Morphology
https://doi.org/10.1007/s11525-021-09382-w

Grounding semantic transparency in context

A distributional semantic study on German event nominalizations

Rossella Varvara[1] · Gabriella Lapesa[2] · Sebastian Padó[2]

Received: 1 April 2019 / Accepted: 15 June 2021

Abstract
We present the results of a large-scale corpus-based comparison of two German event nominalization patterns: deverbal nouns in *-ung* (e.g., *die Evaluierung*, 'the evaluation') and nominal infinitives (e.g., *das Evaluieren*, 'the evaluating'). Among the many available event nominalization patterns for German, we selected these two because they are both highly productive and challenging from the semantic point of view. Both patterns are known to keep a tight relation with the event denoted by the base verb, but with different nuances. Our study targets a better understanding of the differences in their semantic import.

The key notion of our comparison is that of semantic transparency, and we propose a usage-based characterization of the relationship between derived nominals and their bases. Using methods from distributional semantics, we bring to bear two concrete measures of transparency which highlight different nuances: the first one, *cosine*, detects nominalizations which are semantically similar to their bases; the second one, *distributional inclusion*, detects nominalizations which are used in a subset of the contexts of the base verb. We find that only the inclusion measure helps in characterizing the difference between the two types of nominalizations, in relation with the traditionally considered variable of relative frequency (Hay, 2001). Finally, the distributional analysis allows us to frame our comparison in the broader coordinates of the inflection vs. derivation cline.

Keywords Semantic transparency · German · Derivation · Nominalization · Distributional Semantics

✉ R. Varvara
rossella.varvara@unito.it

G. Lapesa
gabriella.lapesa@ims.uni-stuttgart.de

S. Padó
sebastian.pado@ims.uni-stuttgart.de

[1] Dipartimento di Informatica, Università degli Studi di Torino, Torino, Italy

[2] Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart, Germany

Published online: 08 July 2021

1 Introduction

Nominalization is a word-formation process that is highly productive in many languages, and can refer to both the process and the result of "turning something into a noun" (Comrie & Thompson, 2007:334). In this paper, we focus on *deverbal nominalizations*, which turn verbs into nouns (e.g. English *defer-ment*, *activa-tion*), as opposed to de-adjectival (e.g., *red-ness*) or de-prepositional (e.g., *up-ness*) nominalizations. More specifically, we focus on *event nominalizations*, which denote the event itself (*abandon-ment*) or its result state/object (*astonish-ment, contain-ment*) – as opposed to, for example, participant nominalizations, which may denote the agent (*smok-er*) or the instrument (*blend-er*).

Event nominalizations are a subset of event-denoting nouns characterized by a specific derivational history, which distinguishes them from simple action nouns, such as *trip* or *game*. The term nominalization, indeed, points to the transpositional process that takes place when a verb is used as a base for a noun, and conveys the idea that we are talking about a complex word, not a simple one. From a semantic point of view, this class is different from other more prototypical nouns: events are located in time and space, they are perceivable by senses, but their perceptual properties are not constant and stable over time (Lyons, 1977:443).

Languages are usually equipped with multiple affixes which can be applied to verbal roots to produce event nominalizations (Koptjevskaja-Tamm, 1993), and often different patterns apply to different verbal bases. This is also the case for German, the language we focus on in this paper. The set of German event nominalizations include the borrowed *-ion* (*Spekulation*, 'speculate, speculation' from the verb *spekulieren*), the native *-t* (*Fahrt*, 'drive, ride', from *fahren*), *-e* (*Hilfe*, 'help', from *helfen*) and *-ung* (*Verteidigung*, 'defend, defense', from *verteidigen*[1]). In addition, stem-derived nominals like *Fall* ('fall', from *fallen*) and nominal infinitives[2] (e.g. *das Laufen*, 'walking') can also be used to derive an event noun.

In this paper, we focus on the comparison between *-ung* deverbal nouns (henceforth UNGs, *evaluieren*, 'to evaluate' → *die Evaluierung*, 'the evaluation'), and nominal infinitives (henceforth NIs, *evaluieren*, 'to evaluate' → *das Evaluieren*, 'the evaluating'). We selected these two patterns because they are at the same time highly productive (they apply to a large set of verbal bases, making them good candidates for a corpus-based investigation) and rather challenging from a semantic point of view. Given that both UNGs and NIs keep a tight relation with the event denoted by the base verb, we ask what the difference in their semantic import is, and how it is reflected in usage.

This tight relation is demonstrated in the following examples, which show related usages of a base verb (1), a nominal infinitive (2), and an UNG nominalization (3): [3]

[1] Examples from Scheffler (2005:2).

[2] Note that in this article we are not dealing with controversial notions like zero-derivation and conversion. For our purpose, it is not crucial to decide whether nominal infinitives are a case of conversion or derivation since our aim is to describe empirically their semantic differences with respect to other derived nouns (for a summary of the debate on the topic see, for example, Bauer et al., 2013:Chap. 25). In the rest of the paper we will refer to the formation of nominal infinitives as a derivational or word-formation process.

[3] We prefer to show naturally occurring sentences even though they are not completely parallel. Examples taken from https://books.google.de/books?id=qTaMaN2rIesC&pg=PA46,

(1) Die Erfolge der Therapiemaßnahmen können **schnell evaluiert** werden.
‘The successes of the therapeutic activities can be evaluated quickly.’

(2) Die oberste Doktrin ist dabei “fail fast”, also **das schnelle Evaluieren** von Ansätzen.
‘The highest doctrine is “fail fast”, which is the quick evaluation of the approaches.’

(3) Die Eingabe konzeptueller Entwurfsskizzen muss **die schnelle Evaluierung** mehrerer Alternativen ermöglichen [...].
‘The input of conceptual drafts must enable the quick evaluation of multiple alternatives.’

These three examples have parallel argument structures: the evaluation event is realized with a Theme. The Agent is not present, but could be realized in all three sentences as a prepositional phrase headed by *durch* (‘by’). Their semantics is similar and arguably the choice between the base verb (*evaluieren*) and nominalizations (*Evaluierung*, *Evaluieren*) may be due to other syntactic, semantic or discourse factors.

That being said, base verbs and different nominalization patterns often differ clearly with respect to various factors. For example, nominalizations are subject to regularities regarding argument realizations that differ from their base verbs, although there is a lot of debate on the topic (Grimshaw, 1990; Borer, 2005; Alexiadou, 2010). Some studies show an interaction between aktionsart and type of nominalizations. Borer (2005) finds that the English gerundive construction *-ing of* is acceptable with non-culminating events (*the sinking of the ships*, *the falling of the stock prices*) but not with achievements (**the arriving of the train*). Other gerund formations and other derived nominals do not have these restrictions (*the arrival of the train*). Alexiadou (2010) finds the same for Spanish.

Bejan (2007) studies subclasses of German nominal infinitives, finding evidence for a distinction between more nominal and more verbal NIs. The first group can be modified by adjectives and can realize arguments with a post-head *von*-phrase, while the second, more restricted type only allows adverbs and a preceding bare accusative object. Bejan argues that the first type refers to specific events, and the second to generic ones. These conclusions are mirrored in work on Spanish (Schirakowski, 2017) and Romanian (Iordăchioaia & Soare, 2015). These studies, however, do not consider the comparison between NI and UNG nominalizations.

Our work fills (at least partially) the gaps of previous work by a) establishing an empirical comparison between NIs and UNGs, and b) by contrasting the two nominalizations with regard to the relation to their respective base verbs.

The key theoretical notion of our comparison is **semantic transparency** with regard to the base verb, which we expect to differ between the two types of nominalizations. It has been frequently argued in the theoretical literature that whereas inflected forms have highly regular and predictable meanings, derived words, on the other hand, often acquire meanings that are not purely compositional, i.e. are not just

www.freiheitssuche.de/ueber-die-erfolgswahrscheinlichkeit and https://books.google.de/books?id=RLIfBAAAQBAJ&pg=PA156, respectively.

a function of the meaning of their constituents (Booij, 2000; Laca, 2001). In our case, we expect `NIs` to have a more transparent meaning than `UNGs`. `NIs` are less subject to semantic shifts and their semantics remains closer to the one of the original base verb: this property can be directly linked to their inflectional origin (discussed in Sect. 2.1).

The predictability and regularity of meaning of nominalizations has not been tested empirically. The major factor influencing transparency that has received empirical treatment is *(relative) frequency* (Hay, 2001). Hay observes that if the base word is less frequent than the derived word, then its meaning is less accessible to speakers during processing, and the output of the derivational process is likely to become less transparent with respect to the semantics of the base. Given her findings, in our experiments we consider relative frequency as a proxy of semantic transparency. However, even in Hay's study, semantic transparency was not measured in a corpus-based fashion, but approximated as the presence of an explicit referent to the base word in the dictionary definition of the derived word. This method is limited since it accounts for transparency as a binary feature, whereas it is theoretically assumed to be a graded notion. A similar facet of semantic transparency has been investigated by Bonami and Paperno (2018), named by the authors as *stability of contrast*. Their study was aimed at investigating the difference of "stability" between inflectional and derivational morphological processes. Due to the topic and the methodology applied, our work is directly linked to theirs, even if some major differences are present. Our focus is indeed on derivational processes, specifically on nominalizations.

We propose to investigate the relationship between nominalizations and transparency with a methodology that builds directly on large corpora of naturalistic language and can integrate frequency considerations with more fine-grained semantic observations. Concretely, we employ distributional semantics (Harris, 1954; Firth, 1957) which builds on the assumption that the meaning of a word (e.g., *dog*) can be empirically approximated in terms of a list of words which frequently occur in its context (e.g., *bark, bone, run*). The most frequent context words can be interpreted as the most salient semantic features of a word (e.g., typical actions performed by the word referent, typical patients of such actions, locations in which the word referent is typically found, etc.): in this perspective, the list of most salient contexts can therefore be considered as a rich, usage-based counterpart of the lexical entries employed in formal semantics. The meaning of two words can be straightforwardly compared in terms of the extent to which their *usage-based lexical entries* overlap: similar words occur in similar contexts. This approach to the quantification of similarity can also be applied to words that are connected by a derivational history: this is exactly what we do in this paper.

The distributional semantics literature offers various strategies to capture the different nuances of the very broad notion of semantic similarity/relatedness. In this study, we test the predictions with two distributional semantics measures: *cosine similarity*, commonly employed to model *synonymy*, and *distributional inclusion*, commonly employed to model *hypernymy/troponymy*. Furthermore, given the correlation between frequency and transparency shown in previous work in the morphological literature, we integrate absolute and relative frequency into our analysis. Our results show that, despite the shared semantics and the comparable syntactic behavior, the distributional inclusion measure can capture fine-grained distinctions between UNG and NI nominalizations, over and above the effect of frequency, which still plays a major role in supporting the distinction between UNG and NI.

As a final contribution, we extend our comparison of UNGs vs. NIs in two directions. First, we explore the potential of our distributional measures in extending the comparison to the agentive nominalization *-er* (*der Evaluierer*, ‘the evaluator’): in this case, the derivational process almost consistently erases the eventive reading (preserved in UNG and NI) and produces a noun which almost consistently denotes an external argument (agent or instrument, vs. UNG which has a result object reading). Second, we explore the inflectional nature of NI by introducing in the comparison a case of inflection, namely the present participle *-end* (*evaluierend*, ‘evaluating’).

The paper is structured as follows. In Sect. 2, we provide the theoretical background of our work: a description of the target phenomenon and of our research questions (Sect. 2.1), and a definition of the key notions we employ in our analysis, along with their empirical treatment in previous work (Sect. 2.2). In Sect. 3, we outline distributional semantics and conceptualize it in relation to our research questions. In Sect. 4, we present our experiments: we start by introducing our experimental setup and proceed to discuss our main results for the comparison between UNG vs. NI. After that, we incrementally enlarge our picture by introducing the comparison with *-er* derivatives and then the one with the present participles, followed by a comprehensive discussion of our experimental results and of their interpretation. Section 5 wraps up the paper by drawing general conclusions and discussing open questions and future work.

2 Background

In this section we introduce the word-formation processes that are the object of the study: nominal infinitives and derivatives formed with the *-ung* suffix. We compare UNG and NI synchronically in terms of their semantics, as well as productivity and syntactic behavior, and provide a diachronic account of their development. After that, in Sect. 2.2, we define the core theoretical notion of our work, namely semantic transparency.

2.1 Two patterns of German nominalization

In this study we focus on two specific patterns of German event nominalizations which can be frequently formed from the same base: deverbal nouns in *-ung* (UNG) and nominal infinitives[4] (NI).

Deverbal nouns in -ung: UNGs manifest a large range of meanings. Adopting the example of *Absperrung* from Rossdeutscher & Kamp (2010), the word can denote:[5]

[4] In the DerivBase derivational lexicon for German (Zeller et al., 2013), 1667 verbs (out of 2278 verbs covered in the lexicon) have both a UNG nominalization and a nominal infinitive.

[5] All following examples from the SdeWAC corpus (Faaß & Eckart, 2013).

- An event (the event of cordoning off)

(4) Heerespioniere der 6. Armee besorgten die Absperrung der Schlucht
'Pioneers of the 6th army carried out the cordoning off of the gorge'

- A result state (the state of an area having been cordoned off)

(5) Die europäische Kultur hat ihre Stärke unter anderem daraus gewonnen, dass sie keine Absperrungen auf Dauer zuließ [...]
'European culture has drawn its strength, among other things, from the fact that it did not allow permanent states of cordoning off [...]'

- A result object (the barricade that was erected)[6]

(6) [...] die Veranstalter nahmen an der Absperrung umgehend bauliche Verbesserungen vor
'[...] the organizers promptly applied structural improvements to the barricade'

The contextual constraints that disambiguate individual occurrences with regard to the available readings have been object of numerous studies (Ehrich & Rapp, 2000; Hamm & Kamp, 2009; Kountz et al., 2007; Spranger & Heid, 2007; Eberle et al., 2009), but sometimes it is difficult to clearly discern the different readings in the same token, since these meanings are strictly interconnected.

As far as their productivity is concerned, UNGs are considered to be one of the most productive among the class of event-denoting suffixes (Eisenberg, 1994:364, Shin, 2001:297). Yet, there are some restrictions on the formation of UNGs, which have been highly debated in past literature. Esau (1973), Bartsch (1986), and Demske (2002), among others, show that verbs expressing states or verbs referring to the beginning or the repetition of a situation do not allow *-ung* nouns. However, some counterexamples are presented by Knobloch (2003:338),[7] e.g. *Erblindung*, 'loss of sight', *Erkaltung*, 'becoming cold'. Rossdeutscher & Kamp (2010) notice that most *-ung* nouns are derived from transitive verbs. Moreover, they argue that verbs that do not allow *-ung* nominals can be generally defined as activity verbs, like *arbeiten*, 'to work', or *wischen*, 'to wipe'.

The syntactic and morphological behavior of UNGs is typical of common nouns: they can be pluralized, their arguments can be realized either by a possessive pronoun or by a post-nominal genitive, they can be modified by adjectives and preceded by a definite or indefinite determiner (Demske, 2002; Scheffler, 2005).

Nominal infinitives: NIs are truly transpositional, since they keep only the event reading from the base verb. In a few cases, a result state or result object reading is also

[6] The basic distinction between event and result state is parallel to what Grimshaw (1990) called complex and simple event nominals, and has been reused, expanded or re-defined, for various languages in many different studies, such as Ehrich and Rapp (2000), Alexiadou (2001), Heyvaert (2003), Melloni (2007), Bauer et al. (2013), among many others.

[7] Cited in Hartmann (2014).

possible (e.g. *Verstehen*, *Ansehen* and *Schreiben*), but these are usually lexicalized words that are far more frequent than other nominal infinitives.

There are practically no constraints on NI productivity, as they can be formed from every base verb.

From a syntactic point of view, NIs exhibit a clear nominal behavior. They are usually preceded by a definite or indefinite article and can be modified by adjectives, like nouns are:[8]

(7) Das Laufen fiel ihm immer schwerer.
'Walking was getting harder for him.'

(8) Es herrschte ein Laufen und Springen, ein Rennen und Hüpfen.
'There was running and jumping, racing and hopping.'

(9) Das schnelle Zerstören der Stadt war notwendig.
'The rapid destroying of the city was necessary.'

Arguments expressed by a genitive or a possessive pronoun can refer both to the subject or object of the verb. A subject interpretation is preferred (Knobloch, 2003; Scheffler, 2005), even though an object reading is possible:

(10) [Dieser Raum enthält vertrauliches Material.] Sein Betreten ist verboten.
'[This room contains confidential data.] Its stepping-in is forbidden.'

They can be compounded to produce further nominals:

(11) Wir schreiben Briefe → Briefeschreiben
We write letters → letter-writing

Contrary to common countable nouns, they do not pluralize and are non-countable:

(12) *Die Zerstören der Stadt waren notwendig.
The destroyings of the city were necessary.

It is possible to derive NI from the passive form of the base verb:[9]

(13) Das Gesehen-werden ist die Hauptdimension der Kunst.
'The being seen is the main dimension of art'.

The consistently nominal behavior of German NIs is of particular interest from a cross-linguistic perspective, as in other languages NIs exhibit mixed features. In Italian, for example, NIs can have both more verbal patterns (when they are modified by adverbs, ex. 14, or express their direct object as a NP, ex. 15) and more nominal ones[10] (where the subject is expressed by a PP and the NI is modified by an adjective, ex. 16).

[8] Examples from Scheffler (2005:7) and Koptjevskaja-Tamm (2006:656).

[9] Examples from Gaeta (1998:5).

[10] See Skytte (1983), Skytte et al. (2001), and Zucchi (1993) for a more extensive discussion of Italian nominal infinitives.

(14) Il lavorare continuamente di Luigi
'Luigi's working continuously'

(15) Il degustare un buon bicchiere di vino
'(The) tasting a good glass of wine'

(16) Il lavorare continuo di Luigi
'Luigi's continuous working'

A diachronic perspective on German event nominalizations: The nominal behavior exhibited by NIs is the result of a change through history which is, in fact, parallel to the one experienced by UNGs (Werner, 2013).

In the Early New High German (ENHG) period, (around the 16th and 17th century), UNGs showed the same argument structure and event interpretation as their corresponding base verb (Demske, 2002:68, but also Göransson, 1911; Behaghel, 1923[11]). This verbal behavior is, indeed, similar to more verbal infinitives. Only in recent times have they evolved a more noun-like character, with increasing restrictions on their productivity. Nominals derived from verbs of states or from inchoative/ingressive verbs, which are not attested in Present Day German (PDG), are attested in ENHG, as Demske (2002:80) shows with a corpus study on newspapers of the 16th and 17th century. In PDG these missing UNGs seem to have been replaced by NIs.

In similar way, NIs went through a change from the verbal to the nominal pole (and also, as argued by Gaeta, 1998, from the inflectional to the derivational side). In Old and Middle High German, NIs had more mixed properties: they could be modified by adverbs, prepositional phrases and direct objects (Gaeta, 1998:6).

Summing up, NIs used to exhibit a mixed behavior which got lost in present day German, preserving only more nominal properties. Thus, as discussed above, they are closer to other event-denoting nominals than the corresponding nominal infinitives in other languages.

2.2 Semantic transparency in derivation

There is wide consensus in the morphological literature that *transparency* is the major semantic criterion to contrast inflection and derivation.[12] Definitions of transparency build on the notion of compositionality: "A lexeme is said to be transparent if it is clearly analysable into its constituent morphs and a knowledge of the morphs involved is sufficient to allow the speaker-listener to interpret the lexeme when it is encountered in context." (Bauer, 1983:19). Dressler (2005:271) claims that inflection can be completely compositional, while word formation (including derivation) cannot.[13] Plag (2003:15-16) exemplifies this with the word *interview*. He notes that its

[11] Cited in Gaeta (1998).

[12] Numerous works have focused also on the role of semantic transparency in the processing of compounds (for example Bell & Schäfer, 2016, and references therein). However, we will not deal with their findings since the study of compounds semantic transparency implies rather different questions, e.g. the role of the two constituents and the relation between the two.

[13] However, exceptions to the transparency/compositionality status of inflections can be found, as well: an example from English is the difference between the plural *clothes*, which has the meaning of 'garnments', and the singular form *cloth*, which means 'woven material' (Bybee, 1985:88, Booij, 2000:364).

meaning “is not the sum of the meaning of its part. The meaning of *inter-* can be paraphrased as ‘between,’ that of (the verb) *view* as ‘look at something’ (definitions according to the *Longman Dictionary of Contemporary English*), whereas the meaning of (the composed verb) *interview* is ‘to ask someone questions, especially in a formal meeting.’ Thus the meaning of the derived word cannot be inferred on the basis of its constituent morphemes; it is to some extent opaque or non-transparent.”

If a derivational process is compositional in the sense defined above, then the meaning of the derived word should be predictable given the base word. Some authors focus on this property: Aronoff (1976:38) uses the term *coherency*, Bauer et al. (2013:34) call it *semantic regularity* or *meaning semantic constancy*. Bell and Schäfer (2016:158), who investigate transparency in nominal compounds, call this *meaning predictability*. Coherency and compositionality are arguably just different sides of the same coin. In the words of Bybee (1985:88), when the semantic relation between base and derived word is no longer transparent, we have a *lexical split*.

Previous work on the empirical characterization of transparency in derivation has highlighted its relation with frequency. Bybee (1985, also 1995) was the first to posit a connection with the frequency of the derived word: the more frequent the derivative is, the more likely it is subject to lexical split.[14] However, as noted by Hay (2001, 2003), the semantic transparency of a given word is better explained by its relative frequency, defined as the ratio of the derivative’s frequency to the frequency of the base term. Derived words with low relative frequency (i.e., derived words that are less frequent than their bases) are more semantically transparent than words with an high relative one. Hay empirically quantifies lexical split (also called semantic drift) by looking at dictionary definitions. She considers a derived word to be transparent if the base word is cited in its dictionary entry. Thus, the word *dishorn* is considered transparent since its definition in the Websters 1913 Unabridged English Dictionary is “To deprive of horns”. On the contrary, if the base word is not present in the dictionary entry, the derived word is considered opaque and considered as an instance of lexical split. From her study, Hay concludes that “dictionary calculations reveal that derived forms that are more frequent than their bases are significantly more likely to display symptoms of semantic drift than derived forms containing higher-frequency bases.” (Hay, 2001:1041).

Hay’s work, however, does not take into consideration an important theoretical feature of transparency. Since Cruse (1986:39), semantic transparency has been described as a graded feature, not as a binary one. Words may be more or less transparent with respect to their base. Her methodology is not adequate to catch continuous values. What is missing is an empirical way to compare the semantic representations of base and derived words on a large scale and based on naturalistic linguistic data (vs. dictionary definitions). In the following section, we outline a corpus-based methodology which allows us a) to gather usage-based representations for base and derived words from a large amount of natural data and b) to use these representations

[14] The link between frequency and transparency is also motivated by psycholinguistic considerations concerning the way in which the word is stored and processed. In a dual-route theory of processing (see e.g. McQueen & Cutler, 1998), low frequency words are processed in a decomposition route, in which the meaning of the derivative is composed from the meaning of its constituents; high frequency words, instead, are processed in a whole-word route, since their resting activation is higher than for their bases.

to get linguistically motivated insights into the semantic relations between derived words and their bases.

3 Methodology

In this section, we present our corpus-based methodology for the semantic characterization of the relation between base and derived words. In Sect. 3.1, we provide an introduction to distributional semantics which can be skipped by readers already familiar with this method. In Sect. 3.2, we discuss its potential for the investigation of morphological processes. Section 3.3 motivates the transparency measures adopted in our study and discusses our predictions on our target patterns.

3.1 Distributional semantics

Distributional semantics is a widely used method in computational linguistics which builds on the assumption that the meaning of a word (the so-called *target*, e.g. *dog*) can be successfully approximated in terms of its linguistic contexts (e.g. words that are used together with it in a sentence, such as for example *bark, bone, run*).

The foundations of distributional semantics go back to the structuralist take on meaning which characterized early corpus linguistics: in his *Distributional Structure*, Harris (1954), stated that “difference of meaning correlates with difference of distribution”; in the highly-cited words by Firth (1957:11), “You shall know a word by the company it keeps!”. Later on, in the ’90s, the *Distributional Hypothesis* elaborated by George Miller and Walter Charles provided the psychological arguments for a usage-based characterization of meaning on the basis of contextual information[15] (Miller & Charles, 1991).

In practice, distributional approaches to word meaning extract co-occurrence information from large corpora. For each target word, a distributional representation is constructed in the form of a vector of its co-occurrence values with the contexts considered. For example, the target word *dog* can be represented with an ordered list of values of its co-occurrence with a set of context words: *dog*: {*bark*: 100, *pet*: 35, *meow*: 1}. Such representation enables an empirical comparison of the meaning of *dog* to that of other target words, i.e., *cat*: {*bark*: 4, *pet*: 43, *meow*: 97}. Indeed, this *differential* approach to meaning is a signature feature of distributional semantics.

Parameters of distributional models. Crucially, there is not just one method to build a distributional semantic model (i.e., a collection of vectors of contexts for a set of targets). Indeed, the Distributional Hypothesis can (and has been) implemented in

[15] “What people know when they know a word is not how to recite its dictionary definition: they know how to use it (when to produce it and how to understand it) in everyday discourse [...]. Knowing how to use words is a basic component of knowing a language, and how that component is acquired is a central question for linguists and cognitive psychologists alike. The search for an answer can begin with the cogent assumption that people learn how to use words by observing how words are used. And because words are used together in phrases and sentences, this starting assumption directs attention immediately to the importance of context” (Miller & Charles, 1991:4).

Table 1 Five sentences corpus for the word *cat*

ignoring each other, like a	**cat**	that puts its head under
of our beloved dogs or	**cats**	having worms.
They all love Latin, as	**cats**	loves milk.
There are strange frogs,	**cats**,	rats, lizards, and even more
a cheeky, affectionate and sociable	**cat**	and we just love him.

Table 2 Toy distributional vectors for *cat* and *bird*

	dog	worm	philosophical	Latin	love	milk	frog	fly	...
cat	1	1	0	1	3	1	1	0	...
bird	0	0	0	1	2	0	1	5	...

many different ways and concrete implementations, corresponding to different model architectures and parameters.

The most intuitive model architecture is the one based on co-occurrence counts, which gives rise to the so-called bag-of-words (or *count*) Distributional Semantic Models (DSMs), which we also employ in our work. The toy vectors for *cat* and *bird* reported below (Table 2) are an example of such an approach. The construction process of such DSMs involves a large number of design choices (parameters); for a comprehensive overview, see Turney and Pantel (2010). One such parameter is what counts as a linguistic contexts. Frequently, they correspond to the target's collocates. They can be extracted from a specific window of text around the target word: a 5 words window around the target, or the whole sentence, or even a whole document.[16] Another option is to employ syntax to determine the most representative collocates (e.g., objects and subjects for verbs, adjectival modifiers and head verbs for nouns, etc.).

In this paper, we will employ a bag-of-words DSM which is based on co-occurrence within a context window: let us demonstrate the extraction process at work on a concrete example. Table 1 shows five example sentences containing the word *cat*, extracted from a large web-corpus for English, UkWaC (Baroni et al., 2009). To our example sentences, we applied a 5-word window to identify potential contexts (i.e., 5 words to the left and to the right). The extraction process stops at sentence boundaries and ignores punctuation. Closed class words (known also as *function words*, e.g., *a, its, our*, etc.) are excluded from the potential contexts and both target and context words are lemmatized. The extraction process outlined above results in the toy vector for the word *cat* displayed in Table 2 together with a putative vector for a second target, *bird*. Note that negative information is also relevant: for the difference between cats and birds, it is important that only the latter occur in the context of the verb *fly*.

[16] In some approaches (common in information retrieval), contexts correspond to the document in which the target word occurs (e.g. Salton et al., 1975; Landauer & Dumais, 1997). In those cases, two words are seen as similar if they occur in the same documents.

Co-occurrence counts between targets and contexts (e.g., 5 for the word *bird* occurring with *fly*) are the most straightforward way to quantify the salience of a certain context for a certain target: in our toy example, the context *fly* is, unsurprisingly, the most salient feature for *bird*. However, raw co-occurrence is not the most statistically robust option, due to the well known Zipf's law (few very frequent words and a large number of rare ones; Zipf, 1949). To counteract frequency effects on collocations, different forms of weighting functions are usually applied to vectors in order to assign lower values to high frequent contexts which are not informative about the target (i.e., because they are just very frequent in the entire corpus). One of the most common functions is PPMI (positive pointwise mutual information, Church & Hanks, 1990; Bullinaria & Levy, 2007), which computes the (logarithmic) ratio of the actual joint probability of two terms (i.e. their observed co-occurrence value) to the "expected" joint probability if the two were independent (Evert, 2005:ch. 3).

Furthermore, collecting co-occurrences for a large number of targets will necessarily lead to an extremely high dimensional and sparse representation, i.e., vectors will have a large number of contexts (this is the so called "curse of dimensionality") but also a very high number of zeros (for context words not co-occurring with a given target). Moreover, there would still be a high number of dimensions which are not informative enough, despite the application of feature weighting. In this scenario, generalizations brought by similar dimensions can be missed. A common solution to this problem is to reduce the number of dimensions into fewer latent dimensions, which reveal latent abstract features and remove sparsity and noise (Deerwester et al., 1990; Landauer & Dumais, 1997). Frequently used algorithms are singular value decomposition (SVD), principal component analysis (PCA) and nonnegative matrix factorization (NMF). Such methods produce a "statistically grounded" summary of the context dimensions, that is to say, one which has a lower dimensionality (from hundreds of thousands to hundreds) and benefits from the latent relation between the original dimensions. The side effect of this is that context dimensions cannot be mapped to a specific word anymore (e.g., *fly* for *bird*), making the DSM representation opaque and specific measures (including the one we rely on in this paper) are not applicable.

Whether to rely on syntactic or surface co-occurrences, the size of the context window, whether to apply feature weighting with PPMI (or other measures) and to resort to dimensionality reduction (and with which methods) are precisely the design choices hinted at before. Technically, they are referred to as the *parameters* of a DSM, and their manipulation obviously affects the contextual representation encoded in the distributional vectors.

Semantic similarity and relatedness. Even though the interpretation of distributional vectors as direct lists of *conceptual features* for the target words is tempting, there is typically no direct correspondence, and extracting linguistically or cognitive plausible features requires additional steps (Baroni et al., 2010; Rubinstein et al., 2015). Therefore, distributional semantics typically sets itself the more modest goal of modeling the *degree of semantic similarity* or *semantic relatedness* among targets. As sketched above, this is a direct corollary of the distributional hypothesis: complementary occurrence provides negative evidence for semantic similarity.

Besides the technicalities of the computation of similarity based on context overlap, which we will discuss in detail in Sect. 3.3, a natural question to ask concerns the semantic nature of DSM similarity. The answer is that there is not just one “similarity”. Indeed, one of the main criticisms against DSMs is that their implementation of semantic similarity may be just too broad to be useful, as it encompasses a wide range of relations with different linguistic properties (Sahlgren, 2006; Lenci, 2008) and a more neutral word such as “relatedness” should be used.[17]

The refinement of the distributional notion of semantic similarity is an active topic of research (Turney, 2008; Baroni & Lenci, 2010): it has been shown, for example, that DSMs can learn relation-specific representations to discriminate between pairs of candidate relations, e.g. synonymy/antonymy (Scheible et al., 2013; Santus et al., 2014). Among previous work on tailoring DSM representations to specific semantic relations, research on the hypernymy/hyponymy relation in terms of *distributional inclusion* (Weeds et al., 2004; Clarke, 2009; Lenci & Benotto, 2012) is of particular relevance for this paper; see Sect. 3.3 for details.

Distributional semantics beyond co-occurrence counts As sketched above, bag-of-words DSMs are only one of the possible model architectures. The models outlined above (and employed in this paper) have been labelled *count models* (Baroni et al., 2014), because they create a distributional representation by accumulating co-occurrence counts. In recent years, new distributional semantic models that are based on neural networks have become increasingly popular and are state-of-the-art in the majority of the tasks (Baroni et al., 2014; Mandera et al., 2017). They are referred to as *predict models* or word embeddings. Indeed, instead of accumulating co-occurrence counts, these neural architectures are trained in the task of *predicting* the contexts given a target, or a target given the contexts. The low-dimensional, implicit representation built by the network to perform this language prediction task is then employed as a DSM representation for the target words. In this sense, the DSM vectors (now called *word embeddings*) come into being as a by-product of the task performed by the network.

Unfortunately, the interpretability of neural models tends to be even worse than that of (dimensionality reduced) count models (Lenci, 2018). Additionally, a lot of work targeting an explicit comparison between count and predict models has uncovered the underlying mathematical equivalence between the two architectures (Levy & Goldberg, 2014b), leading to a refinement of the scope, and constraints on the superiority of predict models over the count ones (Levy & Goldberg, 2014a; Levy et al., 2015; Sahlgren & Lenci, 2016). At any rate, the specific properties of the similarity measure we decided to employ (see Sect. 3.3 for more details) have restricted our choice to models with interpretable dimensions, and thus to the count models without dimensionality reduction.

[17] Semantic relatedness is assumed as a broader term that subsumes any kind of lexical or functional associations, from meronymy and synonymy to simple semantic field association; semantic similarity is a narrower concept that also takes ontological considerations into account and that defines relations such as synonymy and hyponymy (Budanitsky & Hirst, 2006; Kolb, 2009).

3.2 Investigating morphology with distributional semantics

Distributional semantics approaches to model the meaning of morphological processes fall into two high-level categories. In the first group, we find those studies which aim at mining the semantic import of a derivation by comparing the distributional representation of the input (the base word) with that of the output (the derived word). In the second group, we find the approaches which aim at learning distributional representations for morphemes (instead of inferring them). The present work belongs to the first set of approaches, as we directly compare base and derived words to get more insight into the nature of the corresponding semantic shifts.

Let us then take a closer look at the relevant literature. As anticipated above, the starting point of a distributional investigation of the meaning shifts produced by a morphological process (e.g., the German suffix *-in* which derives a female noun from a male noun, *Bäcker* → *Bäckerin*) is the extraction of the distributional vectors for pairs of base and derived words (*Bäcker*, *Bäckerin*; *Ingenieur*, *Ingenieurin*, etc.). Once a representative sample of base/derived pairs for a specific derivation of interest has been collected, distributional semantics provides two alternative (but not mutually exclusive) ways to exploit these usage-based representations to characterize the underlying meaning shifts.

The first approach is based on the comparison between the vector of base and derived words in terms of distributional properties tailored to the theoretical questions addressed. For example, distributional methods devised to detect the antonymy relation can be employed to model the meaning shifts produced by negating prefixes (e.g., *happy* stands to *unhappy* in the same relation in which it stands to *sad*). There is surprisingly little work which adopts such an "analytic" approach to target specific theoretical questions on the nature of the derivational processes. Wauquier (2016) employs this approach in the comparison between French agent nouns in *-eur*, *-euse* and *-rice* (e.g. *gagneur*, 'winner') and event nominalizations (derived by means of different suffixes). By comparing the cosine distance between the base verb and the two derivatives, she found that `-eur` derivatives were further from the base than the corresponding event noun. Our study goes in the same direction and provides a contribution using this approach by extending the scope of the involved measures (cosine, distributional inclusion, as well as frequency effects).

The second approach frames derivation as a compositional process in the Fregean sense: the meaning of the derived word (e.g., *Bäckerin*, female 'baker') can be predicted as a function of the meanings of its parts (*Bäcker*, 'baker', plus *-in*, female suffix). In practical terms, this means learning a function for each derivational pattern (*-er, -in*, etc.) that takes the representation for the base word as input and returns a representation for the derived word. This approach addresses, from a practical point of view, the problem of data sparsity for rare (but productive) derived words. Very often, derived words are less frequent than their bases, and vectors for them can be of a lower quality or completely unavailable (in the case of low frequency and unattested words, respectively). Compositional models allow us to build a semantic representation also for missing complex words, given vectors of the base and of the affix involved (obtained for example by averaging those of the derived words available).

The compositional approach to derivation and compounding is very popular in distributional semantics and has been studied with some success for a number of languages and morphological processes (see Lazaridou et al., 2013; Marelli & Baroni, 2015 for English, Padó et al., 2016; Cotterell & Schütze, 2018 for German, and Melymuka et al., 2017 for Ukrainian, Günther & Marelli, 2018 on compounding). While it is less common in more traditional, count DSMs (but see Keith et al. (2015) and the other works cited above), this approach has found plenty of application in neural DSMs (Luong et al., 2013; Cotterell & Schütze, 2018), and it also characterizes the discriminative learning approach by Baayen et al. (2019) (who, however, do not focus on subword units such as *-s*, but on semantic units such as PLURAL). The compositional approach does, however, build on the straightforward assumption that derivational shifts are fully learnable – in the sense that they are systematic and predictable based on the meaning of the base word. While the transparency assumption is likely to be met at the phrase level (for which compositional distributional semantic approaches have been devised) it is, however, much less straightforward below word level. Derived words may indeed show different degrees of compositionality.

In the studies by Bonami and Paperno (2018) and Huyghe and Wauquier (2020), the authors use an average among word vectors or a centroid, to distributionally represent a morphological process. Huyghe and Wauquier employ the centroid vector of prototypical agent nouns to represent and investigate this specific derivational pattern. Bonami and Paperno instead compare multiple morphological processes to test distributionally the difference between inflection and derivation. They show that the variance of vector offsets between derivational forms and their corresponding bases is greater than the variance of vector offsets between inflectional forms and their bases. The variance is a measure of what they call *stability of morphosyntactic and semantic contrast*, a facet of semantic regularity. Given their results, they conclude that inflectionally related words differ from each other in a more regular way than derivational ones.

Both the direct comparison of base/derived words and the compositional approach can be used to investigate morphological processes. Reddy et al. (2011) compared the two in modelling human transparency ratings for compound nouns and found only a small improvement with a compositional approach.

In our study, we opt for the direct comparison of nominalizations with their bases because our focus is on the transparency relation holding (or not holding) between the derivative and the base verb. We leave the investigation of the same processes from a compositional point of view as a matter of future work, since we consider this angle to primarily address a question of learnability and predictability, and not one of transparency.

3.3 Distributional measures of transparency

As discussed in Sect. 2, transparency is the core theoretical notion of our work: we are interested in comparing `UNG` and `NI` nominalizations in terms of their transparency to the base terms or, in opposite terms, in quantifying how big the semantic drift introduced by the nominalization process is. In this section, we outline and motivate the two distributional measures of transparency employed in this study.

The first measure of transparency, which has already been employed in the literature, is **cosine similarity**. It is the standard measure in distributional semantics and has proven to be the most robust way to detect meaning relatedness in distributional vectors. Cosine similarity measures context overlap among vectors of words, and it is commonly assumed that higher values correspond to higher semantic similarity or relatedness. We expect cosine similarity to be a possible measure of semantic transparency, given that the notion of transparency is similar to the one of semantic similarity. Specifically, we measure the cosine similarity between the base verb and the corresponding derived term: higher cosine similarity values will indicate higher semantic transparency of the derived word, since its meaning will be shown to be similar to that of the base.

Cosine similarity quantifies the overlap between the contexts in which two words (in our case, the base verb and the nominalization) are used by measuring the cosine of the angle between the two vectors. Values range from 1 for very similar vectors, over 0 for orthogonal vectors, to -1 for opposite vectors. However, if a PPMI transformation has been applied, the values of cosine will range between 0 and 1, since there will not be negative values.

Cosine is based on the dot product operator (also called inner product): the dot product of two vectors will be high when they have large values in the same dimensions; on the contrary, if two vectors have zeros in different dimensions, they will have a dot product of 0, which represents their strong dissimilarity. The cosine similarity measure is the normalized version of the dot product, i.e. it normalizes the value for the vector length. More frequent words have longer vectors, i.e. less zeros and higher values, since they occur more often and then have higher probability to occur with an higher number of different contexts. The raw dot product will be, thus, consequently higher for more frequent words. To overcome this problem, the cosine measure divides the dot product by the lengths of each of the two vectors. The vector length is defined as the squared root of the sum of the quadratic values of each dimension:

$$|\mathbf{t}| = \sqrt{\sum_{i=1}^{N} t_i^2}$$

Consequently, the cosine of vector t and c is computed as follows:

$$\cos(\mathbf{t}, \mathbf{c}) = \frac{\mathbf{t} \cdot \mathbf{c}}{|\mathbf{t}||\mathbf{c}|} = \frac{\sum\limits_{i=1}^{N} t_i c_i}{\sqrt{\sum\limits_{i=1}^{N} t_i^2}\sqrt{\sum\limits_{i=1}^{N} c_i^2}}$$

Cosine is a symmetric measure: $cos(x, y) = cos(y, x)$ for all words x, y. However, frequently a word x can be similar to y, but word y can have other closer words. This is usually the case with hypernyms - hyponyms (Lenci & Benotto, 2012): a hyponym like *lion* occurs in many of the contexts of its hypernym *animal*; on the other hand, *animal* has a wider meaning and it will have a number of contexts which will not be shared by *lion*. The hyponym will be closer to its hypernym, but the vice versa will not be true in the same measure.

Given these facts, we consider **distributional inclusion** (Geffet & Dagan, 2005; Kotlerman et al., 2010) as a second similarity measure. It has been devised to model the relation of nominal hypernymy: distributional inclusion quantifies the extent to which the derived word is used only in a subset of the contexts of the base word. Theoretically, it captures a more fine-grained, asymmetric notion of transparency which comes with an additional nuance of *specificity* which is completely lacking in the cosine measure. It counts how much the vector of a word u is included in the vector of its hypernym v, i.e. in which proportion its contexts are the same of v, but it then takes into consideration how much of the vector of v is not included in u.[18] We expect that a fully transparent derivation should behave like a hyponym: it should exhibit the same range (or a part) of meanings of the base, without acquiring other additional or idiosyncratic meanings. A less transparent derivation (i.e., one that acquires novel meanings with respect to the base) will be less included because the values on some of its dimensions (those corresponding to the novel meaning) will be higher compared to the base.

An entire family of hypernymy-tailored similarity measures is based on distributional inclusion (Weeds et al., 2004; Clarke, 2009; Lenci & Benotto, 2012). We conducted some preliminary experiments and selected the most promising one (both in the state-of-the-art and in our own pilot study), InvCL (Lenci & Benotto, 2012). InvCL is derived from the *ClarkeDE* measure[19] (Clarke, 2009), which computes the degree of inclusion of word x into word y:

$$\text{ClarkeDE}(u, v) = \frac{\sum_{f \in F_u \cap F_v} \min(w_u(f), w_v(f))}{\sum_{f \in F_u} w_u(f)}$$

Note that, where cosine has a dot product in the numerator, which mathematically implements context overlap by symmetric pairwise multiplication, here we have a minimum function which is symmetric (i.e., returns the same value for animal/cat and cat/animal) as well and effectively defines a subset of the overlap. It is the denominator which introduces the asymmetry element, taking the "perspective" of just one of the words (of which it calculates the L1-length, and not the euclidean length as in cosine), namely the candidate for being included.

Given this formula, InvCL is computed as follows:

$$\text{InvCL}(u, v) = \sqrt{\text{ClarkeDE}(u, v) \cdot (1 - \text{ClarkeDE}(v, u))}$$

In order to exemplify how this measure works, let's consider again a toy model (Table 3) for the words *cat* and *animal*.

When computing the inclusion of the vector for *cat* into the vector for *animal*, we first consider for each dimension the lowest value between the two vectors; we then divide this sum by the sum of the values for *cat*:

$$\text{ClarkeDE}(cat, animal) = \frac{25 + 10 + 0 + 1 + 22 + 10 + 3 + 0}{25 + 10 + 1 + 2 + 31 + 24 + 3 + 0} = \frac{71}{96} = 0.74$$

[18] Distributional inclusion is also called *feature inclusion*, since it considers the degree of inclusion for each context feature.

[19] The ClarkeDE measure was in its turn based on the measure proposed by Weeds and Weir (2003) and Weeds et al. (2004).

Table 3 Toy distributional vectors for *cat* and *animal*

	dog	worm	philosophical	Latin	love	milk	frog	fly	...
cat	25	10	1	2	31	24	3	0	...
animal	36	40	0	1	22	10	37	70	

We do the same for the ClarkeDe measure of *animal* into *cat*:

$$\text{ClarkeDE}(animal, cat) = \frac{25+10+0+1+22+10+3+0}{36+40+0+1+22+10+37+70} = \frac{71}{217} = 0.33$$

Given these values, we can compute the distributional inclusion of *cat* into *animal*:

$$\begin{aligned}&\text{InvCL}(cat, animal)\\&\quad= \sqrt{\text{ClarkeDE}(cat, animal) \cdot (1 - \text{ClarkeDE}(animal, cat))}\\&\quad= \sqrt{0, 0.74 \cdot (1 - 0.33)} = 0.70\end{aligned}$$

Inclusion values from the InvCL measure range between 0 and 1 (as well as ClarkeDe values), where 1 indicates perfectly included vectors, i.e. identical vectors.

InvCL can describe asymmetrical relations where $r(u, v)$ holds, but $r(v, u)$ does not. In the original context of hypernym detection, the authors thought that the features of a hyponym should be included in the features of the broader term, i.e. the hypernym. Thus, the distributional inclusion of the hyponym will be higher than the inclusion of the hypernym. This is confirmed by the toy example above: if we compute the inclusion of *animal* into *cat*, we will see that it is lower than the inverse:[20]

$$\begin{aligned}&\text{InvCL}(animal, cat)\\&\quad= \sqrt{\text{ClarkeDE}(animal, cat) \cdot (1 - \text{ClarkeDE}(cat, animal))}\\&\quad= \sqrt{0.33 \cdot (1 - 0, 74)} = 0, 29\end{aligned}$$

Matching the semantic properties of our transparency measures with the theoretical considerations concerning `UNG` and `NI` discussed in Sect. 2 enables us to formulate the following experimental predictions and questions:

- Cosine similarity quantifies to what extent the nominalization is a viable distributional substitute for the base verb, and their overall semantic relatedness. We can hypothesize that NIs will display a larger overlap with their corresponding base verbs since they keep the base event reading more consistently.
- Distributional inclusion quantifies to what extent the nominalization behaves like a hyponym of the base verb. Again, we expect that `NIs`, which are often used in very specific contexts with respect to the base verb (*Das Laufen über die Brücke ist verboten, 'the crossing of the bridge is prohibited'*) and which have a lower degree

[20] Remember that these are only fake toy vectors, whose values are not taken from real corpora and were used for demonstration only. Lenci and Benotto (2012) provide support for these claims.

of polysemy than UNGs, will be more included in their bases than UNG. We then expect the values of *InvCL*(*NI, Base*) to be higher than *InvCL*(*UNG, Base*).

Instead of establishing a competition between cosine and distributional inclusion as measures of semantic transparency, we will investigate their different import as if they were describing two aspects of the notion of transparency. InvCL tells us that *Das Laufen* is used in a very restricted subset of the contexts of *laufen*. Cosine similarity, instead, gives us a more general picture of the semantic relatedness of the two words, without further specification of the kind of relation.

4 Experiments

This section presents the results of our experiments. Section 4.1 spells out the experimental setup: we introduce the dataset, provide the details of our distributional semantic model and define our statistical methodology. Our main experiment, comparing NI and UNG, is presented in Sect. 4.2. Section 4.3 reports on auxiliary experiments that locate NI and UNG with respect to more general 'milestones' for inflection and derivation.

In the article text, we concentrate on our main findings. Additional analyses and details are made available in a series of online Appendices available at https://osf.io/x6ctj/?view_only=96578ccc91cb4f4087fb442790901347.

4.1 Experimental setup

Dataset The experimental items for the dataset are sampled from DErivBase, a large-coverage database of German derivational morphology (Zeller et al., 2013). We extract verbs for which both an *–ung* nominalization and a nominal infinitive are attested.[21] This results in 770 tuples. We considered paired nominalizations to focus on cases where there are no other factors blocking their formation. In cases where only one nominalization is formed, other features may have a role, and these could be linked mainly to the base verb rather than to the nominalization process. On the contrary, when both are attested we can reliably observe the distributional differences between them - independently of other semantic constraints which may have acted as a blocking factor when only one nominalization is attested.

We manually inspected the top 5% most frequent items in both categories to remove lexicalized derivatives. In this study, indeed, we were interested in modeling the prototypical items of the two derivational processes. Usually, very frequent derivatives acquire more lexicalized meanings (Bybee, 1985; Boleda, 2020), which are not in line with the compositional meaning brought by the suffix. For example, the word *Speisen* is mainly used in its lexicalized sense of 'food', rather than as a nominal infinitive meaning 'dining'. We then removed 6 cases of NI and 15 cases of UNG.[22]

[21] Specifically, we extracted all the cases in which both rules dVN09 (for NIs) and dVN07 (for UNGs) were attested. We preferred DErivBase to other derivational lexicons such as Celex (Baayen et al., 1995) because it covers also cases of conversion, like the nominal infinitives in the focus of this paper.

[22] The NIs removed are: *Falten* 'wrinkles', *Gestalten* 'figures', *Speisen* 'foods', *Stärken* 'strengths', *Strahlen* 'rays', *Streben* 'aspiration'. We removed also the following UNG derivatives: *Anforderung* 're-

We then applied frequency filtering, retaining only cases where all the forms have a frequency higher than 50, in order to ensure the reliability of the vectors. Indeed, it has been frequently stated (e.g. Bullinaria & Levy, 2007; Luong et al., 2013; Schnabel et al., 2015; Gong et al., 2018) that low frequency target words have less reliable vectors; notably, Drozd et al. (2015) showed that with frequency increasing from 5 to 100, accuracy in their task improved rapidly. Frequency information for the target words is taken from the same corpus used to build the distributional model, i.e. the 700 million word SdeWaC corpus (Faaß & Eckart, 2013), POS tagged and lemmatized using TreeTagger (Schmid, 1994). The frequency distributions of the three categories before filtering are reported in the online Appendix B, Table 9, available at the URL mentioned at the beginning of this Section. Half of the NIs (381) have a frequency lower than 102 (the median of the category); UNGs have instead a higher median value and only 77 items have a frequency lower than 102. This difference in frequency distribution of UNG vs. NI motivates our decision to establish a frequency filter, as the negative impact of low frequency on the quality of the vectors would have affected mainly the NI vectors.

The end result is a dataset containing 370 triplets of NIs and UNGs, plus their corresponding base verbs. Table 10 in the online Appendix B reports the frequency distributions of the filtered dataset, whereas the complete list of items is provided in online Appendix A.

Distributional model For each word in the dataset, we extracted distributional count vectors from the SdeWaC corpus (Faaß & Eckart, 2013), adopting default settings for this class of DSMs: a symmetric 5-words context window, PPMI as a scoring function, and the lemmas of the 10 thousands most frequent content words (nouns, adjectives, verbs, and adverbs) as dimensions.[23]

Task and regression analysis We investigate the usefulness of the distributional measures to distinguish between `NI` and `UNG` by setting up a binary classification task: given a set of features that describe a pair of a base verb and a derived noun, we train a logistic regression model to predict whether the derived word is a nominal infinitive or an *ung* nominalization of the base verb.[24] The rationale for this setup is that we can interpret the coefficients of each feature as a quantification of its importance for the `NI - UNG` distinction. We use the following features:

- log-transformed **relative frequency** (ratio between the frequency of the derived word and the frequency of the base word);
- log-transformed **frequency** of the derived word;
- **cosine similarity** between the derived word and its base;

quirement', *Anwendung* 'application', *Bestimmung* 'determination', *Beziehung* 'relation(ship)', *Bildung* 'education', *Einführung* 'introduction', *Erfahrung* 'experience', *Herausforderung* 'challenge', *Lösung* 'solution', *Ordnung* 'order', *Prüfung* 'exam', *Regierung* 'government', *Verfügung* 'disposal', *Verfassung* 'constitution', *Wohnung* 'apartment'.

[23] We do not perform dimensionality reduction, since InvCL is only well-defined on vectors from $\mathbb{R}^+$, but dimensionality reduction methods typically map into $\mathbb{R}$.

[24] We implement the logistic regression model with the `glm` function from `R`, with family type equal to binomial.

Table 4 Model selection procedure

- Step 1: frequency predictors
  - `Model 1.a: Category ~ Relative frequency`
  - `Model 1.b: Category ~ Absolute derived word frequency`
  - `Model 2: Category ~ Rel. freq. + Derived word freq.`
- Step 2: distributional predictors
  - `Model 3.a: Category ~ Significant predictors from step 1 + Cosine`
  - `Model 3.b: Category ~ Sign. predictors from step 1 + InvCL`
  - `Model 4: Category ~ Sign. predictors from step 1 + Cosine + InvCL`
- Step 3: Interactions
  - `Model 5: Category ~ (Sign. freq. predictors from step 1) * (Sign. distributional predictors from step 2)`

- **distributional inclusion** between the derived word as included term and the base as including word.

Frequency distributions for our predictors are reported in the online Appendix B, Table 11. All the numerical variables are scaled and centered on their mean. Note that the identity of the base verb cannot be recovered from the features, so the model cannot succeed by learning base verb-specific lexicalization patterns.

We apply a standard feature selection procedure to construct our model, namely forward selection,[25] adopting a theory-motivated order. The method is summarized in Table 4.

We start by testing the significance of the frequency predictors (step 1) since they have been previously indicated in the literature as correlates of semantic transparency (Bybee, 1985; Hay, 2001). We first add them in separate models (`1.a` and `1.b`), and we check if the addition of the second frequency variable (model `2`) significantly improves the model fit, based on the result of likelihood ratio tests perfomed with the Anova function. The more complex model (model `2` in this step) is retained only if it improves both of the previous models. For example, if model 2 is significantly better than model `1.b` but not better than model `1.a`, it means that the addition of the derivative frequency does not improve the fit and thus model `1.a` will be the selected model for this step.

We then proceed with the two distributional measures: cosine, as it encodes a rather unmarked (and symmetric) notion of similarity, and lastly the (asymmetric) inclusion measure. We add them one at a time to the best model resulting from step 1, and (as done for frequency variables) if both are significant we test if the model is improved by adding both of them (model `4`). Lastly, we test if any interactions[26] among the significant frequency and distributional variables selected from previous steps improve the model fit (step 3, model `5`).

[25] We also tested with backward elimination of the variables, i.e. starting from a model with all the variables and their interactions and then removing one at a time the variables whose loss does not worsen the model fit. The best and final model however is the same as with the forward selection.

[26] Following the R syntax, we mark interactions with an asterisk between the names of the interacting variables (e.g. `Rel.freq. * InvCL`).

Once the best model has been identified with forward selection, we will interpret its adjusted R^2 as an indication of how good of a fit the regression model achieves with respect to the target classification – i.e. the NI vs. UNG distinction. In the model summary, the significance of a predictor[27] (its p-value) indicates if it contributes to the targeted distinction, i.e. if it shows a difference between the two morphological patterns that is not due to chance; the sign of the corresponding estimate indicates the direction of the effect. In the logistic regression the category `NI` corresponds to 0, and `UNG` to 1. Therefore, for example, a significant positive estimate for relative frequency in the `NI` vs. `UNG` distinction indicates that `UNGs` have a higher relative frequency than `NIs`.

4.2 Main results: NI vs UNG

Before starting the regression analysis, we inspect the variables in order to detect possible problems of collinearity among them. Correlated variables may lead to severe problems in regression models (e.g. Baayen, 2008:6.2.2; Tomaschek et al., 2018), such as suppression effects or unexpected and theoretically uninterpretable parameter estimates. To assess collinearity, we use multiple means: we first inspect the correlation matrix of the variables considered (reported in the online Appendix C); second, we compute the condition number *k*;[28] lastly, we check the vif values for the regression models. We consider as warnings correlation values higher than 0.7, *k* and vif values respectively higher than 6 and 3. In those cases in which values higher than these thresholds signal possible collinearity problems, we inspect our models and their results further.

With regard to the `NI-UNG` comparison, we observe a high correlation (r=0.84) between the frequency of the derived term and the cosine value (despite the low *k* of 5.65). It has been previously observed that the cosine similarity measure is influenced by the frequency of the compared terms (e.g. Weeds et al., 2004). In the case of our study, the two nominalization patterns largely differ in terms of frequency (with `UNG` nouns much more frequent than `NIs`), and the cosine measure may be affected by this difference. This is an important drawback of this measure. The inclusion measure on the other hand does not suffer of this disadvantage and seems more stable and independent of frequency.[29]

We now illustrate the model selection procedure described above by constructing the best model for the `NI` vs. `UNG` distinction (process presented in Table 21, online Appendix D).

In the first step, we start adding frequency predictors (`m1.a` and `m1.b`). Both of them significantly improve the model, as shown by a likelihood ratio test (LRT)

[27] In the whole study, we adopt a conservative confidence level of $\alpha = 0.01$ due to the number of comparisons we carry out.

[28] We use the function `collin.fnc()` to compute the condition number.

[29] In order to avoid the collinearity between cosine and the derived frequency, we experimented with different samples of `NIs` and `UNGs`, trying to flatten their difference in frequency. However, in all these experiments the two variables always turned out to be highly correlated. The two frequency distributions are so different (in our original sample `NIs` have a median frequency value of 190, whereas the first quantile of `UNGs` has already a value of 824) that the required subsampling would have yelded an unacceptably small and not-representative dataset.

Table 5 Summary of the final model for the NI vs UNG distinction

Predictor	Estimate	Std. Error	z-value	Significance
(Intercept)	-0.072	0.119	-0.606	$p > 0.05$ n.s.
relfq	1.901	0.204	9.337	$p < 0.001$ ***
InvCL	-0.195	0.117	-1.676	$p > 0.05$ n.s.
logfqder	0.995	0.159	6.262	$p < 0.001$ ***
relfq∗InvCL	0.544	0.128	4.253	$p < 0.001$ ***

among `m2` and both `m1.a` and `m1.b`. The significance levels of the comparison to previous models (reported in the last column)[30] indicate that the newly introduced predictors contribute to the distinction between the two target categories. Adding cosine as predictor (`m3.a`) does not improve the model fit: the LRT reveals that the difference with `m2` is not significant. This result may be due to the correlation between cosine and the derived's frequency highlighted above;[31] however, we cannot remove the derived term's frequency from predictors in favour of Cosine, because of the very different frequency of the patterns under investigation: frequency is a natural covariate we want our distributional variables to build upon. We observed, moreover, that the sign of the estimates for cosine turns from negative to positive if the derived term's frequency is removed from the model. This fact is due to the above mentioned correlation and makes us mistrust a model in which cosine is preserved instead of derived frequency.[32]

Given this fact, we do not keep cosine as a variable in the following models. In contrast, the addition of the inclusion variable, InvCL, improved it (m3.b).

The interaction between relative frequency and the inclusion variable significantly improves the fit; model `4.a` is our final model and the adjusted R^2 value shows that this distinction can be predicted rather well thanks to our corpus-based variables.[33] The summary of the model fit, reported in Table 5, confirms that the frequency predictors are highly significant, whereas our inclusion variable alone is not signficant ($0.05 < p < 0.1$), but highly significant if moderated by relative frequency.

Let's look at the direction of the effect of each predictor. As expected, the model picks up the fact that `UNG` nouns are significantly more frequent with respect to `NIs`. Similarly, their relative frequency is higher than that of `NIs`, i.e. `UNGs` are generally more frequent than their bases with respect to `NIs`. If we take relative frequency as

[30] We highlight p values lower than 0.001 with ***, and values lower than 0.01 with **. If the new model is not significantly better than the previous one, we indicate it with the abbreviation *n.s.* (not significant).

[31] Note, however, that all vif values are lower than 3.

[32] We are aware of potential concerns regarding the simultaneous presence of derived and relative frequency in the model and their log transformations (Tomaschek et al., 2021:10-11). However, we find that the model is stable despite the presence of the two predictors, and most importantly, without collinearity issues.

[33] We check for collinearity, using the `vif()` function from the `car` package. The results indicate that the final model does not suffer from collinearity issues (vif values: derivative's frequency = 1.20; inclusion = 1.38; relative frequency = 1.90; inclusion ∗ relative frequency = 1.64).

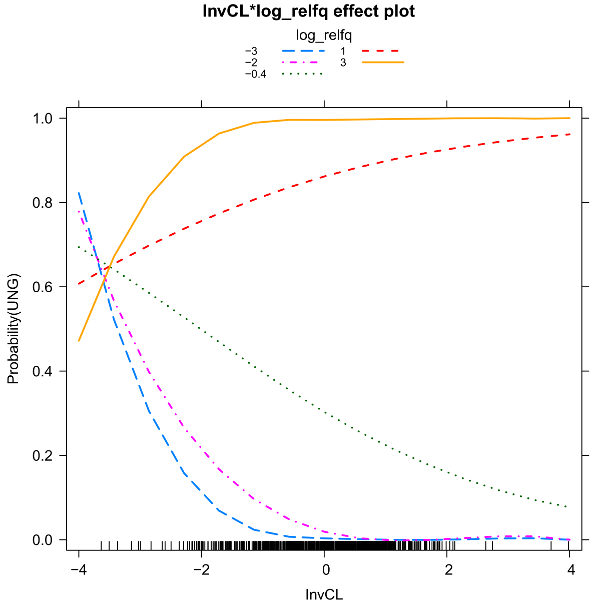

Fig. 1 Effect plot for the interaction between relative frequency and inclusion in the NI-UNG comparison (Color figure online)

a correlate of semantic transparency (following statements made by Hay, 2001, for example), we would interpret UNGs to be less transparent than NIs.

Distributional inclusion shows a significant effect only when considered in interaction with relative frequency. Figure 1 shows that NIs vectors are more included in the vectors of the base verbs than those of UNG nominalizations only when their relative frequency is low (longdash blue and dotdash purple lines in the plot). When the relative frequency is higher (solid orange line) we find the opposite trend: UNGs show higher inclusion values than NIs. The effect of InvCL is absent for derivatives with relative frequency around 0 and 1 (i.e. those with frequency equal to their bases, dotted green and dashed red lines).

4.2.1 Discussion

The interaction between relative frequency and inclusion measure brings new insights and partially contradicts our expectations about frequency effects on transparency. Given these data, we can argue that higher relative frequency does not always imply a semantic shift and, vice versa, a lower relative frequency is not always associated with semantic transparency. A more complex view emerges, in which morphological patterns show specific trends. In particular, the two nominalizations exhibit opposite behaviour once related to their relative frequency. Low-frequency NIs actually preserve the meaning of their corresponding base verb. In those cases in which their relative frequency is higher, they acquire a meaning shift that brings their vectors away from those of their base verb. Consider for example the NIs *Verschulden* (‘fault, blame’) and *Bestreben* (‘ambition, aspiration’), whose meanings are not totally transparent with respect to those of the base verbs *verschulden* (‘to get into debt, to cause something’) and *bestreben* (‘to endeavor’). The behavior of NIs thus matches previous observations concerning relative frequency and transparency.

UNG nouns behave in the opposite way: when their frequency is lower than those of the base verb, they are less transparent. When they are frequent, they are actually similar to their base verbs in meaning; a behaviour that is in contrast with previous observations concerning (relative) frequency and lexical split (a.o. Bybee, 1985; Hay, 2001).

In order to better understand the general results of the model, we manually inspect the items that were correctly and wrongly predicted by the model. Starting from the most confidently predicted UNGs and NIs (i.e. those with a probability respectively higher than 0.9 and lower than 0.1), we notice that in most cases there is a sortal distinction between the two nominalizations. The model has done a better job at predicting cases in which the UNG derivative has a predominant result reading (either a result object or resulting state reading), that clearly differs from the event reading of the NI. Some examples extracted among the top 20 best predicted items are:

- *Abbilden* ('the depicting, representing') vs. *Abbildung* ('reproduction, image');
- *Äußern* ('the uttering') vs. *Äußerung* ('statement, comment');
- *Beleuchten* ('the illuminating') vs. *Beleuchtung* ('light');
- *Beschreiben* ('the describing') vs. *Beschreibung* ('description, report');
- *Einsparen* ('the economizing') vs. *Einsparung* ('savings').
- *Absichern* ('the safeguarding') vs. *Absicherung* ('safeguard');
- *Abstimmen* ('the voting, tuning') vs. *Abstimmung* ('harmonization');
- *Bewundern* ('the admiring') vs. *Bewunderung* ('admiration');
- *Darstellen* ('the representing') vs. *Darstellung* ('representation');
- *Einschätzen* ('the estimating') vs. *Einschätzung* ('assessment');
- *Einschränken* ('the restricting') vs. *Einschränkung* ('limitation')

In a few cases, the two nominalizations seem to disambiguate a polysemous base verb by inheriting only part of its senses. Consider as examples the verbs *annähern* and *ausstrahlen*. *Annähern* has two main senses, 'to approach' and 'to approximate', and each nominalization tends to express only one of these two meanings: the NI *Annähern* mostly means 'approximating', whereas the 'approach' meaning is arguably more dominant for the UNG *Annäherung*. Similarly, from the verb *ausstrahlen* ('to emanate, to radiate'), the UNG noun *Ausstrahlung* dominantly inherits the metaphorical meaning of 'charisma', whereas the NI *Ausstrahlen* is usually used for 'radiating'.

There are, however, cases correctly predicted by the model that do not show a clear semantic difference between UNG and NI. For example the following pairs do not exhibit a sortal difference or other meaning differences: *Aktivieren* vs. *Aktivierung* ('activation'); *Beschaffen* vs. *Beschaffung* ('acquisition'); *Bombardieren* vs. *Bombardierung* ('bombing'); *Durchsetzen* vs. *Durchsetzung* ('assertion, implementation'); *Einbeziehen* vs. *Einbeziehung* ('inclusion').

If we look at the errors of the model, it is more difficult to find patterns. Among the pairs that were wrongly predicted (NI's probability higher than 0.8 and UNG's probability lower than 0.2), there are cases in which the UNG noun has acquired an idiosyncratic meaning which does not correspond to the act or result of the verb, as in the following cases:

- *Heizen* ('the act of heating') vs. *Heizung* ('heater');
- *Lesen* ('the act of reading') vs. *Lesung* ('an author reading event');
- *Schreiben* ('the act of writing') vs. *Schreibung* ('orthography').

4.3 Further comparison: NI vs. UNG in the inflection/derivation cline

The two nominalization patterns above show some peculiar characteristics: they have similar meanings since they can both be used to refer to events, but they largely differ in terms of frequency. The previous experiment highlightes these properties, but it also shows that there is a fine-grained difference in terms of semantic transparency between the two processes.

In this section, we want to apply our distributional predictors to the comparison with two other morphological processes, namely *-er* deverbal nouns (ERs) and present participles in *-end* (ENDs).

Our aim is twofold. First, we want to verify the reliability of our transparency measures outside the event nominalizations domain. For this reason, we chose a derivational process (ER) that produces a type-shift from event to agent or instrument nouns, and an inflectional process (ENDs) that only maintains the event reading and that does not imply a conversion to nouns.[34] In this way we check that our transparency measures are not simply measuring the relatedness to the base verb (an agent noun is still related to the verb it is formed by), nor are they influenced by the syntactic status of nouns.

Second, we want to relate our results to the general distinction between inflection and derivation. In Sect. 2.1 we have seen that NIs maintain some features of their inflectional origin, since they do not show all the nominal properties. UNG nouns instead are a more typical case of derivation, since they acquired through time all the syntactic and morphological features of typical common nouns. This different origin is mirrored in their different degrees of semantic transparency: as reported above, it has been frequently stated that derived words often acquire idiosyncratic meanings and non-unique form-meaning associations (Booij, 2000; Laca, 2001), whereas inflection can be completely compositional (Dressler, 2005:271). Given these premises, we expect our measures to capture the difference in semantic transparency between inflectional and derivational processes, showing a graded distinction between the two sides. We acknoweldge that other additional processes are needed to fully investigate transparency into the inflection-derivation continuum, but this is out of the scope of our paper. We refer the reader to Bonami and Paperno (2018) for a similar analysis.

In Sect. 4.3.1, we investigate the modulation of our transparency measures in connection with type-shift introducing *-er* deverbal nouns into the picture. Second, in Sect. 4.3.2 we introduce an inflectional comparison term, the present participles in *-end*.

4.3.1 A derivational landmark: ER deverbal nouns

In German, the suffix *-er* is used to create deverbal nouns that denote agents (*Läufer*, ‘runner’) or instruments (*Öffner*, ‘opener’) (Plag, 1998). In a few cases, they may also indicate patient nouns (e.g. *Aufkleber* ‘sticker’, from *aufkleben* ‘to stick’) or semelfactive events (*Seufzer* ‘sigh’) (Luschützky & Rainer, 2011; Scherer, 2011;

[34] Moreover, contrary to other inflectional processes, END participles are lemmatized, thus directly comparable to other lemmas in our distributional space.

Müller, 2011) – however, the event reading is marginal for ER, particularly in comparison to UNG nominals. In the context of our experiments, we take ER as a derivational landmark in which the event reading of the base verb is commonly lost: we thus expect ERs to be less included in the base verb's vector than UNGs and NIs.

From the dataset used in 4.2, we extracted for each base verb the corresponding -er nominals from DerivBase (when present), thus restricting the list to 76 quadruples. We report descriptive statistics for the variables considered in online Appendix B.

We then repeated the experimental setup from Sect. 4.2 twice, once using the NI/ER distinction and once with the UNG/ER distinction as the dependent variable (ER being coded as 0). From the correlation matrices (online Appendix C, Table 16 and 17), we observe that there may be some collinearity issues between Cosine and InvCL, since their correlation coefficient is equal (for NI-ER) or above (for UNG-ER) the threeshold of 0.7, and between InvCL and relative frequency (r=0.72) for the UNG-ER comparison.[35]

Considering these values, we proceed with caution analyzing the results of our regression models. We refer the reader to the online Appendix D (Tables 22 and 23) for the detailed model selection procedure for NI/ER and UNG/ER, respectively. For the NI-ER prediction task, the possible correlation between our distributional variables is not relevant, since the addition of Cosine (model m3.a, Table 22 in online Appendix D) yields only a marginally significant improvement ($p > 0.01$). The final model uses relative frequency and InvCL as predictors (model m3.b). With regard to the UNG-ER comparison, the frequency variables are not significant in the final model, and thus the correlation with InvCL is not an issue. However, the correlation between our distributional measures may be problematic, causing unreliable model. We thus analyze their import in separate models, without considering them together in a single one.

In both comparisons, the frequency of the derived term had no impact on the improvement of model accuracy, whereas relative frequency was significant for the ER-NI comparison. InvCL significantly improved the models, whereas Cosine played a role only in the ER-UNG comparison. Interactions among the variables were discarded because they did not improve the fit. We thus end with relative frequency and InvCL as predictors for the ER-NI distinction (model m3.b), and two distinct models with our distributional variables for the ER-UNG one (model m1.c and m1.d). A collinearity check confirms that the effects of the model can be trusted.[36] The estimates for the best models are reported in Table 6.

The adjusted R^2 values (0.44 and 0.38) are lower than in the NI vs. UNG experiments (R^2=0.63), showing that distinguishing the two nominalizations in the core of this paper from ER is less straightforward. This observation indicates that the common event semantics of UNGs and NIs does not make the task more difficult. Our inclusion variables help indeed in discriminating the two event nominalizations.

In the comparisons with ER, the role of relative frequency in the model fit is weaker: the adjusted R^2 for the models with only the frequency variables are much

[35] The *k* condition number does not suggest any collinearity problem (4.38 for the UNG-ER prediction, 3.91 for the NI-ER one.)

[36] VIF values for ER vs. NI: relative frequency 1.93 ; inclusion, 1.93.

Table 6 Final models for the ER-NI (m4) and ER-UNG (m1.c and m1.d) comparisons

ER vs. NI ($R^2 = 0.44$)			ER vs. UNG (m1c: $R^2 = 0.38$; m1d: $R^2 = 0.34$)		
Predictor	Estimates		Predictor	Estimates	
Relative freq.	-1.859	$p < 0.001$ ***	m1c: Cosine	1.334	$p < 0.001$ ***
InvCL	1.720	$p < 0.001$ ***	m1d: InvCL	1.288	$p < 0.001$ ***

lower than those of the best model with distributional measures (0.13 vs 0.44 for the ER-NI comparison; see online Appendix D); for the ER-UNG comparison relative frequency is not significant. This suggests that relative frequency is not enough to explain the difference among derivational patterns and to measure semantic transparency, whereas our distributional measures are better at catching it.

Even if the model fit is lower, our distributional variables are significant in the distinction of ERs from the two event nominalizations. Let us look at the estimates of the logistic regression model displayed in Table 6.

Looking at the effect of the distributional inclusion measure, ER is significantly less included than NI and UNG nouns. Given their event semantics, we expected the latter to have a stronger semantic relationship with the base verb. It suggests that the InvCL measure grasps the strong event semantics of this class of nominalizations, rather than a simple relation in meaning with the base verb semantics (as could be in the case of agent nouns).

The same is not true for the Cosine measure: it is not significant in the ER-NI task, despite the semantic difference between the two patterns. In the ER-UNG comparison instead, cosine is significant and it also explains a larger variance than the inclusion measure (R^2=0.38 vs 0.34). If we look at cosine as a measure of substitutability, UNG is confirmed as a better candidate substitute for the base verb than ER. However, the results from the ER-NI comparison do not allow us to fully trust a similar interpretation. Therefore, we can make the following considerations: a) cosine similarity is not a mere correlate of type shift (otherwise NI would have been significantly more similar to the base than ER) and b) it is probably too dependent on frequency.

With regard to relative frequency, its effect is significant only in the comparison between ERs and NIs. The relative frequency of NIs is significantly lower than that of ERs, whereas the latter is not significantly different from that of UNGs. A one-unit increase in relative frequency is associated with the decrease in the log odds of the derivative being an NI vs. an ER nominal in the amount of 1.859. Interestingly, ERs and UNGs, at the derivational extreme of our cline, show the same behaviour with respect to relative frequency.

4.3.2 An inflectional landmark: Present participles

Let us now extend our study to present participles, our inflectional landmark, by introducing in our comparison the present participle END.

Present participles (inflectional verbal adjectives, Haspelmath, 1994) can be formed for every base verb and are mainly used as adjectives: *das kochende Wasser*

Table 7 Summary of the best model for the comparisons END vs UNG, END vs NI, and END vs ER

END vs. UNG (R^2 = 0.62)		END vs. NI (R^2 = 0.04)		END vs. ER	
Predictor	Estimates	Predictor	Estimates	Predictor	Estimates
Relative freq.	1.928 ***	Cosine.	-0.507 **	**m3.b** (R^2 = 0.48):	
Cosine	0.986***			Relative freq.	1.695 ***
InvCL	-0.913 ***			InvCL	-2.018 ***
Rel.fq*InvCL	0.698 ***			**m3.a** (R^2 = 0.23):	
				Relative freq.	1.213 ***
				Cosine	-0.980 ***

(‘the boiling water’), *das laufende Jahr* (‘the current year’). Some of them (like *spannend* ‘gripping’, *umfassend* ‘comprehensive’, *zwingend* ‘mandatory’) have become true adjectives and can also occur after the auxiliary *sein*, ‘to be’ (Durrell, 2003). Given their inflectional nature and the fact that they only retain the event reading, we expect ENDs to pattern with NIs.

We extracted from DerivBase the corresponding END derivatives: we obtained 185 items matching NIs and UNGs and 59 matching the previously selected ERs. We report descriptive statistics for the two dataset in the online Appendix B, Table 13 and 14 respectively. In the online Appendix D (Tables 24, 25 and 26), we report the result of model selections for the three comparisons, obtained following the same procedure described in previous sections.

The low R^2 values for the comparison between NI and END (best model: R^2 = 0.04) show that our predictors are not able to tease apart these two categories: indeed, the inflectional extreme of our cline appears to be quite tight. Instead, END can be better distinguished from UNG (R^2 = 0.62) and (to a lower extent, somewhat surprisingly) from ER (R^2 = 0.48).

In Table 7, we report the effects of our predictors in the best model for the three comparisons.[37] ENDs correspond always to category 0, whereas ERs, NIs or UNGs to 1. For the END-ER comparison, the two distributional variables seem correlated (r=0.73, vif values > 3; see Table 20, online Appendix C). For this reason, we do not add the two predictors in the same model, but we consider them in two separate final models.

InvCL contributes significantly in the distinction between END and UNG, and in the distinction between END and ER. On the contrary, for the END vs NI comparison, only Cosine has a role (but consider that the R squared is really low). Cosine is significant in the other two tasks, even if in the END-ER prediction the model with Cosine achieves a lower R^2 than that with InvCL.

The InvCL measure is not significant in the comparison between END and NI. In the other two tasks, higher values of inclusion are associated with a decrease in the

[37] As for previous tasks, we check for collinearity, confirming that results from the richer models can be trusted. Vif values for
END vs UNG : relative frequency= 1.77 ; cosine= 1.99; invcl= 2.13; invcl*relative frequency = 1.29.
END vs ER: m3.a: Relative frequency = 1.84 , Cosine = 1.84; m3.b: Relative frequency= 2.03, InvCL= 2.03.

log odds for UNGs and ERs, and thus with an increase for ENDs. This result suggests that there is no difference in terms of inclusion between the two inflectional patterns, whereas it is displayed between inflectional and derivational ones.

Considering the case of END vs UNG, the effects of inclusion and cosine similarity go in the opposite direction: END is associated with higher inclusion values, UNG with higher cosine values. On the contrary, in the distinction between END and ER the effects of InvCL and Cosine go in the same direction: END formations are more included and more similar to the base verb than ER nouns. The behaviour of UNG nouns is a *unicum*: like the other derivational process, they are less included into the base verbs than inflectional patterns;[38] however, they are more similar to the base than ENDs when similarity is computed with the Cosine measure.

Relative frequency is not significant in the comparison between END and NI, whereas in the other two tasks higher values are associated with higher log odds for UNG and ER. A higher relative frequency appears to be associated with true derivational processes, while inflectional processes are in general less frequent than their bases. The mixed nature of NIs is not sufficient to associate them to derivational processes, as far as frequency predictors are concerned.

We cannot exclude that, in these last three comparisons, an additional factor is coming into play: the difference of part of speech. While all other tasks involved a V–N derivational pattern, END brings into the picture a V–Adj process.

4.4 General discussion

We have seen so far that in modeling the distinction between two morphological processes our distributional predictors did have a significant role, even if differences due to the processes involved were present. In this Section, we provide a comprehensive view on the findings of our experiments.

First of all, looking at the R^2 values, we can see a difference in the fit for the various regressions. We can order the comparisons as follows, from the highest R^2 value to the lowest one:[39]

$$\mathtt{NI/UNG} > \mathtt{END/UNG} > \mathtt{END/ER} > \mathtt{NI/ER} > \mathtt{UNG/ER} > \mathtt{END/NI}$$

The comparison between NIs and UNGs, which is at the core of this paper, is the task (directly followed by the END/UNG prediction) in which our variables do the best in teasing apart the two categories (R^2=0.63). We take this as an indication that their distributional contexts (considered in relation to those of their base verbs) can lead us to a distinction between the two categories. However, only one of the distributional predictors was adequate: cosine did not contribute in improving the

[38] However, consider that in the comparison with NI, UNG is more included when more frequent than its base verb.

[39] We acknowledge that our tasks have been conducted on datasets with different sizes and that this fact may influence R^2 values. For this reason, we repeated the whole analysis on a dataset in which, for each base verb, all four derivatives are present. Given the resulting dataset (consisting in 51 items for each category), the differences among models fit remain the same. We report here the R^2 values for this smaller dataset: NI-UNG: 0.74; END-UNG: 0.74; END-ER: 0.49; NI-ER: 0.45; UNG-ER: 0.39; END-NI: 0.02. However, even if the order of R^2 values is the same, we will continue in the rest of the paper to refer to values coming from the previous sections.

model, contrary to the distributional inclusion measure. We reached a high R^2 value (R^2=0.62) also for the other comparison between UNGs and the other inflectional process, i.e. END. For the comparison between the two inflectional patterns (END/NI), we reached the lowest R value (R^2=0.04). The two derivational processes (ER/UNG) were better identified (R^2=0.38), even if this value was lower than those of the other four tasks. The fact that the comparisons involving ER are in the middle range of our predictability cline confirms, once again, that our measures of transparency are not just sensitive to type shift: had that been the case, ER (which retains the event reading only for a small minority of cases) would have been the easiest pattern to spot.

As regards frequency predictors (relative frequency and derivative frequency), we note that relative frequency is not significant in the comparison between the two inflectional patterns (END/NI) and the two derivational ones (ER/UNG). This suggests that the two classes of morphological processes have different behaviours with regards to frequency: derivational processes produce more items whose frequency is higher than those of their base verbs, but this is not the case for inflectional ones. Considering also the other comparisons, we can order the effect of relative frequency as follows:

$$\mathtt{UNG,\ ER} > \mathtt{NI,\ END}$$

Higher relative frequency values are highly significantly associated with the increase of the log odds for derivational processes (UNG vs END, UNG vs NI, ER vs END, ER vs NI). Considering these results in the perspective of Hay's claims (2001), we observe that they match our theoretical expectations about semantic transparency. We expected derivational patterns to be less transparent than inflectional ones (included NIs), and consequently, given Hay's predictions, we expected higher relative frequency for derivational patterns.

The relation between relative frequency and transparency is once more supported if we compare these results to the effects of the distributional inclusion measure. Given our predictions, we can argue that the inclusion measure is able to quantify the degree of semantic transparency between two words, specifically between a derived term and its base verb. The effects we observed go in the opposite direction to those of relative frequency, as claimed by Hay: higher relative frequency corresponds to lower transparency, i.e. to lower inclusion.

$$\mathtt{NI,\ END} > \mathtt{UNG} > \mathtt{ER}$$

We found no difference in inclusion in the comparison between END and NI, suggesting that categories belonging to the inflectional side of the cline share the same degree of semantic transparency, and are difficult to be distinguished. By contrast, on the derivational side, UNGs were more included than ERs: given their event semantics, they preserve more the meaning of the base verbs with respect to ER agent nouns.

The absolute frequency of the derived term was significant only in the distinction between UNG and NI. Relative frequency turned out to be more relevant than derivative frequency: Hay's proposal of relative frequency as a better measure of transparency with respect to absolute frequency is thus supported by our empirical results.

However, the case of the UNG/NI comparison can be seen as an exception to the relation between relative frequency and transparency. In this task, relative frequency has a moderator effect on the inclusion variable. NIs with low relative frequency are more included (and thus more transparent) than UNG nouns, as we expected; but with high relative frequency the two patterns behave in the opposite way: UNGs with high relative frequency are more included than NIs. In other words, NIs are more included when less frequent, whereas UNGs are more included when more frequent than their bases. While we expected NIs to be more included, the association between high relative frequency UNGs and higher inclusion surprised us leaving room for further investigation concerning the relation between frequency and semantic transparency.

Consequently, we can hypothesize that there are some morphological processes whose more frequent outputs do not lose semantic transparency, but rather preserve it. We should however be cautious about this final conclusion, since the difference in relative frequency among the two categories was really high and only further investigation could help us in accepting or refuting this claim. Moreover, we can hypothesize that our results depend on the very particular characteristics of the NIs pattern, which exhibits an ambiguous status with respect to the inflection vs. derivation distinction.

The effect of Cosine does not match our expectations and its behaviour is difficult to interpret. It is, indeed, not significant either in the distinction between the two event nominalizations, or in the NI vs ER one. It is interesting, moreover, to note that in the END-UNG comparison, the effect of the Cosine measure goes in the opposite direction of that of the inclusion measure: with respect to ENDs, UNG nouns are more similar to the corresponding base verb, but less included. However, in other comparisons (ER-UNG, END-ER), the effects of InvCL and Cosine converge. Given these inconsistencies, difficult to interpret, and the relation between Cosine and frequency (shown in this study and in previous works) we believe that the InvCL inclusion measure offers a more reliable measurement of semantic transparency. Cosine similarity can be interpreted as a more general measure of semantic relatedness.

Within the larger context of the inflection/derivation distinction, our results support the picture found by Bonami and Paperno (2018): namely, a difference between inflection and derivation in terms of transparency (and in terms of corpus distribution in general), without a clear categorical boundary.

5 Conclusion

Semantic transparency in derivational morphology has been traditionally interpreted in terms of decomposability (or lack of it) and empirically related to relative frequency. In this paper, we have proposed an extension of this analysis with two measures based on distributional semantics. These measures quantify two relevant dimensions in the transparency of nominalizations with respect to their base verbs: the degree of semantic substitutability (quantified by cosine similarity) and the tendency of the derived word to be used in a subset of the contexts of the base (quantified by distributional inclusion).

Our discussion has shown that distributional inclusion contributes a finer-grained semantic characterization of two event nominalizations (nominal infinitives and deverbal nouns in *-ung*), whereas our measure of substitutability (cosine similarity) is not suitable to catch differences in their semantic import, probably because its close relation with frequency.

Moreover, our results show an interaction between distributional inclusion and relative frequency that partially confirms previous claims about the relation between frequency and semantic transparency (e.g. Hay, 2001).We confirmed relative frequency as a better correlate of semantic transparency than absolute frequency. However, derived terms with higher relative frequency are not always less transparent: our results show that very frequent UNG nouns preserve the base verb meaning, thus contradicting the general tendency. Additionally, we compared event nominalizations to two other morphological processes, showing that the four patterns mostly differ in their distributional profiles, and confirming similar distributional characterization of the inflection-derivation continuum (Bonami & Paperno, 2018).

Another contribution of this work is methodological. We successfully applied distributional semantic measures targeted at specific semantic relations to a question from theoretical linguistics, by giving concrete linguistic interpretations to the outcomes of our quantitative analysis. While interpreting our results, however, we also came up against the downside of our methodology, which aggregates all instances of the words in question, namely that it does not provide easy access to more qualitative aspects. While sentence-level analysis was out of the scope of this paper, it is without any doubt the next necessary step in our research on nominalizations.

Acknowledgements We acknowledge the partial support of Deutsche Forschungsgemeinschaft for Sonderforschungsbereich 732, project B9. G. Lapesa further acknowledges the funding by Bundesministerium fur Bildung und Forschung (BMBF) through E-DELIB (Powering up E-DELIBeration: towards AI-supported moderation). We thank the anonymous reviewers and the editor for their constructive criticism, the audiences of SFB colloquiums and the ESSLLI DSALT workshop as well as Ingo Plag, Roberto Zamparelli, and Sascha Alexeyenko for discussions and comments. We are grateful to Cathy-Ann Heitmeier for proofreading. All remaining errors are, of course, our own.

Further notes The study was conceived and designed jointly as presented in Sect. 1. RV assumed the lead on Sect. 2, and GL on Sect. 3. RV and GL jointly carried out the main data analysis for Sect. 4. SP supervised the work, provided competence as a native speaker and regarding the interface between linguistics and distributional analysis. Earlier versions of this study were presented in Varvara et al. (2016) and Varvara (2017). The present submission represents a major revision both in terms of empirical method and in terms of interpretation over the earlier versions.

Funding Note Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement.

References

Alexiadou, A. (2001). *Functional structure in nominals: nominalization and ergativity*. Amsterdam: John Benjamins Publishing.

Alexiadou, A. (2010). Nominalizations: A probe into the architecture of grammar part I: the nominalization puzzle. *Language and Linguistics Compass*, *4*(7), 496–511.

Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge: The MIT Press.

Baayen, H. R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*, *2019*.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (cd-rom). *Linguistic Data Consortium.*

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, *43*(3), 209–226.

Baroni, M., Dinu, G., & Kruszewski, G. (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)* (pp. 238–247). Baltimore, Maryland: Association for Computational Linguistics.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 1–49.

Baroni, M., Murphy, B., Barbu, E., & Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, *34*(2), 222–254.

Bartsch, R. (1986). On aspectual properties of Dutch and German nominalizations. In V. Lo Cascio & C. Vet (Eds.), *Temporal structure in sentence and discourse* (pp. 7–39). Dordrecht: Foris.

Bauer, L. (1983). *English word-formation*. Cambridge: Cambridge university press.

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

Behaghel, O. (1923). *Deutsche Syntax* (Vol. *1*). Heidelberg: Winter.

Bejan, C. (2007). Two types of German psych nominals: -ung vs. -en. In P. Mos (Ed.), *A building with a view. Papers in honor of Alexandra Cornilescu* (pp. 327–337). Editura Universității din București.

Bell, M. J., & Schäfer, M. (2016). Modelling semantic transparency. *Morphology*, *26*(2), 157–199.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, *6*, 213–234.

Bonami, O., & Paperno, D. (2018). Inflection vs. derivation in a distributional vector space. *Lingue E Linguaggio*, *17*(2), 173–196.

Booij, G. (2000). Inflection and derivation. In G. Booij, C. Lehmann, & J. Mugdan (Eds.), *Morphologie / morphology. Ein internationales Handbuch zur Flexion und wortbildung. An international handbook on inflection and word-formation* (Vol. 1, pp. 360–369). Berlin: De Gruyter Mouton.

Borer, H. (2005). *Structuring sense vol. II: The normal course of events*. London: Oxford University Press.

Budanitsky, A., & Hirst, G. (2006). Evaluating Wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, *32*(1), 13–47.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510–526.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins Publishing.

Bybee, J. L. (1995). Diachronic and typological properties of morphology and their implications for representation. In L. B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 225–246). Hillsdale: Erlbaum.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Clarke, D. (2009). Context-theoretic semantics for natural language: an overview. In *Proceedings of the EACL GEMS workshop on GEometrical Models of natural language Semantics* (pp. 112–119). Athens, Greece: Association for Computational Linguistics.

Comrie, B., & Thompson, S. A. (2007). Lexical nominalization. In *Language typology and syntactic description vol. III: Grammatical categories and the lexicon* (pp. 334–381). Cambridge: Cambridge University Press.

Cotterell, R., & Schütze, H. (2018). Joint semantic synthesis and morphological analysis of the derived word. *Transactions of the Association for Computational Linguistics*, *6*, 33–48.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Demske, U. (2002). Nominalization and argument structure in Early New High German. In E. Lang & I. Zimmermann (Eds.), *Nominalisations, Number 27 in ZAS Papers in Linguistics. ZAS*.

Dressler, W. U. (2005). Word-formation in natural morphology. In P. Štekauer & R. Lieber (Eds.), *Handbook of word-formation* (pp. 267–284). Berlin: Springer.

Drozd, A., Gladkova, A., & Matsuoka, S. (2015). Discovering aspectual classes of Russian verbs in untagged large corpora. In *2015 IEEE international conference on data science and data intensive systems* (pp. 61–68). New York: IEEE.

Durrell, M. (2003). *Using German: A guide to contemporary usage*. Cambridge: Cambridge University Press.

Eberle, K., Faaß, G., & Heid, U. (2009). Corpus-based identification and disambiguation of reading indicators for German nominalizations. In *Proceedings of the fifth corpus linguistics conference*, Liverpool.

Ehrich, V., & Rapp, I. (2000). Sortale Bedeutung und Argumentstruktur: ung-Nominalisierungen im Deutschen. *Zeitschrift für Sprachwissenschaft*, *19*(2), 245–303.

Eisenberg, P. (1994). German. In E. König & J. Van Der Auwera (Eds.), *The Germanic languages* (pp. 349–387). London: Routledge.

Esau, H. (1973). *Nominalization and complementation in modern German*. Amsterdam: North Holland.

Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations*. Ph.D. thesis, University of Stuttgart.

Faaß, G., & Eckart, K. (2013). SdeWaC - a corpus of parsable sentences from the web. In *Language processing and knowledge in the Web* (pp. 61–68). Berlin: Springer.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. In *Studies in linguistic analysis* (pp. 1–32). Oxford: The Philological Society.

Gaeta, L. (1998). The inflection vs. derivation dichotomy: The case of German infinitives. In B. Caron (Ed.), *Proceedings of the XVI international congress of linguists*, Elmsford: Pergamon.

Geffet, M., & Dagan, I. (2005, June). The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics* (pp. 107–114), Ann Arbor, Michigan: Association for Computational Linguistics.

Gong, C., He, D., Tan, X., Qin, T., Wang, L., & Liu, T.-Y. (2018). Frage: Frequency-agnostic word representation. In *Proceedings of the 32nd conference on neural information processing systems (NeurIPS 2018)*, Montréal, Canada (pp. 1341–1352).

Göransson, C. E. (1911). *Die doppelpropositionalen Infinitive im Deutschen. Wald. Zachrissons Boktrykeri*.

Grimshaw, J. (1990). *Argument structure*. Cambridge: MIT Press.

Günther, F., & Marelli, M. (2018). Enter sandman: Compound processing and semantic transparency in a compositional perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(10), 1872–1882.

Hamm, F., & Kamp, H. (2009). Ontology and inference: The case of German ung–nominals. *Disambiguation and Reambiguation*, *6*, 1–67.

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

Hartmann, S. (2014). The diachronic change of German nominalization patterns: An increase in prototypicality. In *Selected papers from the 4th UK cognitive linguistics conference* (pp. 52–171). Lancaster: Cognitive Linguistics Association.

Haspelmath, M. (1994). Passive participles across languages. In B. Fox & P. J. Hopper (Eds.), *Voice: Form and function* (pp. 151–178). Amsterdam: John Benjamins Publishing Company.

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, *39*(4), 1041–1070.

Hay, J. (2003). *Causes and consequences of word structure*. London: Routledge.

Heyvaert, L. (2003). *A cognitive-functional approach to nominalization in English*. Berlin: Walter de Gruyter.

Huyghe, R., & Wauquier, M. (2020). What's in an agent? *Morphology*, *30*, 185–218.

Iordăchioaia, G., & Soare, E. (2015). Deverbal nominalization with the down operator. In E. Aboh, J. Schaeffer, & P. Sleeman (Eds.), *Romance languages and linguistic theory 2013*. Amsterdam: John Benjamins.

Keith, J., Westbury, C., & Goldman, J. (2015). Performance impact of stop lists and morphological decomposition on word–word corpus-based semantic space models. *Behavior Research Methods*, *47*(3), 666–684.

Knobloch, C. (2003). Zwischen Satz-Nominalisierung und Nennderivation: -ung-Nomina im Deutschen. *Sprachwissenschaft*, *27*(3), 333–362.

Kolb, P. (2009). Experiments on the difference between semantic similarity and relatedness. In *Proceedings of the 17th Nordic conference of computational linguistics (NODALIDA 2009)* (pp. 81–88).

Koptjevskaja-Tamm, M. (1993). *Nominalizations*. London: Routledge.

Koptjevskaja-Tamm, M. (2006). Nominalizations. *Encyclopedia of language and linguistics*, *2*.

Kotlerman, L., Dagan, I., Szpektor, I., & Zhitomirsky-Geffet, M. (2010). Directional distributional similarity for lexical inference. *Natural Language Engineering*, *16*(4), 359–389.

Kountz, M., Heid, U., & Spranger, K. (2007). Automatic sortal interpretation of German nominalisations with-ung towards using underspecified representations in corpora. In *Proceedings of corpus linguistics 2007*, Birmingham, England, UK.

Laca, B. (2001). Derivation. In M. Haspelmath, E. König, W. Oesterreicher, & W. Raible (Eds.), *Language typology and language universals* (Vol. 2, pp. 1214–1227). Berlin: de Gruyter.

Landauer, T., & Dumais, S. (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.

Lazaridou, A., Marelli, M., Zamparelli, R., & Baroni, M. (2013). Compositionally derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (Volume 1: Long papers)* (pp. 1517–1526), Sofia, Bulgaria: Association for Computational Linguistics.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(1), 1–31.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151–171.

Lenci, A., & Benotto, G. (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of SEM 2012 – the first joint conference on lexical and computational semantics* (pp. 75–79), Montréal, Canada: Association for Computational Linguistics.

Levy, O., & Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning* (pp. 171–180).

Levy, O., & Goldberg, Y. (2014b, 01). Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, *3*, 2177–2185.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225.

Luong, M.-T., Socher, R., & Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the seventeenth conference on computational natural language learning, CoNLL* (pp. 104–113).

Luschützky, H. C., & Rainer, F. (2011). Introduction. *STUF – Language Typology and Universals*, *64*(1), 3–7.

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

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

McQueen, J. M., & Cutler, A. (1998). Morphology in word recognition. In A. Spencer & A. Zwicky (Eds.), *The Handbook of morphology* (pp. 406–427). Oxford: Blackwell.

Melloni, C. (2007). *Polysemy in word formation: the case of deverbal nominals*. Ph.D. thesis, University of Verona.

Melymuka, M., Lapesa, G., Kisselew, M., & Padó, S. (2017). Modeling derivational morphology in Ukrainian. In *Proceedings of IWCS*, Montpellier, France.

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

Müller, P. O. (2011). The polysemy of the German suffix-er: aspects of its origin and development. *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung*, *64*(1), 33–40.

Padó, S., Herbelot, A., Kisselew, M., & Šnajder, J. (2016). Predictability of distributional semantics in derivational word formation. In *Proceedings of COLING 2016* (pp. 1286–1296).

Plag, I. (1998). Morphological haplology in a constraint-based morpho-phonology. In W. Kehrein & R. Wiese (Eds.), *Phonology and morphology of the Germanic languages* (pp. 199–215). Tübingen: Niemeyer.

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

Reddy, S., McCarthy, D., & Manandhar, S. (2011). An empirical study on compositionality in compound nouns. In *Proceedings of 5th international joint conference on natural language processing* (pp. 210–218).

Rossdeutscher, & Kamp, H. (2010). Syntactic and semantic constraints in the formation and interpretation of ung-nouns. In M. Rathert & A. Alexiadou (Eds.), *The semantics of nominalizations across languages and frameworks* (pp. 169–214). Berlin: Mouton de Gruyter.

Rubinstein, D., Levi, E., Schwartz, R., & Rappoport, A. (2015 July). How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (Volume 2: Short papers)* (pp. 726–730). Beijing, China: Association for Computational Linguistics.

Sahlgren, M. (2006). *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, University of Stockolm.

Sahlgren, M., & Lenci, A. (2016). The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 975–980).

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Santus, E., Lu, Q., Lenci, A., & Huang, C.-R. (2014). Unsupervised antonym-synonym discrimination in vector space. In *Proceedings of the first Italian conference on computational linguistics CLiC-it 2014* (pp. 328–332).

Scheffler, T. (2005). *Nominalization in German*. Unpublished manuscript, University of Pennsylvania.

Scheible, S., Schulte im Walde, S., & Springorum, S. (2013). Uncovering distributional differences between synonyms and antonyms in a word space model. In *Proceedings of the international joint conference on natural language processing* (pp. 489–497).

Scherer, C. (2011). Polysemy and productivity in German. *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung*, *64*(1), 41–52.

Schirakowski, B. (2017). Methodological challenges in investigating nominalized infinitives in Spanish. In *Proceedings of the 8th tutorial and research workshop on experimental linguistics*, Heraklion, Greece.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of NeMLaP*, Manchester, UK.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 298–307).

Shin, S.-S. (2001). On the event structure of -ung-nominals in German. *Linguistics*, *39*, 297–319.

Skytte, G. (1983). *La sintassi dell'infinito in italiano moderno*. Charlottenlund: Museum Tusculanum Press.

Skytte, G., Salvi, G., & Manzini, M. (2001). Frasi subordinate all'infinito. In L. Renzi, G. Salvi, & A. Cardinaletti (Eds.), *Grande grammatica italiana di consultazione, Volume II: I sintagmi verbale, aggettivale, avverbiale. La subordinazione* (pp. 483–569). Il Mulino.

Spranger, K., & Heid, U. (2007). Applying constraints derived from the context in the process of incremental sortal specification of German ung-nominalizations. In *Proceedings of the 4th international workshop on constraints and language processing* (pp. 65–77).

Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, *71*, 249–267.

Tomaschek, F., Tucker, B. V., Ramscar, M., & Baayen, R. H. (2021). Paradigmatic enhancement of stem vowels in regular English inflected verb forms. *Morphology*, *31*, 1–29.

Turney, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of COLING* (pp. 905–912).

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*(1), 141–188.

Varvara, R. (2017). *Verbs as nouns: empirical investigations on event-denoting nominalizations*. PhD thesis, University of Trento.

Varvara, R., Lapesa, G., & Padó, S. (2016). Quantifying regularity in morphological processes: An ongoing study on nominalization in German. In *Proceedings of the ESSLLI DISSALT workshop: DISSALT Workshop on Distributional Semantics and Semantic Theory*, Bolzano, Italy.

Wauquier, M. (2016). *Indices distributionnels pour la comparaison sémantique de dérivés morphologiques*. Master's thesis, Université Toulouse Jean Jaurés, Toulouse.

Weeds, J., & Weir, D. (2003). A general framework for distributional similarity. In *Proceedings of the EMNLP 2003*, Sapporo, Japan (pp. 81–88).

Weeds, J., Weir, D., & McCarthy, D. (2004). Characterising measures of lexical distributional similarity. In *Proceedings of COLING 2004*, Geneva, Switzerland (pp. 1015–1021).

Werner, M. (2013). Nominalizing infinitives in the history of German: state-of-the-art & open questions. *Bavarian Working Papers in Linguistics*, *3*, 127–134.

Zeller, B., Šnajder, J., & Padó, S. (2013). DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (Volume 1: Long papers)* (pp. 1201–1211), Sofia, Bulgaria: Association for Computational Linguistics.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Reading: Addison-Wesley.

Zucchi, A. (1993). *The language of propositions and events. Issues in the syntax and the semantics of nominalization*. Berlin: Springer.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.