing substantial inter-annotator agreement, with a final Cohen's κ of 0.82 and 0.72 for ontological and relational types, respectively. The rest of the tokens were analyzed by the three annotators separately, although the most problematic cases were discussed and annotated collectively.

4. OBSERVED AMBIGUITY OF NOMINALIZATIONS

In this section, we present the results of the semantic annotation of the nominalizations selected in the study, based on lexical and contextual data. We describe the role of the semantic classifications used for the definition of senses and for the assessment of lexical ambiguity, the differences between the lexical and the contextual annotations, as well as differences among suffixes. The annotation described will be used in Section 5 as a gold standard for the assessment of ambiguity in contextualized embeddings.

4.1 Degrees of ambiguity

Based on the annotation described in the previous section, we obtain for each noun in our sample a number of semantic types that reflects its degree of ambiguity. As already mentioned, we rely on different annotation strategies (lexical and contextual), as well as on 3 semantic classifications (ontological, relational, combined) and 3 granularity levels (fine, medium, coarse). The combination of these variables results in 18 values of ambiguity for each noun.
In Table 3, we report ambiguity rates for the 90 nouns under scrutiny, considering the different typologies of annotation. The average ambiguity rate per noun is lower for the lexical annotation (Mean = 1.44, SD = 0.66) than for the contextual annotation (Mean = 1.91, SD = 1.03). As expected, coarse-grained typologies are associated with the lowest rates of ambiguity, since they merge together in a single class semantic types that are distinguished in the more fine-grained typologies. Relational types exhibit less variation not only between lexical and contextual annotations, but also across semantic granularities: the difference in ambiguity rate among granularities of relational types is limited to 0.07 for the lexical annotation and 0.19 for the contextual one (vs. 0.22 and 0.90 for ontological types, and 0.18 and 0.76 for combined types). This difference shows that ambiguity is defined at a more general semantic level with relational than with ontological types, i.e., the number of meanings is more stable over semantic granularities for relational than for ontological properties. By comparison, differences in meaning at the ontological level (and consequently, for combined types) are frequently due to subtle semantic differences that fade with coarse labels. The contrast between semantic classifications is amplified in the contextual vs. lexical annotation, since the difference of ambiguity rate between the two annotations is higher for ontological and combined fine types (0.89 and 0.96, respectively) than for relational fine types (0.27). This indicates that more fine-

grained labels are used for ontological types than for relational types in the corpus annotation, where more subtle semantic components are identified in contextual uses. This result was expected because, unlike the type of referent denoted by a noun, the relationship between a derivative and its base verb does not depend directly on linguistic context. Ambiguity degrees are thus affected by the type of semantic description involved and by differences between lexical and contextual annotations.

SEMANTIC TYPOLOGY	LEX ANNOTATION		CONTEXT ANNOTATION		RATE DIFF
	NO. OF MEANINGS	RATE	NO. OF MEANINGS	RATE	
ONTOLOGICAL FINE	136	1.51	216	2.40	0.89
ONTOLOGICAL MEDIUM	132	1.47	182	2.02	0.56
ONTOLOGICAL COARSE	116	1.29	135	1.50	0.21
RELATIONAL FINE	130	1.44	154	1.71	0.27
RELATIONAL MEDIUM	125	1.39	147	1.63	0.24
RELATIONAL COARSE	123	1.37	137	1.52	0.16
COMBINED FINE	141	1.57	228	2.53	0.97
COMBINED MEDIUM	135	1.50	192	2.13	0.63
COMBINED COARSE	129	1.39	159	1.77	0.38
AVERAGE	129	1.44	172	1.91	0.48

TABLE 3: TOTAL NUMBER OF WORD MEANINGS OBSERVED FOR 90 NOMINALIZATIONS, WITH AVERAGE AMBIGUITY RATES PER WORD AND RATE DIFFERENCES BETWEEN ANNOTATION TYPES, FOR VARIOUS SEMANTIC TYPOLOGIES.

As shown in Table 4, the number of meanings per noun in the lexically annotated dataset ranges from 1 to 4 (except for ontological and relational coarse types, which range from 1 to 3). However, nouns with up to 8 different meanings can be found in the contextually annotated dataset. The highest number of nouns with more than 3 meanings can be observed in the contextual annotation with combined fine-grained semantic types, and amounts to 19 out of 90 nouns. By contrast, there are more monosemous nouns (i.e., nouns with only one semantic type) than ambiguous nouns in the lexical annotation, since they represent on average 59% of the sample. The number of monosemous nouns is lower in the contextual annotation (41% on average). The fact remains that nouns with only one semantic type are more frequent than all other nouns in the sample, whatever the typology.

In addition to observations on the variation and distribution of ambiguity degrees in the annotated datasets, we investigated whether the ambiguity of a noun was correlated with its frequency in the corpus, with the hypothesis that more frequent words have a higher degree of ambiguity. The results of a Spearman correlation test between frequency and number of meanings assigned to a noun were not significant ($p > 0.05$), regardless of the semantic typology or annotation type involved – the only exception being combined fine types in the lexical annotation, for which a very weak correlation could be observed ($\rho = 0.19$, $S = 97957$, $p < 0.05$). Quite unexpectedly, the level of lexical ambiguity does not appear to be dependent on use frequency for the nominalizations we analyzed.

SEMANTIC TYPOLOGY	ANNOTATION	NUMBER OF MEANINGS							
		1	2	3	4	5	6	7	8
COMBINED FINE	LEXICAL	52	27	9	2				
	CONTEXTUAL	32	20	19	6	7	3	2	1
COMBINED MEDIUM	LEXICAL	57	22	10	1				
	CONTEXTUAL	37	25	15	8	3	1	1	
COMBINED COARSE	LEXICAL	61	24	4	1				
	CONTEXTUAL	43	28	17	1	1			
ONTOLOGICAL FINE	LEXICAL	55	25	9	1				
	CONTEXTUAL	35	19	18	8	5	4		1
ONTOLOGICAL MEDIUM	LEXICAL	59	21	9	1				
	CONTEXTUAL	38	27	15	7	1	2		
ONTOLOGICAL COARSE	LEXICAL	65	24	1					
	CONTEXTUAL	52	31	7					
RELATIONAL FINE	LEXICAL	59	24	5	2				
	CONTEXTUAL	41	36	11	2				
RELATIONAL MEDIUM	LEXICAL	61	24	4	1				
	CONTEXTUAL	43	38	8	1				
RELATIONAL COARSE	LEXICAL	61	25	4					
	CONTEXTUAL	47	39	4					

TABLE 4: NUMBER OF NOUNS PER NUMBER OF MEANINGS IN LEXICAL AND CONTEXTUAL ANNOTATIONS, BASED ON VARIOUS SEMANTIC TYPOLOGIES.

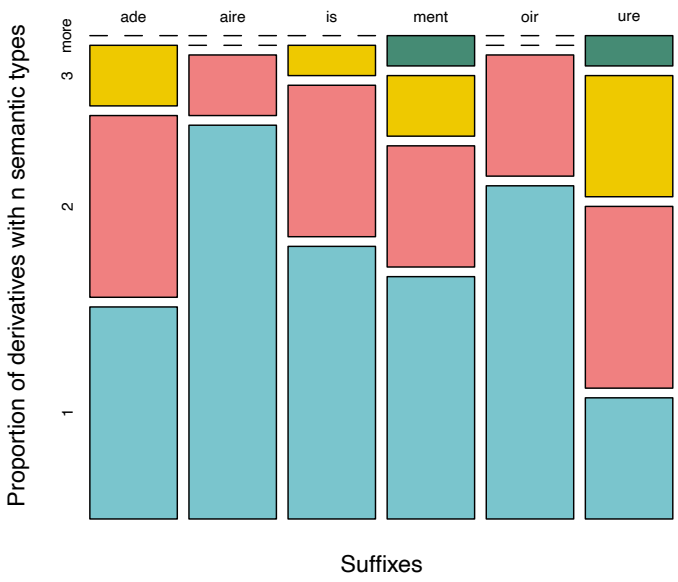
4.2 The role of suffixes in lexical ambiguity

As mentioned in Section 2, word-formation processes can have an influence on the ambiguity of complex words. Affixes in particular can differ in their ability to form ambiguous words, which has only been marginally investigated in previous research. Our annotated data can provide information about the differences between nominalizing suffixes with regard to the derivation of ambiguous nouns.

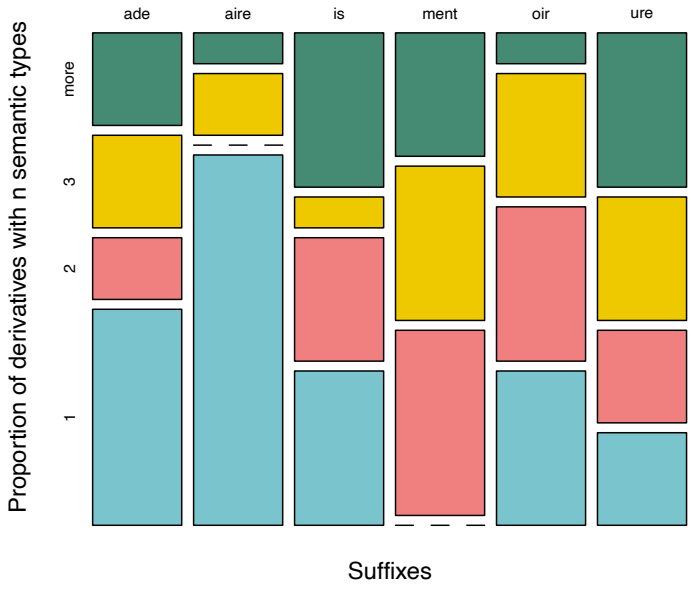
Overall, the proportion of monosemous and ambiguous nominalizations varies according to the different suffixes considered. Figure 1 presents the distribution of derivatives per suffix and per level of ambiguity, considering finegrained combined types. Some suffixes (e.g., *-ure*) tend to form more ambiguous nouns than others (e.g., *-ade*), and despite their polyfunctionality, suffixes such as *-aire* are used to derive mostly monosemous nouns. Further, some important differences can be noted between the lexical and contextual annotations. While more semantic flexibility is generally observed in contextual uses, this property is more pronounced for some suffixes than for others. For instance, the number of meanings for nominalizations in *-ment* and *-oir* increases significantly in contextual uses, whereas *-aire* is relatively stable with respect to the distribution of monosemous and ambiguous derivatives across both annotations. The higher numbers of ambiguous words observed in the contextual annotation are to some extent related to complex types. This assumption is supported by the fact that the differences between lexical and contextual annotations are smaller when considering medium-grained semantic types, in which complex types are conflated (see Tables 11 and 12 in Appendix). In the case of *-ment* for instance, some nominalizations are lexically annotated as having a complex meaning

composed of two semantic facets, but these nouns are often disambiguated when used in discourse, with the contextual selection of one facet or the other. Such is the case of the derivative *pleurnichement* 'whining', for example. While it has been assigned the complex type COGNITIVE*EVENT at the lexical level, some of its tokens have been considered to instantiate a specific facet at the contextual level: cognitive in (3), event in (4). It remains that both facets are sometimes indistinguishable, as in (5).

3) *King Kong Théorie ressemble à un livre qu'aurait pu écrire une féministe extrémiste des années 60: rempli de clichés, de raccourcis, de pleurnichements d'une nana maltraitée par LES hommes, sans remise aucune du rôle de la femme dans l'état actuel de la société ...*
   'King Kong Theory looks like a book that could have been written by an extremist feminist of the 60's: full of clichés, shortcuts, *whining* of a girl mistreated by men, without any consideration of the role of women in the current state of society.'
4) *Pour le dire autrement: les pleurnichements, les lamentations et les frémissements ne sont pas l'ultime mode d'action sur le monde.*
   'To put it another way: *whining*, lamenting and quivering are not the ultimate mode of action on the world.'
5) *Début mars 2013, malgré ses pleurnichements d'innocence, jurant n'avoir commis aucune erreur, Steven Vanackere présente sa démission sous le prétexte de subir trop "d'insinuations persistantes et injustifiées" (sic)!*
   'In early March 2013, despite his *whining* of innocence, swearing that he had made no mistakes, Steven Vanackere submitted his resignation under the pretext of being subjected to too many "persistent and unjustified insinuations" (sic)!'



(a) LEXICALLY ANNOTATED



(b) CONTEXTUALLY ANNOTATED

FIGURE 1: DISTRIBUTION OF DERIVATIVES PER SUFFIX AND LEVEL OF AMBIGUITY (WITH FINE-GRAINED COMBINED SEMANTIC TYPES).

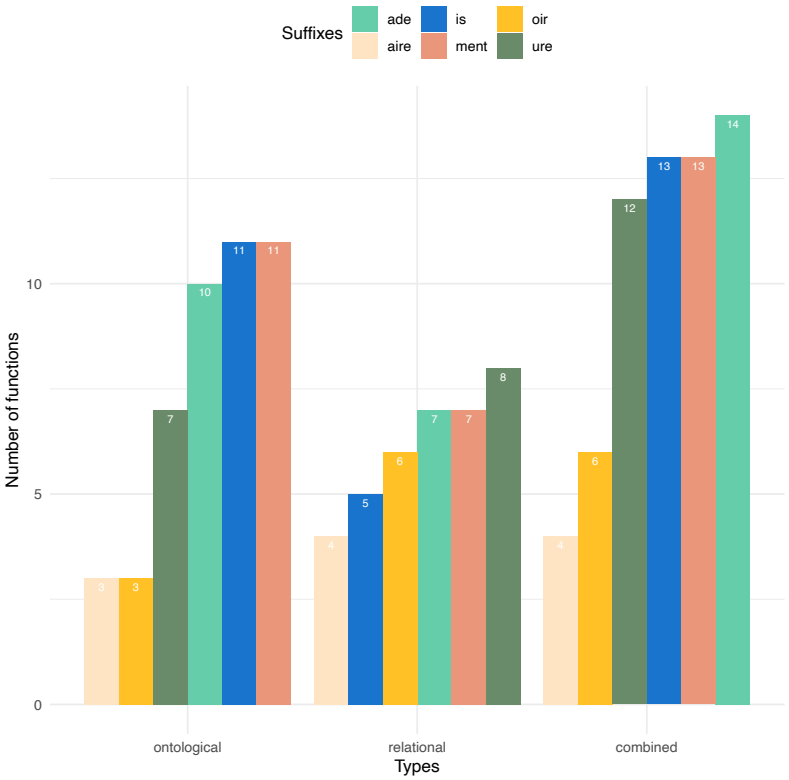
Differences among suffixes can be observed not only in relation to the ambiguity of their derivatives, but also with respect to their degree of polyfunctionality, i.e., the number of different functions they can realize. Although positively correlated,[6] these two properties are not equivalent and variation can be observed among nominalizing suffixes with respect to how the multiple functions of a suffix tend to be instantiated in ambiguous derivatives (Salvadori & Huyghe 2023). Such variation is noticeable when comparing Figure 1 and Figure 2, which indicate the number of semantic functions per suffix based on the fine-grained annotation of the 90 nominalizations. At the lexical level for instance, *-ure* has the highest proportion of ambiguous derivatives (Figure 1a), but it ranks only 4th for polyfunctionality (Figure 2a). At the contextual level, *-ment* outdoes *-ure* in terms of average ambiguity of derivatives (Figure 1b), whereas *-ure* remains first in terms of polyfunctionality (Figure 2b).

Figure 2 provides further information about the semantic functionality of nominalizing suffixes. In particular, the distinction between ontological and relational categories proves to capture different aspects of the semantics of morphological processes. Considering the lexical annotation at the fine ontological level (Figure 2a), *-is* and *-ment* present the highest number of different functions, but *-is* has one of the lowest numbers of fine relational functions. It seems that *-is* has a less cohesive referential meaning, but a more stable derivational relationship with base verbs. The suffix *-ure* in comparison ranks 4th based on ontological categories, but it is the one with the highest number of relational functions. Considering the merged category of combined functions, these differences are leveraged: suffixes that have more ontological or relational functions show a higher number of combined functions. It should be noted that we cannot observe a uniform relation between ontological and relational functions across suffixes. There is no inverse correlation between the number of ontological functions and the number of relational functions, since there are suffixes that behave similarly in both categories, having low (*-aire*) or high (*-ment*) numbers of functions in both of them. The relation between the number of ontological and relational functions realized seems to be specific to each suffix. Similar considerations can be made for the contextually annotated data in Figure 2b.

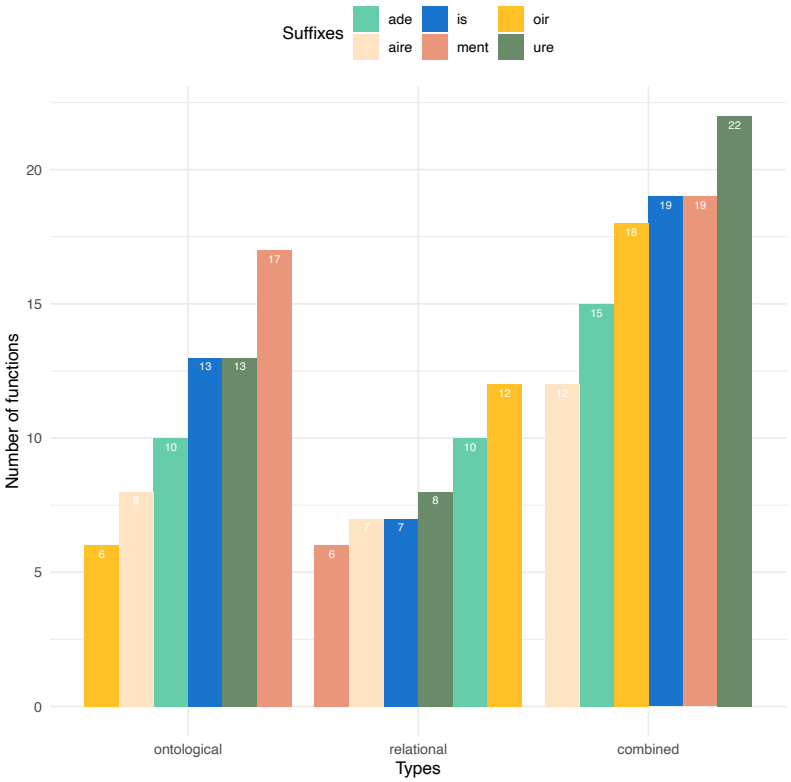
5. EVALUATING SELFSIM AS A MEASURE OF AMBIGUITY

In this section, we investigate the predictability of the lexical ambiguity of deverbal nouns using pre-trained contextualized embeddings from FlauBERT (Le *et al.* 2020). Following previous works on the topic (Garí Soler & Apidianaki 2021; Haber & Poesio 2021), we hypothesize that the average cosine similarity between token vectors is lower for ambiguous than for monosemous nouns. The underlying idea is that distributional representations of monosemous nouns are more consistent and closer across contexts than those of ambiguous nouns, since all occurrences of the former express the same sense. To test this hypothesis, we rely on the different degrees of ambiguity presented in the previous section.

[6] A correlation of $\rho = 0.74$ ($S = 6820.2$, $p < 0.001$) is observed with fine-grained combined types in the lexical annotation, and of $\rho = 0.72$ ($S = 7225.8$, $p < 0.001$) with fine-grained combined types in the contextual annotation.



(a) LEXICALLY ANNOTATED



(b) CONTEXTUALLY ANNOTATED

FIGURE 2: SUFFIX POLYFUNCTIONALITY WITH FINE-GRAINED SEMANTIC TYPES.

We make use of both the lexical and contextual annotations, and investigate which one is better predicted by the FlauBERT model. We also compare the different semantic typologies based on ontological and relational description, as well as degrees of granularity. Finally, we include in the analysis the suffixes of the derived nouns, since as previously observed, the degree of ambiguity of nominalizations varies with derivational processes.

5.1 Methodology

As already mentioned, pre-trained contextualized models can be used to obtain embeddings for single tokens of a word. In such models, algorithms consider the knowledge of a word acquired during the pre-training to represent new occurrences of the word. For this study, we extracte from FlauBERT token representations for each of the 4,500 corpus occurrences we annotated manually. Then, for each of the 90 nouns considered, we compute the pairwise cosine similarity among the embeddings of the 50 tokens and we average them to obtain one value per noun. We refer to this measure as the SelfSim measure,[7] as proposed by Ethayarajh (2019) and Garí Soler & Apidianaki (2021). We compute SelfSim values for each noun considering embeddings from all 12 hidden layers of FlauBERT (plus the initial layer), looking for differences in their ability to represent lexical ambiguity.

In order to evaluate the distributional measure, we compare the SelfSim value obtained for each noun with its degree of ambiguity, i.e., its number of semantic types based on our annotation. More precisely, we fit an ordinal logistic regression model to predict the ambiguity degree considering SelfSim as independent variable. Ambiguity degrees are reduced to three main levels, corresponding to 1, 2, and 3+ word meanings, since higher numbers of meanings are rare. We consider nominalizing suffixes as an additional control variable in the equation. We opt for an ordinal logistic regression instead of a linear regression because of the nature and distribution of ambiguity scores.[8] Ordinal regression indicates how much the probability of having multiple semantic types increases as SelfSim decreases. We expect monosemous words to be associated with higher values of SelfSim, and more ambiguous words to be associated with lower values of SelfSim. Higher values of SelfSim indeed indicate a more consistent semantic space and a closer semantic relationship between tokens.

We build our model with forward selection of the variables, starting from suffix as categorical predictor, then adding the SelfSim variable and the interaction between the two predictors. At each step we test model significance through a chi-squared test performed with the ANOVA function and we keep the additional variable if it significantly improves the model's fit. We perform this analysis considering the 9 semantic typologies to determine ambiguity degrees and the 12 BERT layers as source for the embeddings. Moreover, we consider ambiguity as described in both the lexical and the contextual annotations.

[7] Note that this measure is equivalent to the average cosine similarity between the centroid of the class and every instance of the class, and that this computation is more efficient and would be preferable for larger samples.

[8] The residuals of a linear regression built with the ambiguity score as a numerical response variable were not normally distributed, thus violating the assumptions of this model.

The embeddings have been extracted using Python, the 'huggingface' library, as well as parts of the code provided by Garí Soler & Apidianaki (2021), whereas the statistical analysis has been conducted in R. [9]

5.2 Results: predicting lexical ambiguity

We first present the results for ambiguity annotated at the lexical level. To select the best predicting model, we start by evaluating the suffix as a predictor in comparison to a null model (i.e., a model with no predictors and just an intercept). The fit of the model is significantly improved only when the degree of ambiguity is computed with fine and medium ontological types, as well as with fine and medium combined types. Adding SelfSim as a predictor to these models significantly improves the fit in all cases ($p < 0.05$), but only when computed on Layer 9 (plus Layers 10, 11, and 12 for medium ontological types). The addition of SelfSim computed on other layers did not result in a significant improvement. The best fit is obtained with fine ontological types (Adjusted $R2 = 0.27$). The interaction between the suffix and SelfSim is not significant in any model.[10]

In Table 5, we report the results of the best model for lexical annotation. This model includes ambiguity degrees with fine ontological types as the response variable, and suffixes and SelfSim at Layer 9 as predictors. As shown in Figure 3, the highest values of SelfSim are associated with monosemous words, i.e., words that have been annotated with only one fine ontological semantic type. This result goes in the expected direction. Differences in SelfSim values are not significantly associated with a change in the probability of having 2 or more meanings. Lastly, the differences among suffixes are significant only in the case of *-aire* and *-oir*, which are more frequently associated with monosemous than with ambiguous words, as shown in Figure 4.

		ESTIMATE	STD. ERROR	*t*-value	*p*-value
COEFFICIENTS	*-aire*	-2.02	0.95	-2.13	0.05
	-is	-0.24	0.74	-0.33	n.s.
	-ment	0.20	0.73	0.27	n.s
	-oir	-2.06	0.94	-2.19	< 0.05
	-ure	0.94	0.70	1.34	n.s.
	selfsim.9	-8.22	4.18	-1.97	< 0.05
INTERCEPT	1|2	-5.96	3.12	-1.91	n.s.
	2 | more	-4.01	3.09	-1.30	n.s.

TABLE 5: SUMMARY OF THE BEST MODEL WITH DEGREE OF AMBIGUITY AS RESPONSE, COMPUTED ON LEXICALLY ANNOTATED FINE ONTOLOGICAL TYPES (WITH SUFFIX *-ade* AS REFERENCE LEVEL).

[9] The dataset and the scripts used in this study are available at https://osf.io/8k3de/?view_only=24d7917d33474ca08ca70feb266490bf.

[10] We also tested models with the log-transformed frequency of nouns as a predictor, which did not significantly improve the fit in comparison with the null model.

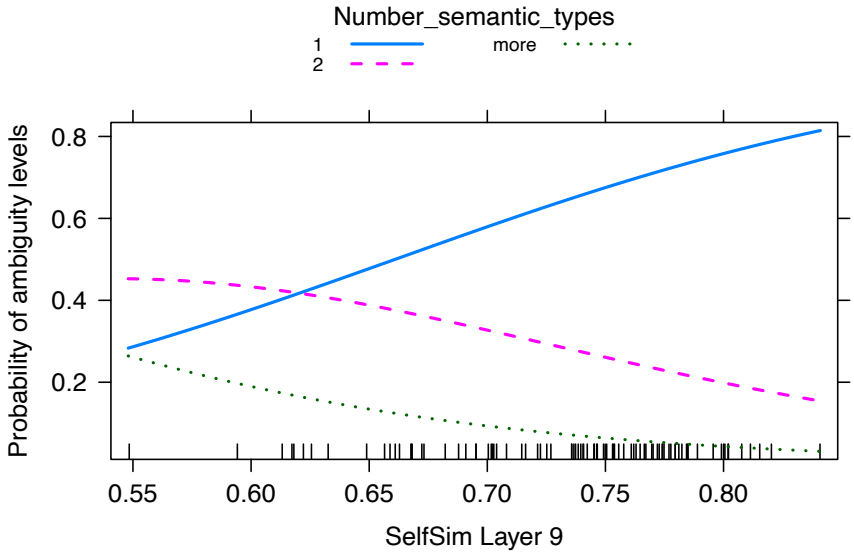


FIGURE 3: EFFECT PLOT FOR SELFSIM AS A PREDICTOR OF AMBIGUITY (BEST MODEL FOR LEXICALLY ANNOTATED DATA WITH FINE-GRAINED ONTOLOGICAL TYPES).

We follow the same procedure to predict ambiguity degrees inferred from the contextual annotation. We build an ordinal regression model with forward selection of variables, considering the different semantic typologies and the different BERT layers. The addition of suffixes as a predictor with respect to the null model is significant when considering all semantic types. Adding SelfSim improves the fit only marginally ($p < 0.1$) for coarse ontological types and only for some layers (4, 5, 6, 10, and 12). The best model to fit the data (i.e., with ambiguity degrees computed on relational coarse types as the response variable, and suffixes as the predictor) achieves an adjusted *R2* of 0.25, followed by the model with ambiguity based on fine ontological types (adjusted *R2* = 0.21). The results of the best model are reported in Table 6. Contrary to what we observe for lexically annotated data (Table 5), all the suffixes are significant for the prediction of ambiguity. The suffixes *-ade* and *-ure* are associated with higher probabilities of having more ambiguous derivatives with respect to the other suffixes, whereas *-aire* more probably derives monosemous nouns.

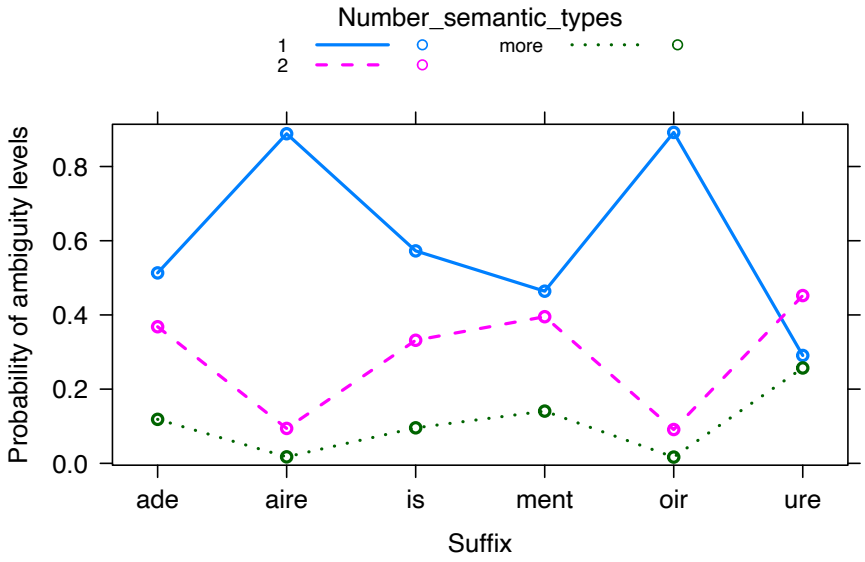


FIGURE 4: EFFECT PLOT FOR SUFFIXES AS A PREDICTOR OF AMBIGUITY (BEST MODEL FOR LEXICALLY ANNOTATED DATA WITH FINE-GRAINED ONTOLOGICAL TYPES).

		ESTIMATE	STD. ERROR	*t*-value	*p*-value
COEFFICIENTS	*-aire*	-1.50	0.92	-1.63	< 0.001
	-is	-0.67	0.78	-0.86	< 0.001
	-ment	-0.48	0.79	-0.60	< 0.001
	-oir	-0.67	0.78	-0.86	< 0.001
	-ure	0.69	0.71	1.97	< 0.001
INTERCEPT	1|2	-0.38	0.52	0.74	< 0.001
	2 | 3	2.84	0.70	4.08	< 0.001

TABLE 6: SUMMARY OF THE BEST MODEL WITH DEGREE OF AMBIGUITY AS RESPONSE, COMPUTED ON CONTEXTUALLY ANNOTATED COARSE RELATIONAL TYPES (WITH SUFFIX *-ade* AS REFERENCE LEVEL).

5.3 Evaluating affix polyfunctionality

In the previous section, we investigated the role of the SelfSim measure in predicting the degree of ambiguity of nominalizations. In this section, we evaluate how well it correlates with affix polyfunctionality. We defined affix polyfunctionality from a theoretical point of view in Section 2, distinguishing it from lexical ambiguity, and we investigated it in Section 4.2 by listing the different semantic types associated with the 6 suffixes under study. As in the previous analysis, we consider two different gold standards, depending on whether the annotation was performed at the lexical or at the token level. We compute the SelfSim measure for the different suffixes considering together all the token embeddings of the nouns ending with the same suffix. We determine cosine similarity for each pair of tokens of nouns ending with a given suffix before averaging the values for all the token pairs of the suffix.

Given the low number of data points available for this analysis (i.e., 6 suffixes), we do not use regression models, but evaluate the correlation between SelfSim and polyfunctionality values by means of a Spearman correlation test. As in the case of lexical ambiguity, we expect SelfSim to be inversely correlated with affix polyfunctionality, with higher values related to lower numbers of semantic functions. For the lexical annotation, we observe a significant correlation ($p < 0.05$) between polyfunctionality and SelfSim only when considering Layer 1 and medium/fine relational types. More precisely, we observe a negative correlation of $\rho = -0.80$ ($S = 63.08$) for the medium granularity and $\rho = -0.75$ ($S = 61.38$) for the fine granularity. For the token-based annotation, we observe a negative correlation between polyfunctionality and SelfSim with the input layer (Layer 0) for medium ontological ($\rho = -0.75$, $S = 44.27$), medium combined ($\rho = -0.75$, $S = 61.38$) and fine combined ($\rho = -0.81$, $S = 63.409$) semantic types. As expected, higher degrees of polyfunctionality correlate with lower values of SelfSim. However, as already mentioned, this correlation is restricted to relational types for lexically annotated data and to ontological and combined types for contextually annotated data, while applying only to one layer of the model in each case.

The suffix SelfSim is also significantly correlated with the average ambiguity of the derivatives of a suffix (Spearman $\rho = -0.75$, $S = 61.38$), but only for combined coarse types and Layer 1, considering the lexical annotation. With the contextual annotation, SelfSim (Layer 0) correlates with average lexical ambiguity for ontological medium, combined coarse and combined medium semantic types ($\rho = -0.77, -0.77, -0.75$; $S= 62, 62, 61.38$).

5.4 Partial discussion

In the above, we have presented the results of the evaluation of a distributional measure (SelfSim) computed on BERT embeddings as a measure of lexical ambiguity. We expected SelfSim to be negatively related to lexical ambiguity: the higher the ambiguity, the lower the SelfSim among the token vectors of a lexeme. We tested this hypothesis computing SelfSim on the different hidden layers of BERT and considering as gold standard for lexical ambiguity the number of semantic types assigned to nominalizations, according to different types of annotation. For the evaluation, we fitted an ordinal logistic regression model that predicts the level of ambiguity by means of SelfSim and depending on the suffix involved in the nominalization process. In what follows, we discuss the results taking into consideration the role of BERT layers, the variety of semantic classifications, the type of annotation performed and the interaction between these three variables.

First, SelfSim is a significant predictor of ambiguity only when computed on the embeddings of some BERT layers. The last four layers were significant predictors for lexically annotated ambiguity, and Layer 9 in particular showed significant results with different semantic typologies. This seems to be in line with previous studies, which related semantic knowledge to the last layers of BERT but also pointed out that the very last layers had more context-specific representations (e.g., Ethayarajh 2019). Layer 9 may be interpreted as a point at which the model has acquired lexical knowledge, but has not associated it with specific contextual features. The significant results with respect to ontological types in the last layers can be related to the fact that these types are more sensitive to contextual variation than relational and therefore combined types—given that relational types describe with the semantic relationship between bases and derivatives, which is not context-dependent.

Regarding the correlation between SelfSim and affix polyfunctionality, the first layers are the only ones associated with significant results. The same is observed for the relation of SelfSim

with average lexical ambiguity. The importance of the first layers in providing information about affix polyfunctionality can be seen as a confirmation of previous studies, which have described BERT first layers as related to word order and morphological information (e.g., Lin *et al.* 2019; Rogers *et al.* 2021). It is worth noting, however, that this basic morphological information may include semantic elements, since polyfunctionality is related to the meaning of affixes or of derivational patterns associated with affixes.

As far as semantic classifications are concerned, ontological and combined types yielded significant results when predicting nominal ambiguity with the lexical data, even if BERT performed better with medium ontological types at a larger number of layers. Based on the contextual annotation, only marginally significant results were observed, and only for coarse ontological types. When predicting suffix polyfunctionality, (medium and fine) relational types were significantly correlated with SelfSim based on the lexical annotation, whereas (medium) ontological and (fine and medium) combined types could be predicted based on the contextual annotation. Suffix average ambiguity was correlated with SelfSim when encoded with the same semantic types as polyfunctionality for the contextual annotation, whereas with the lexical data the correlation was significant when considering combined coarse types. These results confirm the need to keep ontological and relational types separated in the semantic description of nominalizations. Ontological types are informative about the lexical semantics of a noun, whereas relational types provide further information about the semantic aspects of the derivational process. With regard to semantic granularity, we obtained the best results in the main task with fine and medium types, which suggests that a coarse semantic classification is not necessarily the optimal choice to describe semantic features based on distributional data. In some cases at least, distributional information can capture medium-grained or fine-grained semantic elements included in lexical structures. It remains true that, despite significant results, the fit of the regression models was not very high, since the best models we used achieved an adjusted *R2* of 0.27. This result indicates that lexical ambiguity (and lexical semantics in general) is only a part of the linguistic information encoded in BERT embeddings, and that other factors influence the distributional profile of words.

Surprisingly, SelfSim performed better on the data from the lexical annotation than on those from the contextual annotation, achieving only marginally significant results in the latter case. Our initial hypothesis was that the performance would be better for contextually annotated data, given that the BERT embeddings used in the study were based on the set of sentences manually annotated. However, BERT embeddings result from pre-training on a large corpus and, as noted by Garí Soler & Apidianaki (2021), they include previously acquired knowledge in the representations computed for new tokens. This might explain why BERT is better at predicting lexical ambiguity based on a systematic and complete listing of meanings, rather than based on a small sample of sentences. The size of the sample of tokens we annotated was limited by the time required to complete manual annotation, and obviously a larger sample would have offered a more complete picture of the ambiguity of a noun, probably allowing for a better distributional assessment. It may also be the case that the lack of results for contextually annotated ambiguity is due to the specificity of the contextual annotation (and to the challenges raised by the annotation of nominalizations, especially when complex types are involved), or to the nature of the semantic classification used in our study. To test the latter, we decided to further inspect the semantic representation provided by the embeddings, and to investigate the distributional consistency of the different semantic typologies we used to describe the ambiguity of nominalizations. This investigation is presented in the next section.

6. SEMANTIC TYPES IN CONTEXTUAL EMBEDDINGS

Despite significant results obtained in the evaluation of the relationship between SelfSim and lexical ambiguity, the experiment reported in Section 5 showed less promising results than those obtained in previous work (e.g., Garí Soler & Apidianaki 2021).[11] In particular, we did not find significant results when ambiguity was approached through the semantic annotation of samples of corpus occurrences. The reasons for this may be diverse, ranging from the annotation scheme used to encode ambiguity to the size of the samples annotated. We do not suspect the parameters of the distributional model to be the main cause of this negative result (although they may obviously have an effect), since the same model achieved significant results when tested against the lexically annotated data. Moreover, due to the cost of manual annotation, we cannot investigate whether considerably enlarging the annotated samples would improve the results.

Given these premises, we check whether the semantic types used to encode the ambiguity of nominalizations can be identified and discriminated in BERT embeddings. More precisely, we wish to test whether tokens annotated with the same semantic type can be discriminated on the basis of their contextual distributional representations. In order to do so, we train a classifier to distinguish one semantic type from the others, and we evaluate its performance as a probe of its ability to encode this semantic information.

We use as materials the same dataset as in the previous sections, considering the semantic types annotated at the token level described in Section 3.1. As a classifier, we use a Gradient Boosting Machine (Friedman 2001), a popular machine learning algorithm that is based on decision trees and that has proved highly accurate in various linguistic tasks with high efficiency and interpretability (Athanasiou & Maragoudakis 2017; Guzmán Naranjo & Bonami 2023). Specifically, we use lightGBM (Ke *et al.* 2017) in the implementation made available as a Python package. As in the previous experiment, we test the model with the three different semantic classifications (using ontological, relational and combined types), with their three granularities (coarse, medium, and fine), and with representations extracted from the 12 hidden layers of FlauBERT. For each of these settings, we train as many classifiers as there are semantic types. Each classifier has to predict for each of the 4,500 tokens in our sample whether it is an instance of the considered semantic type or not. We therefore frame the problem as a binary classification task where one label is predicted against all the others, in order to reduce as much as possible data sparsity due to the large number of labels.

To evaluate the performance of the classifiers and avoid overfitting, we perform a ten-fold cross-validation on our sample of 4,500 tokens, i.e., we randomly divide the dataset into 10 groups, training the classifier on 9 of the subsets and testing its performance on the subset not used for training, while repeating this process for each of the 10 subgroups. The performance is evaluated by computing the accuracy of the classifier on each fold and averaging it across the 10 folds. To interpret the accuracy values, we compare them to a baseline score obtained through a classifier that always assigns the most frequent value observed in the training set to all the elements in the test set. Also in this case, the values are averaged among the ten folds.

[11] It should be considered, however, that the correspondence between SelfSim and lexical ambigu- ity was evaluated using different methods. We relied on logistic regression, whereas previous work computed correlation scores (Haber & Poesio 2021) or t-test statistics (Garí Soler & Apidianaki 2021).

CLASS	GRANULARITY	MEAN ACCURACY	MEAN IMPROVEMENT	MEAN F1	MEAN PRECISION	MEAN RECALL
ONTOLOGICAL	fine	0.941	0.009	0.925	0.928	0.941
	medium	0.917	0.027	0.903	0.914	0.917
	coarse	0.775	**0.108**	0.772	0.825	0.775
RELATIONAL	fine	0.905	0.014	0.883	0.888	0.905
	medium	0.825	0.024	0.808	0.834	0.825
	coarse	0.840	0.044	0.834	0.855	0.840
COMBINED	fine	**0.951**	0.001	0.934	0.928	0.951
	medium	0.927	0.008	0.909	0.912	0.927
	coarse	0.878	0.021	0.865	0.89	0.878

TABLE 7: SUMMARY OF THE RESULTS OF THE LIGHTGBM CLASSIFIER TRAINED TO PREDICT SEMANTIC TYPES BASED ON FLAUBERT EMBEDDINGS. THE BEST MEAN ACCURACY AND IMPROVEMENT FROM THE BASELINE ARE HIGHLIGHTED.

A summary of the results grouped by classification type and averaged across all layers is given in Table 7. We report the mean accuracy and the improvement from the baseline (i.e., the difference between the accuracy of the model and the accuracy of the baseline), their standard deviations, as well as averaged precision, recall, and F1 scores. The classifier reaches an average accuracy of 0.911, with an average improvement from the baseline of 0.017. As far as semantic classifications are concerned, the combined classification reaches the highest mean accuracy (0.966), whereas the best improvement from the baseline is observed for the relational classification (0.035). In terms of semantic granularity, the fine granularity reaches the highest mean accuracy (0.955), while the best improvement from the baseline is observed with the coarse granularity (0.053). Considering both the semantic classifications and the different granularities, we observe the highest accuracy for fine-grained combined types (0.951) and the best improvement from the baseline for coarse- grained ontological types (0.108). The high accuracy but low improvement observed in the case of fine-grained semantic types may be explained by their low frequency in our dataset. Specific fine-grained types may not be found among the 450 tokens used in a test set, thus bringing the baseline to perfect accuracy and leaving no room for improvement in the classification task.

Lastly, with respect to the layers from which the embeddings were extracted, we observe that accuracy increases linearly, with the exception of the last layer that performs worse than the previous four ones. Layer 11 for fine-grained types reaches the highest accuracy (0.941), whereas Layer 11 for coarse-grained types obtains the best improvement from the baseline (0.072), with an accuracy of 0.871. In Table 8 we report the results aggregated by layer (averaged across classifications and granularities).

We can infer from the classifier results that the semantic typology we used to describe the meaning of nominalizations is captured to some extent by BERT embeddings. Contrary to what could be expected, the last layer of BERT is not the one that provides the best results. It appears that ontological semantic types are better predicted than relational ones, probably because of their stronger link to contextual information, as already noted in the results of the first experiment. With regard to differences in granularity, we believe that the results provided in this section are not conclusive, given the low frequency of fine-grained labels in our sample.

LAYER	MEAN ACCURACY	MEAN IMPROVEMENT	MEAN F1	MEAN PRECISION	MEAN RECALL
1	0.901	0.007	0.881	0.893	0.901
2	0.901	0.007	0.883	0.892	0.901
3	0.904	0.010	0.886	0.895	0.904
4	0.906	0.012	0.889	0.899	0.906
5	0.911	0.017	0.895	0.901	0.911
6	0.911	0.017	0.895	0.904	0.911
7	0.912	0.018	0.897	0.907	0.912
8	0.916	0.022	0.901	0.907	0.916
9	0.916	0.022	0.903	0.908	0.916
10	0.918	0.024	0.904	0.91	0.918
11	**0.919**	**0.026**	0.907	0.912	0.919
12	0.914	0.020	0.899	0.907	0.914

TABLE 8: SUMMARY OF THE RESULTS OF THE LIGHTGBM CLASSIFIER TRAINED TO PREDICT SEMANTIC TYPES BASED ON FLAUBERT EMBEDDINGS AGGREGATED BY LAYERS. THE BEST MEAN ACCURACY AND IMPROVEMENT FROM THE BASELINE ARE HIGHLIGHTED.

7. GENERAL DISCUSSION

As observed in Section 5, the degree of lexical ambiguity in nominalizations could be only partially predicted by the SelfSim measure computed on contextual embeddings. Lexically annotated ambiguity was better predicted than contextually annotated ambiguity, which was unexpected since the contextual annotation was performed on the same sample as the one used to compute the contextual embeddings. However, as noted above, BERT models such as FlauBERT rely on knowledge already acquired during the pre-training on a large sample. The distributional representation these models provide for a specific corpus occurrence can be reminiscent of the whole semantics of the word, thus explaining the difference observed between the two predictions.

With regard to lexical annotation, the SelfSim measure proved to be a good predictor of ambiguity, but only when computed on some layers of the model. Previous studies (e.g., Lin *et al.* 2019; Hewitt & Manning 2019; Vulic´ *et al.* 2020) have shown how the layers are related to different levels of linguistic knowledge, although there is no absolute consensus. Supposedly, the last layers would be the most related to the semantics of words and the most affected by contextual effects (and we find significant results for ontological types at the last four layers), but in our case Layer 9 was the one that yielded significant results with all semantic types. We may hypothesize that this layer captures mostly lexical semantics properties, whereas the last ones encode more pragmatic/context-dependent information (as observed for example by Ethayarajh 2019). Such differences are difficult to demonstrate, but we may infer from our results that the last layer is not necessarily the optimal one for semantic tasks, and that the different layers do include different pieces of information that should not be wasted by considering only one of them. The same conclusion can be drawn from the experiment in Section 6, where a classifier trained on representations from the last layer did not reach the best accuracy score while predicting semantic types. In general, the fit of the regression model

predicting ambiguity was not very high. We considered several possible reasons for this in Section 6, including the irrelevance of the semantic classification used for distributional evaluation, which was the object of our second experiment. The significant results presented in Section 6 allow us to conclude that the semantic classification used to assess lexical ambiguity is not responsible for the predictive weakness of the distributional measure. Other factors should be taken into account, such as sample size and the fact that contextual embeddings capture other linguistic information than core lexical semantics. For example, the distributional properties of a word may be determined not only by its number of senses, but also by usage differences related to genre, style, sociolinguistic variation, or domains of lexical specialization. In addition to the evaluation of the BERT model, we have provided insights into the semantics of nominalizations and its levels of analysis. We have shown that the morphological process used to derive nouns from verbs has an impact on their degree of ambiguity. Some suffixes (e.g., *-ure*) form ambiguous derivatives more frequently than others (e.g., *-aire*). It follows that the lexical ambiguity of affixed words should not be studied without considering the contribution of the different derivational processes, and in that respect, nominalizations are challenging for computational models of lexical ambiguity. As a consequence, we also investigated affix polyfunctionality, showing that it was correlated with our distributional measure. Interestingly, in this case, only the lowest layers of BERT were involved in the significant results, i.e., layers that have been associated with morphological information in previous studies on BERT models. It appears that the distinction we made between affix polyfunctionality and lexical ambiguity in derivatives is captured by BERT, as the layers that best represent these two properties are different.

As shown in Section 4, the description of lexical ambiguity is dependent on the method used to list word senses. A lexical annotation based on lexicographic resources brings different results than an annotation based on a sample of corpus tokens. For example, some nouns annotated as monosemous at the lexical level have more than one semantic type when annotated at the contextual level. These discrepancies can be related to the flexibility of lexical meaning in discourse and to the absence of non-lexicalized occasional meanings in the lexicography. They are also caused by the existence of complex types, identified as cases of inherent polysemy in the lexical description, but frequently disambiguated in context, where semantic facets can be isolated. Consequently, the lexical and contextual annotations should be compared rather than opposed to gain a better understanding of the ambiguity of nominalizations, highlighting the importance of contextual variation in their interpretation.

In this study, we have proposed a specific annotation scheme to describe the ambiguity of nominalizations, involving the distinction between ontological and relational types. The semantic structure of nominalizations is decomposed into referential features and base-related properties inferred from the derivational process. The high inter-annotator agreement for both classifications on the one hand, and the results of our second experiment on the other hand, which provide distributional evidence for the two classifications, make them an appropriate basis for the description of the ambiguity of nominalizations. Comparing the lexical and the contextual annotations, we have observed that relational types tend to be more stable across contexts than ontological types, which are more affected by contextual variation because of referential differences in specific

uses of nominalizations. As far as affix polyfunctionality is concerned, the best correlation was obtained with relational types at the lexical level and with ontological types at the contextual level. This result shows that the semantics of derivational processes involves not only relational types, as could be expected from the fact that these concern the semantic relationship between

bases and derivatives, but also ontological types, thus confirming previous findings on the semantics of deverbal nouns (Salvadori & Huyghe 2023).

A similar remark can be made about semantic granularity. We tested different levels of granularity for ontological, relational and combined semantic types to describe the meaning of nominalizations. Overall, our observations suggest that medium- and fine-grained classifications are distributionally relevant. On the one hand, distributional semantics may be able to capture various aspects of lexical semantics with a certain degree of precision. On the other hand, the semantics of nominalizations appears to be appropriately described with medium- and fine-grained semantic types. The fact remains that an optimal description of polysemy patterns in nominalization should be based on a level of granularity that also matches the semantic specifications of derivational pro- cesses. Such a description would allow an accurate analysis of the combination of lexical and morphological patterns and their respective role in the formation of ambiguous nominalizations. In the current state of research, some uncertainty remains about the semantic granularity of derivational operations, and more investigation is needed to determine the granularity that best characterizes the semantics of nominalizing processes.

8. CONCLUSION

In this paper, we have investigated the ability of contextualized word embeddings to predict the ambiguity of nominalizations in French. We found significant results that are dependent on the hidden layers of the computational models used to extract the embeddings, as well as on the semantic description used to define cases of ambiguity. To evaluate the performance of contextualized word embeddings, we designed a specific annotation scheme and provided a gold standard based on the lexical and contextual annotation of a sample of French nominalizations ending with different suffixes. Our findings contribute to the understanding of the lexical semantic component of contextual embeddings, giving insights into the information they encode and the kind of semantic knowledge they are able to acquire. They also provide an opportunity to discuss issues related to the meaning of nominalizations, including the relevant classification and the appropriate granularity to account for their semantic regularities. The distributional analysis of nominalizations supports the distinction between ontological and relational information in their semantic structure, and indicates the essential role of the suffix in their semantic variability. Accordingly, the ambiguity of nominalizations appears to be determined by a combination of derivational patterns and regular sense extensions that operate in the lexicon and influence the contextual use of the nouns.

REFERENCES

Alexiadou, A. (2019). Event/result in morphology. In M. Aronoff (ed.) Oxford Research Encyclopedia of Linguistics. Oxford: Oxford University Press.

Apidianaki, M. (2023). From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation. Computational Linguistics, 49(2). 1–59. https://doi.org/10.1162/coli_a_00474.

Apothéloz, D. (2002). La construction du lexique français: principes de morphologie dérivationnelle. Paris: Editions Ophrys.

Apresjan, J.D. (1974). Regular polysemy. Linguistics, 12(142). 5–32.

Asher, N. (2011). Lexical meaning in context: A web of words. Cambridge: Cambridge University Press.

Athanasiou, V. & M. Maragoudakis (2017). A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: a case study for Modern Greek. Algorithms, 10(1). 34.

Baroni, M., G. Dinu & G. Kruszewski (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 238–247.

Barque, L., P. Haas & R. Huyghe (2014). La polysémie nominale événement/objet: Quels objets pour quels événements? Neophilologica. 170–187.

Bauer, L. (2017). Metonymy and the semantics of word-formation. In N. Koutsoukos, J. Audring & F. Masini (eds.) Morphological variation: Synchrony and diachrony. MMM11 Online Proceedings, volume 11. 1–13.

Bengio, Y., R. Ducharme & P. Vincent (2000). A neural probabilistic language model. Advances in neural information processing systems, 13.

Bisetto, A. & C. Melloni (2007). Result nominals: A lexical-semantic investigation. In G. Booij, L. Ducceschi, B. Fradin, E. Guevara, A. Ralli & S. Scalise (eds.) On-line Proceedings of the Fifth Mediterranean Morphology Meeting (MMM5), Fréjus. 393–412.

Bojanowski, P., E. Grave, A. Joulin & T. Mikolov (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5. 135–146.

Boleda, G. (2020). Distributional semantics and linguistic theory. An nual Review of Linguistics, 6(1). 213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303.

Booij, G.E. (1986). Form and meaning in morphology: The case of Dutch 'agent nouns'. Linguistics, 24(3). 503–518.

Camacho-Collados, J. & M.T. Pilehvar (2018). From word to sense embeddings: A survey on vector representations of meaning. Journal of Artificial Intelligence Research, 63. 743–788.

Collobert, R. & J. Weston (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning. 160–167.

Cruse, D.A. (1995). Polysemy and related phenomena from a cognitive linguistic viewpoint. In P. St Dizier & E. Viegas (eds.) Computational lexical semantics, 33–49. Cambridge: Cambridge University Press.

De Vries, H. (2021). Collective nouns. In P.C. Hofherr & J. Doetjes (eds.) The Oxford handbook of grammatical number, 257–275. Oxford: Oxford University Press.

Devlin, J., M.W. Chang, K. Lee & K. Toutanova (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (Long and Short Papers), 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dubois, J. (1962). Étude sur la dérivation suffixale en français moderne et contemporain: essai d'interprétation des mouvenents observés dans le domaine de la morphologie des mots construits. Paris: Larousse.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP- IJCNLP). Hong Kong, China: Association for Computational Linguistics, 55–65.

Ferret, K. & F. Villoing (2015). French N-age instrumentals: Semantic properties of the base verb. Morphology, 25(4). 473–496.

Flaux, N. (1999). A propos des noms collectifs. Revue de linguistique romane, 63(251-52). 471–502.

Flaux, N. & D. Van de Velde (2000). Les noms en français: esquisse de classement. Paris: Editions Ophrys.

Fradin, B. (2012). Les nominalisations et la lecture'moyen'. Lexique, 20. 129–156. Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics. 1189–1232.

Garí Soler, A. & M. Apidianaki (2021). Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. Transactions of the Association for Computational Linguistics, 9. 825–844.

Godard, D. & J. Jayez (1993). Types nominaux et anaphores: le cas des objets et des événements. In W. De Mulder, L. Tasmowski-De Ryck & C. Vetters (eds.) Anaphores temporelles et (in-) cohérence, Cahiers Chronos, volume 1, 41–58. Amsterdam: Rodopi.

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. arXiv preprint arXiv:1901.05287.

Grimshaw, J. (1990). Argument Structure. Cambridge, MA: The MIT Press.

Guzmán Naranjo, M. & O. Bonami (2023). A distributional assessment of rivalry in word formation. Word Structure, 16(1). 87–114.

Haas, P., L. Barque, R. Huyghe & D. Tribout (2023). Pour une classification sémantique des noms en français appuyée sur des tests linguistiques. Journal of French Language Studies, 33(1). 52–81.

Haber, J. & M. Poesio (2021). Patterns of polysemy and homonymy in contextualised language models. In Findings of the Association for Computational Linguistics: EMNLP 2021. 2663–2676.

Harris, Z.S. (1954). Distributional structure. Word World, 10(2-3). 146–162.

Hewitt, J. & C.D. Manning (2019). A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (Long and Short Papers). 4129–4138.

Huyghe, R. (2015). Les typologies nominales: présentation. Langue française, 185(1). 5–27.

Huyghe, R., A. Lombard, J. Salvadori & S. Schwab (2023). Semantic rivalry between French deverbal neologisms in-age,-ion and-ment. In S. Kotowski & I. Plag (eds.). The Semantics of Derivational Morphology, 143–176. Berlin: De Gruyter.

Iacobacci, I., M.T. Pilehvar & R. Navigli (2015). Sensembed: Learning sense embeddings for word and relational similarity. In Proceedings of the 53rd Annual Meeting of the

Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, volume 1: Long Papers. 95–105.
Jacquey, E. (2006). Un cas de “polysémie logique”: modélisation de noms d’action en français ambigus entre processus et artefact. TAL, 47(1). 137–166.
Jawahar, G., B. Sagot & D. Seddah (2019). What does BERT learn about the structure of language? In ACL 2019-57th Annual Meeting of the Association for Computational Linguistics.
Jezek, E. (2008). Polysemy of Italian event nominals. Faits de Langue, (30). 251–264.
Kawaletz, L. (2021). The semantics of English -ment nominalizations. Unpublished, Heinrich-Heine-Universität Düsseldorf.
Kawaletz, L. & I. Plag (2015). Predicting the semantics of English nominalizations: A frame-based analysis of -ment suffixation. In L. Bauer, L. Körtvélyessy & P. Štekauer (eds.) Semantics of complex words, 289–319. Berlin: Springer.
Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye & T.Y. Liu (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.
Kleiber, G. (1999). Problèmes de sémantique: La polysémie en questions. Villeneuve d’Ascq: Presses Universitaires du Septentrion.
Lammert, M. (2006). Sémantique et cognition: les noms collectifs. Ph.D. thesis, Université Marc Bloch (Strasbourg)(1971-2008).
Landauer, T.K. & S.T. Dumais (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2). 211–240.
Pelevina, M., N. Arefyev, C. Biemann & A. Panchenko (2016). Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. 174–183.
Pennington, J., R. Socher & C.D. Manning (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
Peters, M.E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee & L. Zettlemoyer (2018). Deep contextualized word representations. In M. Walker, H. Ji & A. Stent (eds.) *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2227–2237.
Petukhova, V. & H. Bunt (2008). LIRICS semantic role annotation: Design and evaluation of a set of data categories. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (eds.) *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. Marrakech, Morocco: European Language Resources Association (ELRA), 39– 45.
Pilehvar, M.T. & N. Collier (2016). De-conflated semantic representations. *arXiv preprint arXiv:1608.01961*.
Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.
Rainer, F. (1996). La polysémie des noms abstraits: historique et état de la question. In N. Flaux, M. Glatigny, D. Samain *et al.* (eds.) *Les noms abstraits. Histoire et théories*, 117–126. Villeneuve-d’Ascq: Presses universitaires du Septentrion.
Rainer, F. (2014). Polysemy in derivation. In R. Lieber & P. Štekauer (eds.) *The Oxford handbook of derivational morphology*, 338–353. Oxford University Press.

Reif, E., A. Yuan, M. Wattenberg, F.B. Viegas, A. Coenen, A. Pearce & B. Kim (2019). Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems*, 32.

Rogers, A., O. Kovaleva & A. Rumshisky (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8. 842–866.

Salvadori, J. & R. Huyghe (2023). Affix polyfunctionality in French deverbal nominalizations. *Morphology*, 33(1). 1–39.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*. Mannheim: IDS, 28–34.

Schäfer, R. & F. Bildhauer (2012). Building large corpora from the Web using a new efficient tool chain. In N. Calzolari, K. Choukri, T. Declerck, M.U. Dogan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association, 486–493.

Schuler, K.K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1). 97–123.

Tenney, I., D. Das & E. Pavlick (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4593–4601.

Thiele, J. (1987). *La formation des mots en français moderne*. Montréal: Presses de l'Université de Montréal.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser & I. Polosukhin (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vulic´, I., E.M. Ponti, R. Litschko, G. Glavaš & A. Korhonen (2020). Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7222–7240.

Wauquier, M. (2022). Apports de la sémantique distributionnelle pour la morphologie dérivationnelle. *Corpus*, (23).

Wiedemann, G., S. Remus, A. Chawla & C. Biemann (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*. 161–170.