Sven Kotowski & Martin Schäfer

Quantifying semantic relatedness across base verbs and derivatives: English *out*-prefixation

Abstract: Most authors on derivational semantics agree that the meaning of complex words builds on components contributed by the base, components contributed by the word formation process, as well as contextual knowledge components. There is far less agreement on how these components interact or on their relative contributions. In this paper, we approach the question of relative contributions by looking at relatedness between bases, between derivatives, as well as between bases and derivatives. We use the English scalar-comparative verb prefix *out-* as a test case. We show that classes of bases serve as predictors for the resolution of systematically underspecified parts of the semantics of *out*-derivatives. By means of distributional similarity measures, we also show that the word formation process's derivatives exhibit a high degree of uniformity relative to the similarities between other components of the process. Finally, distributional measures for four further prefixes suggest that this rich contribution of the affix is peculiar to scalar-comparative *out-* rather than a characteristic of prefixation in general.

Keywords: verb semantics, English prefixation, distributional measures, quantitative semantics

1 Introduction

This article deals with the relative importance of base semantics and affix semantics in English *out*-prefixation. All accounts of derivational semantics agree that the meaning of complex words is in some way derived via an interplay of the meaning of the base, the meaning introduced by the word formation process, and contextual or extra-linguistic information (see e.g. Rainer et al. 2014, Lieber 2004). Also, it is common ground in the literature that the derivatives of any productive derivational process share a semantic core, as can be seen for example in the notions of 'semantic coherence' (Aronoff 1976), 'systematic form-meaning correspondences' (Haspelmath & Sims 2013), or 'lexical relatedness' (Spencer 2013). How-

Sven Kotowski & Martin Schäfer, SK: Heinrich-Heine-Universität Düsseldorf; MS: Eberhard-Karls-Universität Tübingen, e-mail of corresponding author: sven.kotowski@uni-duesseldorf.de

https://doi.org/10.1515/9783111074917-007

ever, lexical relatedness can come in various guises, and so can the prominence of the semantic contribution of different word formation processes relative to the contribution of their bases.

As argued by Spencer (2010a; 2010b; 2013), lexical relatedness is a continuum. It ranges from closely related word forms with minimal morphosyntactic differences ('canonical inflection' such as agreement morphology on verbs; *to draw* → *draws*) to more distantly related lexemes with differences in form, syntactic category, and semantic make-up ('canonical derivation', as in *to draw* → *drawable*).

In between these poles, we find less clear-cut categories. These include transpositional operations, i.e. syntactic category change with little or no semantic contribution (such as action nominalizations of the kind *to arrive* → *arrival*, as well as semantic shifts without formal marking (such as conversion phenomena of the kind *to dump* → *a dump*). In the following, we will focus on semantic relatedness in the realm of processes that are clearly derivational.

The general perspective on lexical relations just sketched is obviously process-related, i.e. it primarily looks at the relatedness of a given base and its derivative. However, the notion can also be extended to the relatedness between bases of a given morphological process or between its derivatives. The idea that the products of a derivational process are semantically related is a truism. If there was no discernible meaning shared by a set of derivatives, and if this meaning could not be attributed to the derivational process, we would not speak of such a process in the first place. Consider, for example, the commonalities between deverbal *-ible*/*-able*-adjectives in English. The vast majority of these adjectives indicate the capacity of the base verb's PATIENT/THEME argument to undergo the process denoted by that verb (e.g. *downloadable*, *searchable*). Quite evidently, the shared modal meaning component is a fairly specific product of *-ible*/*-able*-suffixation, rather than of the bases. At the same time, for this output to show such coherence, *-ible*/*-able* imposes the restriction that its base verbs license PATIENT/THEME arguments (see Plag 2003, 94f.; Bauer et al. 2013, 307f.).

However, morphological processes differ with respect to how specific and how uniform their meaning contributions are (see e.g. Koefoed & van Marle 2000; Wauquier 2020, ch.8). For example, the English person/participant-deriving suffix *-er* gives rise to various semantic categories, such as AGENT, EXPERIENCER, STIMULUS, INSTRUMENT, LOCATION, MEASURE etc. As a consequence of this multiplicity of meaning, the semantics of *-er* is difficult to pin down, rather underspecified, and highly dependent on the meaning of the base (Bauer et al. 2013, ch.11; Plag 2003, 89; Rappaport Hovav & Levin 1992). As shown in several studies on derivational polysemy, semantic classes of bases often predict or narrow down the possible readings of derivatives (see e.g. Aronoff & Cho 2001; Kawaletz 2021; Lieber 2016; Plag et al. 2018; Plag et al. 2023).

In this article, we build on the notion of relatedness and investigate the semantics of the English verbal comparative prefix *out-*, as in *to outsing s.o.* (i.e., roughly, to defeat s.o. by singing better/louder/more frequently etc.). We present four studies that quantitatively approach semantic relatedness between derivatives, between bases, and between bases and derivatives of this word formation process. First, we make use of traditional corpus studies that tap into the role of the base in the resolution of systematically underspecified elements of meaning in derivatives. Second, we employ distributional semantic measures and operationalize relatedness as distributional similarity across bases, between bases and derivatives, and across derivatives. Finally, we contrast the distribution of *out-* to that of several other English prefixes. The picture emerging from these studies is that comparative *out-* is fairly rich in semantic content and that derivatives show a relatively high degree of semantic relatedness to each other, both in comparison to the relatedness to their bases and in comparison to other prefixes. At the same time, we also provide quantitative evidence for base–derivative relatedness and show that both base lemmas as well as base verb classes are informative for spelling out underspecified meaning components.

The following section introduces the main characteristics of *out*-prefixation, motivates the studies, and provides an outline of the paper.

2 Background and objectives

2.1 Characteristics of comparative *out*-prefixation

A number of studies deal with comparative *out*-verbs, partly disagreeing on fundamental questions (see in particular Ahn 2022; Kotowski 2021; Talmy 2000; Tolskaya 2014). This section largely confines itself to uncontroversial properties of the word formation process. It shows why comparative *out-* is a highly interesting testing ground for exploring the questions of semantic relatedness, and thereby motivates the studies to follow.

Synchronically, comparative *out-* is among the most productive locative prefixes in English (see Schröder 2011), and all claims on base restrictions for this process have been shown to be far too restrictive (Kotowski 2021). Although there seems to be a preponderance of activity and semelfactive base verbs (as in *to outrun s.o.* and *to outfart s.o.*, respectively), verbs from all aspectual classes are attested as bases, such as stative verbs (e.g. *to outweigh s.o.*), achievements (e.g. *to outspot s.o.*), or change-of-state/accomplishment verbs (e.g. *to outcrush s.o.*). Comparative *out-* is also one of the few clearly category-changing prefixes in English

and regularly occurs with nominal (e.g. *to outtechnology s.o.*) and adjectival bases (e.g. *to outabsurd s.o.*). At least prima facie, *out-*'s bases thus do not seem to show systematic lexical relatedness.

In contrast, all derivatives exhibit a number of strikingly homogeneous features, irrespective of the respective base. First, all derivatives have a comparative meaning component. (1-a) is a fairly typical attestation, and the context clearly indicates the relevant scalar meaning component for *outsing*: the subject argument singing louder than the object argument, i.e. the lead vocalist. Interrelated with their comparative semantics, *out*-verbs are syntactically invariably transitive and, passive sentences aside, occur in frames of the kind [$\text{NP}_{subject}$–*out*-X–NP_{object}]. (1-b) shows the oddness of *out*-forms in intransitive contexts.

(1) a. Good vocal control is essential, a backing singer must not try to **'outsing' the lead vocalist** [...] it is important for singers who have strong voices to remember to back off the microphone a bit (iWeb)
b. ??He outsang.

Second, this syntactic configuration is a clear reflex of argument structural effects (see Haspelmath & Sims 2013, ch.11; Wunderlich 2012 on morphology inducing changes of argument structure). For example, the lemma SING occurs either as a one-place verb (e.g. *He sang*), or as a two-place verb with a THEME-argument that denotes some form of song or tune (e.g. *He sang a song*). (2-a) is odd, however, as a vocalist cannot easily be construed as the THEME of a singing-event. In contrast, the derived lemma OUTSING does not readily allow *a song* as direct object, in particular not as the THEME of a singing event, as indicated by (2-b).

(2) a. ??He sang the lead vocalist.
b. ??He outsang a song.

The example in (3-a) shows *out-* to have the same argument structural effect on a causative change-of-state (and thus necessarily transitive) base verb such as *to crush sth.*, as the object argument is not interpreted as the PATIENT of a crushing-event. Further, the denominal example in (3-b) illustrates that argument structure is also created in the case of base forms that do not readily support arguments on their own, such as *technology*.

(3) a. ...that woman has toes that could **outcrush a boa constrictor**![1]
b. Global big data competitors can **out-technology you** [...] (iWeb)

1 Bishop, David. 2006. Honour be Damned. Black Flame Publishing. Retrieved from http.//books.google.com, n.p.

Thus, comparative *out-* creates its own argument structure. It is applicative in the sense that it adds an object-argument or induces a change of thematic role of this argument. However, the specific nature of the introduced argument structure is contested. Let us very briefly sketch the two alternative analyses in the literature with the help of example (1-a). The first analysis would treat *outsing the lead vocalist* as a pure comparison between two distinct singing events (see e.g. Ahn 2022, Tolskaya 2014). Thus, *the lead vocalist* functions as the mere threshold surpassed by the subject argument regarding some event property (in this case, the loudness of singing events). The second kind of analysis understands *out*-prefixation as a competition construction that introduces a causative macro-event (see Kotowski 2021; Marchand 1969; McIntyre 2003). On this treatment, subject arguments are CAUSERS, i.e. *out-* introduces a further change in argument structure, while object arguments such as the *the lead vocalist* are licensed by a resultative sub-event. Conceptually, these object-arguments 'lose out' or are surpassed in a contest described by the morphological base (in the case of (1-a), by singing less loudly).

Either way, *out*-derivatives have an event structure that differs from the structure of the event denoted by the base. Simplified schematic semantic templates for the purely comparative approach and for the causative approach are given in (4) and (5), respectively.

(4) $\textsc{scalar property}(\textsc{event}_1(\textsc{participant}_1)) \xrightarrow{\text{exceed}} \textsc{scalar property}(\textsc{event}_2(\textsc{participant}_2))$

(5) $[\textsc{scalar property}(\textsc{event}_1(\textsc{participant}_1)) \xrightarrow{\text{exceed}} \textsc{scalar property}(\textsc{event}_2(\textsc{participant}_2))] \xrightarrow{\text{cause}} \textsc{defeated/surpassed}(\textsc{participant}_2)$

On either of these decompositional analyses, comparative *out-* can be regarded as a semantically rich word formation process, whose derivatives denote events of ontologically different types than the events denoted by their respective bases. This can be illustrated via a comparison to the far less productive spatial sense of *out-*. [2] Consider the derivative OUTHAUL, which is attested with both senses. In the comparative sense in (6-a), two eventualities or properties are compared (roughly, ships regarding their load capacities), and the object argument (*indus-*

2 Terminologically, we will always refer to the comparative sense unless indicated otherwise, i.e. we will always use 'spatial *out-*' when referring to the spatial sense. We will not deal with the historical relationship between spatial and comparative *out-* in this article and will remain agnostic regarding the question of whether they are analyzed best as two senses of one prefix or as two divergent word-formation processes; see Bauer et al. (2013, 347) for some discussion.

trial ships) is not interpreted as the THEME of a hauling-event. Rather, either of the two interpretations just sketched applies, i.e. analyses as an exceeding event or as a causation event. In contrast, the spatial prefix sense in (6-b) modifies the motion event encoded by the base by bounding its path, and the derivative still denotes a hauling-event (see e.g. Rappaport Hovav & Levin 2001, 780ff.; Zwarts 2008 on directed motion and event complexity). Unsurprisingly then, as in (6-c), the verbal base allows for the same argument structure as the form prefixed with spatial *out-*.

(6) a. With a great base cargohold and four low slots, it easily **outhauls all other racial Tech I industrial ships** at Industrial IV… (iWeb)
b. "haulback" means the cable used to **outhaul the rigging** or grapple when yarding… (iWeb)
c. I can vividly recall helping him on Sunday mornings to **haul the rope** that rang the church bell (iWeb)

In total, comparative *out*-prefixation thus seems semantically both rich and uniform: it introduces fairly predictable argument and event structures, and contributes comparative semantics. This latter component, however, remains systematically underspecified, as the scalar dimensions required for any form of comparison are not fixed on the lexical level (see Kennedy & McNally 2005; Solt 2015 for overviews of scalarity). Unlike in (1-a), for example, singing-events in (7) are compared along the dimension of QUALITY rather than LOUDNESS, as can be gathered from the underlined explication.

(7) …you can't deny [LBT's] <u>vocal abilities</u> […] nobody out there can **outsing them** <u>from a technical standpoint</u>. (iWeb)

The ambiguity of *out*-derivatives is not accidental, as the majority of verbs do not encode a single dimension (arguably, degree achievements such as *to heat* or *to widen* do; see Kennedy & Levin 2008). This does not mean that derivatives are systematically polysemous in the narrow sense. Rather, the morphological process introduces comparison but typically leaves underspecified the domain to which said comparison is applied. The ambiguity at hand is thus a matter of indeterminacy or vagueness (on indeterminacy, see Maienborn 2019; see also Mititelu et al. 2023).

At the same time, there is a strong intuition that an anything-goes approach to comparison in *out*-attestations is misguided. An immediate question that arises concerns the extent to which bases nevertheless determine, constrain, or help to identify possible scalar dimensions (see in particular Ahn 2019). The idea that the semantics of the base is crucial for dimension resolution is backed up by

asymmetric comparisons. Consider the item in (8), which clearly suggests a contest between an eagle that is flying and Mr. Paxton who is running. The example thus nicely illustrates that there are always at least two distinct sub-events in *out*-prefixation contexts. In (8), however, and unlike in the examples above, we deal with events of different kinds, which nonetheless allow for comparison (while the scalar dimension, possibly SPEED or DISTANCE, remains undetermined).

(8) "I wasn't going to run," Mr. Paxton said later after the game. "I figured I'm not going to **outrun an eagle**, so we might as well just see what happens." (forbes.com)

As argued by Kotowski (2021), it is the common conceptual nature of, or similarity between, items from the same lexical or ontological class that allows for asymmetric comparisons in the first place. Building on the example in (8), the reasoning can be summarized as follows: events of motion in space can be measured out along several parameters, some of which are salient, such as SPEED or DISTANCE (see e.g. Herweg 2020). Both running- and flying-events are subkinds of motion events. Given this shared property, they can be compared in general, and they also suggest salient dimensions for comparison (such as SPEED or DISTANCE). In contrast, we would not expect *out*-attestations to give rise to comparisons between, say, running- and singing-events.

2.2 Outline and objectives of studies

The objective of this paper is to empirically test several of the predictions derived from the characteristics of comparative *out*-prefixation just sketched. The four studies we present below aim at quantifying the contributions of the morphological base as well as the contributions of the word-formation process to the semantics of *out*-derivatives. Our approach across all of these studies is grounded in measures of semantic relatedness. Both the data and their coding (for Studies 1 & 2) as well as the code used in the distributional studies (Studies 3 & 4) are available at https://doi.org/10.6084/m9.figshare.20367735.v1.

In our first study, we are interested in a specific aspect of semantic relatedness of base and derivative. We quantify the role of different semantic classes of base verbs for resolving the underspecification of scalar dimensions in *out*-derivatives. To this end, we cull data from iWeb (Davies 2018) for all base lemmas from seven different VerbNet classes (Kipper et al. 2008), i.e. from classes whose respective members share semantic and conceptual structure. We annotate the data for con-

textual cues in order to find out whether base classes predict dominant scalar dimensions or distinct dimension profiles in derivatives.

The second study is very similar in nature to the first one, but focuses on individual base lemmas rather than classes of lemmas. We investigate the role of the individual lemma for resolving dimensional ambiguity in *out*-derivatives. Again using iWeb, a total of roughly 1,000 tokens from 12 base verb lemmas are annotated in their sentential contexts in order to investigate whether the class-based behavior established in Study 1 is reflected on the level of individual base verbs. In Study 2, we also take into consideration dimensionally unspecified tokens and thus quantify non-resolved underspecification of derivatives. By doing so, we address the question of how prominent specific interpretations of comparison are in the first place.

Our third study takes a more holistic approach to relatedness and analyzes the semantic coherence of *out*-derivatives relative to the semantic coherence of bases as well as relative to the relatedness of bases and derivatives. We make use of the ukWaC web corpus (Baroni et al. 2009), expanded by using derivatives from iWeb, and employ distributional semantic measures (see Boleda 2020) for the same verbs that we used in Study 2. We operationalize coherence as cosine similarities between the vectors of different components of the morphological process. Building on the hypothesis that *out-* is a semantically rich word-formation process with a fairly uniform output, we assume that creating argument/event structure and adding comparative semantics are reflected in distributional measures. We therefore expect relatively high degrees of similarity between *out*-derivatives when compared to the similarities between bases or between bases and derivatives.

Finally, our fourth study establishes whether the patterns of similarity found for *out-* are peculiar in the grander scheme of English prefixation, or whether these are a common feature of verb-to-verb derivation in English. To this end, we compare the distributional behavior of *out-* established in Study 3 with similarity measures for four further prefixes (spatial senses of *over-* and *out-*, reversative *un-*, and iterative *re-*). All of these additional processes presumably differ from *out*-prefixation regarding how rich their respective semantic contributions are: derivatives (mostly) encode the same event types as their bases and show no or weaker argument structural effects than *out-*. We therefore hypothesize that *out*-derivatives show more pronounced semantic coherence than the derivatives of the other prefixes investigated.

Tab. 1: Properties of the seven VerbNet classes used in Study 1 (see Kipper et al. 2008).

VerbNet-class	Description	No. of members (examples)
RUN	manner of movement	159 (*crawl*, *creep*, *run* etc.)
PERFORMANCE	performance as effected object	29 (*chant*, *play*, *dance* etc.)
EXIST	existence at some location	26 (*dwell*, *exist*, *live* etc.)
CARRY	caused accompanied motion	20 (*carry*, *drag*, *draw* etc.)
HIT	bringing an entity into contact with another entity	40 (*bang*, *hit*, *jab* etc.)
SPRAY	covering of surfaces	48 (*baste*, *sprinkle*, *splash* etc.)
SOUND EMISSION	emission of sound	129 (*babble*, *cry*, *rap* etc.)

3 Study 1: Predicting scalar dimensions via base verb classes

3.1 Rationale

In this study, we are interested in quantifying the role of the base verb class for the resolution of underspecification in derivatives. We probe the connection between base verb classes and the specific scalar dimensions encoded in *out*-derivatives:

Research question 1a (RQ1a): Do semantic classes of base verbs show preferences for specific (sets of) scalar dimensions in the *out*-derivatives based on members of this class? Can we predict unique distributions or profiles via class membership?

In order to answer RQ1a, we culled a large sample of *out*-formations from iWeb with bases from members of seven VerbNet classes. VerbNet is a lexical data base that categorizes verbs into semantic classes based on their compatibility with syntactic frames and semantic argument specifications. VerbNet classes have been successfully employed as partial predictors for polysemy patterns in English nominalizations (see Plag et al. 2018; Kawaletz 2021; Plag et al. 2023). In this study, we use the classes introduced in Table 1.

The classes in Table 1 were chosen for three reasons. First, all the classes needed to be large enough (at least 20 and up to 159 members). Second, previous corpus searches had shown that several of their members are attested as bases to *out*-derivatives. Third, we wanted the classes to be conceptually fairly different and therefore made sure that no two classes are from the same superordinate

class. This ensured that the events the respective members of our classes encode differ in salient properties.

Ambiguity of verbs in VerbNet, including both homonymy and polysemy, is characterized by multiple class membership. For example, SING is cross-listed in three classes, as shown in Table 2. As different verbs show different degrees of ambiguity, VerbNet classes also encode different levels of ambiguity once we generalize over their respective sets of members. In order to assess the influence of ambiguity on the dimensions encoded in *out*-contexts, we therefore also investigate multiple class membership of base forms along the lines of RQ1b.

Tab. 2: The different VerbNet classes in which *to sing* is listed.

Example	VerbNet-class	Syntax/argument structure
Susan sang to the children.	MANNER-SPEAKING	AGENT V {+$_{DIRECT}$} RECIPIENT
Sandy sang a song.	PERFORMANCE	AGENT V THEME
The street sang with horns.	SOUND EMISSION	LOCATION V {with} THEME

RQ1b: Does the degree of ambiguity of the members of a base class predict how informative this class is regarding the scalar dimensions encoded in *out*-derivatives?

3.2 Methodology

All attestations used in Study 1 were culled from iWeb via its web interface (see https://www.english-corpora.org/iweb/). Queries were performed individually for all potential base verbs. An example query string for the base lemma RUN is provided in (9) and returns (among nonpertinent hits) all possible word forms of the lemma OUTRUN, both with and without hyphens, i.e.: *outrun*, *out-run*, *outruns*, *out-runs*, *outran*, *out-ran*, *outrunning*, and *out-running*. Mutatis mutandis, this search was performed for all verbs from the VerbNet classes described in Table 1 above.

(9) outrun*|out-run*|outran|out-ran

A total of 451 searches were performed (corresponding to the total number of verbs in all seven classes), which yielded 104 different *out*-verb types, i.e. 104 different base verbs as input to *out*-. We did not distinguish between hyphenated and non-hyphenated examples. For example, *outrunning* and *out-running* were counted as

tokens of the same lemma OUTRUN. For lemmas with more than 100 pertinent hits, we selected the first 100 hits and included different word forms for a given lemma relative to their frequencies.[3] For lemmas with fewer than 100 hits, all tokens with explicit information on dimensions were included in the preliminary data set.

Subsequently, we used iWeb's expanded context menu and coded the data for explicit information on the specific dimensions on which the comparisons were based. The data in (10) illustrate our procedure for the lemma OUTRUN. On the basis of the underlined material, we classified SPEED as the target dimension for the context in (10-a) and coded the item as the lemma–dimension combination OUTRUN-SPEED. In contrast, the item in (10-b), again based on the underlined material, was coded as the combination OUTRUN-DISTANCE. Finally, items such as (10-c) were discarded from the data, because their contexts do not resolve dimension ambiguity. As this study looks at the predictive power of verb classes rather than lemmas, all lemma–dimension combinations, such as OUTRUN–SPEED for (10-a), were counted only once. All other combinations of OUTRUN and SPEED were discarded from the data. This method resulted in a total of 148 lemma–dimension combinations.

(10) a. He immediately ran away before I could get there myself and **outran me** (I didn't pick up the elven swiftness skill, usually if I want speed I just mount a beast so running isn't my characters forte).
b. Arsenal have been **outran by all of the Premier League opponents**. [...] Arsenal players have clocked in less kilometres than their rivals...
c. Jacquelyn Sertic [...] retired Oklahoma in order in the bottom of the ninth. That included an over-the-shoulder catch from DeCamp, who **outran the ball** into left center to make the catch.

Coding for the variable 'dimension' is not always straightforward. This holds in particular as there is no predefined list of possible values, i.e. dimensions, and as it is impossible to a priori define what constitutes specific contextual information on a dimension for any case imaginable. We used the following strategy: The annotations were carried out by the first author. In order to establish both the difficulty and the reliability of the task, a second annotator went through the whole data set

3 That is, if for the lemma OUTRUN there were 900 tokens of word form *outruns* and 100 tokens of word form *out-ran*, we would have kept the first 90 *outruns* tokens and the first 10 *out-ran* tokens from iWeb's context menu on the web interface. We discarded obviously corrupted corpus attestations and, if possible, topped up the data with further attestations in their stead. To keep our procedure as reproducible as possible, we did not randomize the hits. That is, later iWeb searches will return the same order of hits for a word form. We think this procedure is all the more reasonable, as whatever ordering iWeb uses internally, this will be used across all items.

used in Study 2 ($N = 948$), which is a subset of the data that we looked at in Study 1. The two annotators agreed in 64% of all cases, using a total of eleven different values (Cohen's Kappa for two raters: 0.394, $z = 22.4$, $p = 0$). Although this amounts to only fair to moderate agreement (see e.g. McHugh 2012), we decided to include all data from the first author's original annotation for the following reasons: First, the percentage of cases agreed upon is comparable to annotator agreement in similar semantic tasks (see e.g Maguire et al. 2007 with 68% for two raters on compound relations). Second, and more importantly, disagreement between the two annotators concerns almost exclusively the question of whether an item is contextually specified for a dimension or not. There are only five cases for which the annotators disagreed regarding a specified dimension (for example, an item which one rater classified as OUTSWIM–STAMINA, and the other rater as OUTSWIM–DISTANCE). Reporting only those items both annotators agreed on thus primarily leads to higher relative percentages of unspecified cases and the loss of a substantial amount of data. We therefore decided to keep all items in the data set.

In order to address the ambiguity problem (see RQ1b), we operationalized multiple class membership in VerbNet as a measure of uncertainty. We quantified how ambiguous the members of a given class are on average by, first, counting the number of VerbNet classes each base lemma in the data set is listed in. For example, the lemma SING occurs in three classes in total (PERFORMANCE, SOUND-EMISSION, and MANNER-SPEAKING; see Table 2). We therefore assigned SING a cross-listing score of 2, i.e. besides being a member of the class under investigation, it is listed as a member of two further classes. Second, we calculated the mean of the individual base verb scores for each class in our study and used this as the cross-listing score for the respective classes. This value can then be understood as a measure of uncertainty: the higher a class's cross-listing score, the less sure we can be that the *out*-formation's base is in fact from this class.

The classes' cross-listing scores were set in relation to values assessing dimensional variability. First, higher percentages for a dominant dimension for a given class are taken as indicators of the class's homogeneity. In contrast, higher absolute numbers of different dimensions and a higher dimension entropy for a class are taken as indicators for heterogeneity. Dimension entropy is a measure of the overall level of uncertainty about the dimensions that are associated with a class. We calculated it via probabilities derived from the dimension distribution for each class. The more dimensions with similar frequencies there are for a class, i.e. the more difficult it is to predict which dimensions are linked to a verb type, the higher the entropy. The fewer dimensions we find for a class, and the more frequent one dimension is compared to others, the lower the entropy. The entropy is at zero if there is only one dimension.

Tab. 3: Dimension distribution in % over 7 VerbNet-classes (*N* = 148; percentages are rounded; percentage of majority dimension per class underlined; EMISSION is short for SOUND EMISSION).

	PERFORMANCE	RUN	EXIST	SPRAY	EMISSION	CARRY	HIT
QUALITY	48	14	/	33	15	5	20
SPEED	5	51	/	7	4	32	27
DURATION	14	/	67	/	8	/	/
QUANTITY	5	/	/	60	4	21	/
LOUDNESS	14	/	/	/	50	/	20
DISTANCE	/	21	/	/	/	16	/
FREQUENCY	/	5	/	/	12	5	20
IMPACT	/	/	22	/	/	5	/
other	14	9	11	/	8	16	13

3.3 Results

The distribution of lemma–dimension combinations over the seven VerbNet classes is shown in Table 3. Base verb classes are represented in table columns and dimensions in lines.[4] The respective majority dimension for each class is indicated by underlined percentages. For example, read top-down, the column for the PERFORMANCE base class (leftmost) encodes the following information: 48% of lemma-dimension combinations co-occur with the majority dimension QUALITY, 5% with SPEED, 14% with DURATION, 5% with QUANTITY, 14% with LOUDNESS, and 14% with further dimensions (i.e. category 'other').

All classes include different dimensions and for five out of the seven classes, there is a clearly dominant, class-specific dimension that approaches or exceeds 50% of all type-dimension combinations of this class. Only the CARRY and HIT classes behave differently in this respect, with SPEED as the most dominant category in both cases (32% and 27%, respectively). The dimension profiles are unique to all base classes, i.e. no two classes co-occur with the same dimensions.

Several values are necessary for assessing how the ambiguity of a class's members influences the variability and the distribution of dimensions. Information on the dominant dimensions per class can be retrieved from Table 3. In Table 4, the first row shows the cross-listing scores for the base classes, dimension entropy is displayed in the second row, and the third row shows the number of different dimensions per class.

4 The category 'other' exists for reasons of readability and includes various low frequency dimensions. Only dimensions that made up more than 10% for at least one base class were included in the table as individual dimensions.

Tab. 4: The cross-listing scores (CL score), the dimension entropies (Entropy), and the number of different dimensions (Dimensions) for the 7 VerbNet-classes (EMISSION is short for SOUND EMISSION).

	PERFORMANCE	RUN	EXIST	SPRAY	EMISSION	CARRY	HIT
CL score	3.1	1.0	1.1	0.7	2.2	3.0	3.3
Entropy	2.13	1.89	1.22	1.24	2.23	2.49	2.29
Dimensions	8	7	3	3	8	8	6

We tested for correlations between the classes' cross-listing scores and the three variability values. Our data show that the more ambiguous the members of a class are, the more different and the more diverse scalar dimensions we find encoded in *out*-forms based on these members. As depicted in Figure 1, the higher its cross-listing score, the weaker a class's potential for predicting a dimension profile. First, the cross-listing score is strongly negatively correlated with dominant dimensions: the lower the cross-listing score, the higher the percentage of the dominant dimension (Pearson's r: -0.83, $p = 0.022$). Second, the negative correlation is even stronger, and more significant, for cross-listing score and dimension entropy (Pearson's r: 0.85, $p = 0.016$). Last, the absolute number of dimensions per class also correlates negatively with the cross-listing score, but this correlation is not significant(Pearson's r: 0.67, $p = 0.09$).

3.4 Discussion

Study 1 set out to investigate the extent to which semantic classes of base verbs predict the scalar dimensions of the comparative meaning component of *out*-derivatives. The results clearly show lexical semantic relatedness between verb classes and derivatives. The base classes we investigated are associated with different dimension profiles in *out*-formations and we find clearly dominant dimensions for five out of the seven classes. With respect to RQ1a, these findings suggest that VerbNet classes serve as suitable vantage points for predicting the intended interpretations of *out*-derivatives.

For example, our results suggest that the preference for the dimension SPEED for *out*-lemmas based on RUN-class bases, as opposed to, for example, PERFORMANCE-class bases, is no coincidence. This finding is in line with the assumption in section 2.1 that the word-formation process remains underspecified for scalar dimensions, and reliant on its bases in this regard (see Ahn 2022; Kotowski 2021). Therefore, it appears reasonable to assume that the respective dominant dimen-

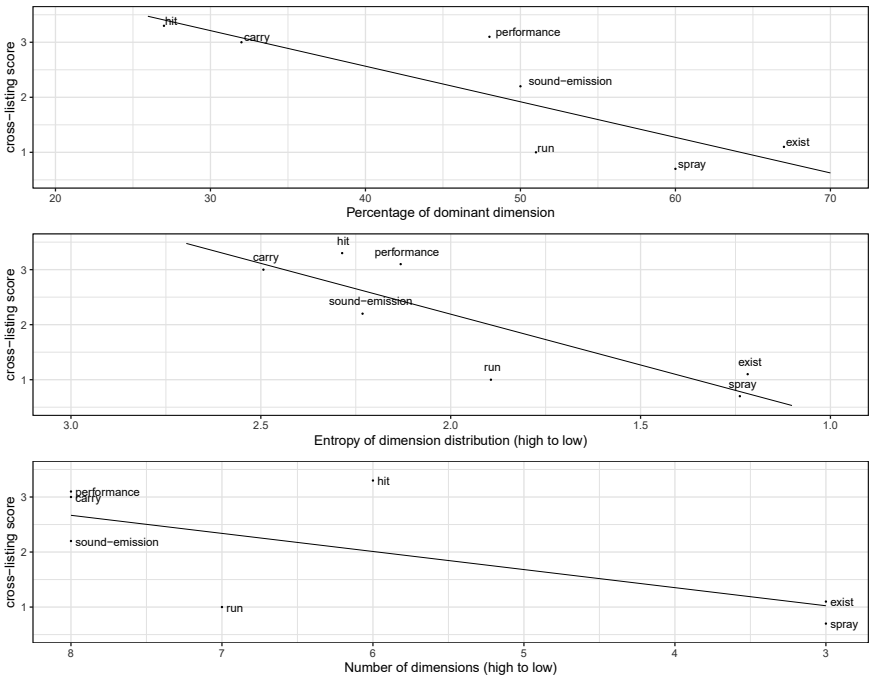

Fig. 1: From top to bottom: correlation dominant dimension–cross-listing score, Pearson's *r*: -0.83, *p* = 0.02), correlation dimension entropy–cross-listing score, Pearson's *r*: 0.85, *p* = 0.016, and correlation of number of dimensions—cross-listing score; Pearson's *r*: 0.67, *p* = 0.09. The lines show the corresponding linear regression lines.

sions are salient properties of the events denoted by the verbs from classes with unambiguous dimension profiles, such as SPEED for RUN-verbs.

However, all classes have the potential to give rise to a number of dimensions (between 3 and 8) and two classes, HIT and CARRY, do not exhibit dominant dimensions. Regarding the question of the influence of ambiguity on a class's predictive potential (see RQ1b), two different correlations are insightful: the degree of ambiguity of a VerbNet class's members, quantified as a class's cross-listing score, is strongly negatively correlated with both the class's dominant dimension and the entropy of the dimension distribution. If we take multiple membership as an indicator of polysemy (or partly homonymy), dimension profiles thus partially depend on how ambiguous a class is: the more polysemous a given VerbNet class's members are overall, the more difficult it is to predict a dominant dimension based on this class.

Building on these class-based findings, let us now move on to the closely related lemma-based investigation in Study 2.

4 Study 2: Predicting scalar dimensions via base verb lemmas

4.1 Rationale

This study investigates the role of different base lemmas as predictors of scalar dimensions in *out*-formations. Unlike Study 1, it does not count types of lemma–dimension combinations, but looks at the frequencies of token–dimension combinations:

Research question 2a (RQ2a): To what extent do tokens of individual verbs reflect the dimension profiles of their respective verb class?

In this study, we are also interested in how frequently ambiguity is in fact explicitly resolved in context. Recall that neither of the analyses introduced in section 2.1 doubt that *out-* includes a comparative core. However, large numbers of tokens in which dimensional ambiguity remains unresolved would possibly cast doubt on comparison as the word formation process's sole semantic contribution, and would make causative analyses that do not rely on the prominence of comparison more plausible:

RQ2b: How frequently is dimensional underspecification resolved and is such resolution the default case?

In order to answer RQ2a and RQ2b, we used large samples of tokens from twelve *out*-lemmas based on three of the VerbNet classes used in Study 1.

4.2 Methodology

Methodologically, Study 2 is very similar to Study 1. All attestations were culled from iWeb via its web interface. We used the same search format as above (see the search string in example (9)), but performed searches for twelve base verbs. These verbs and their respective VerbNet classes are shown in (11). If available, 100 attestations were culled from iWeb. For lemmas with less than 100 hits, we

included all items. For lemmas with more than 100 hits, we used the first 100 hits and included different word forms of a given lemma relative to their frequencies (see footnote 3 for the procedure). Upon discarding hits based on obvious corpus corruptions (e.g., *Hardly anything comes outWritten*), we ended up with a data set of 949 tokens; the number of tokens per lemma is given in parentheses in (11):

(11) PERFORMANCE: outdance (51), outrap (21), outwrite (32), outsing (100)
RUN: outrun (100), outfly (100), outswim (100), outsprint (100)
EXIST: outlive (100), outstay (100), outwait (100), outsurvive (45)

For any individual token, we annotated a specific dimension in case concrete contextual information was available in iWeb's expanded context menu.

Unlike in Study 1, we kept items without explicit dimension information. Examples such as (10-c) were coded as "unspecified" regarding their dimension. The data set created for Study 2 further contrasts with the data set used in Study 1 as multiple occurrences of a particular *out*-lemma and a particular dimension were counted, e.g. multiple tokens of OUTRUN specified for the SPEED-dimension.

4.3 Results

The distribution of dimensions across tokens of the twelve *out*-lemmas are shown as horizontally stacked bars grouped by base verb class in Figure 2. Each bar represents all tokens of one *out*-lemma. The topmost row in Figure 2-a), for example, encodes all 100 attestations of the lemma OUTSPRINT. Bar sections indicate the percentages of attestations co-occurring with a scalar dimension as well as with unspecified dimensions (black sections). For OUTSPRINT, the bar shows the following distribution: 32% of the tokens are unspecified regarding a scalar dimension and 68% explicitly refer to SPEED.[5]

For nearly all lemmas, we find a general pattern of a relatively high proportion of unspecified tokens (on average, 52% of a lemma's total) and a clear majority dimension among the specified cases (on average, 42% of a lemma's total). Moreover, we find a clear pattern of intra-class homogeneity, i.e. lemmas from a class behave similarly, and the preferred dimensions are the same ones we find for the class-based investigation (see Table 3). For example, just as the EXIST-class itself,

5 The category 'other' exists for reasons of readability and includes various low frequency dimensions. Only dimensions that make up more than 5% of dimensions for at least one verb of a base class were included in the table as individual dimensions.

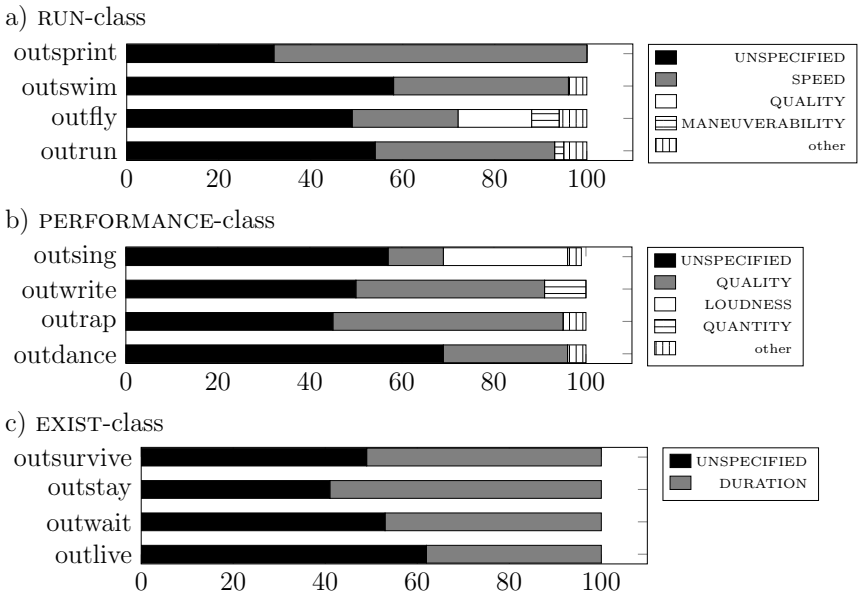

Fig. 2: Distribution in % of scalar dimensions in *out*-forms for lemmas based on VerbNet's a) RUN-class, b) PERFORMANCE-class, and c) EXIST-class; $N = 949$.

the tokens of all four *out*-lemmas based on EXIST-verbs have a preference for the dimension DURATION.

Only one lemma does not follow this overall pattern. The verb OUTSING from the PERFORMANCE-class is attested most with LOUDNESS (27%). Given the results of Study 1 as well as the behavior of other PERFORMANCE-verbs, one would expect QUALITY to feature as the verb's majority dimension, for which we only find 12% of attestations. To a lesser degree, OUTFLY stands out as well, as it shows a less clear preference for SPEED compared to the other RUN-lemmas.

4.4 Discussion

Regarding RQ2a, the class-based findings from Study 1 are clearly reflected in the results of the token-based investigation. In other words, the investigated lemmas cluster by class with respect to their preferred dimensions, and these dimensions are the same ones that are preferred by the respective classes. Leaving aside 'unspecified' cases, all four RUN-lemmas occur most often with SPEED, all four EXIST-lemmas with DURATION, and three out of four PERFORMANCE-lemmas with QUAL-

ITY. More generally speaking, the results thus indicate base–derivative relatedness regarding salient event properties.

Only the lemmas OUTSING and, to a lesser degree, OUTFLY deviate from the behavior their class membership would suggest. A plausible explanation for these two lemmas' preferences can be found in the ambiguity of their respective base forms. For example, SING is cross-listed in VerbNet as a PERFORMANCE, a MANNER-OF-SPEAKING, and a SOUND-EMISSION verb (see Table 2). The reason that OUTSING-contexts are mostly measured out along the scale of LOUDNESS (27%), rather than along the scale of the expected dimension QUALITY, appears to be that *out-* preferentially takes the verb's SOUND-EMISSION sense as base. As shown in section 3.1, members of the SOUND-EMISSION class productively serve as base to *out*-prefixation and LOUDNESS is this class's preferred dimension for comparisons.[6]

With respect to RQ2b, we find a large proportion of unspecified cases for all lemmas, i.e. attestations without explicit contextual clues on scalar dimensions. For eight of the twelve lemmas, these unspecified cases constitute the majority among the tokens and their nature is in need of comment. Importantly, having been coded as 'unspecified' does not mean that an item does not suggest a dimension. Consider, for example, the attestations in (12) (all from iWeb):

(12) a. When was the last time we <u>made more noise</u> than the away support (we've been **outsung** by fewer supporters than Wrexham brought)? [coded as LOUDNESS]
b. I thought our fans were terrific. They showed the world that when they fill any stadium they will not get **outsung**. [coded as UNSPECIFIED]
c. The most memorable scene of unadulterated Allied jingoism occurs when free-French patrons of Ricks Place (sic) **out-sing** the Nazi patrons in a battle of national anthems. [coded as UNSPECIFIED]

Based on the underlined material (*make more noise*), the dimension for the item in (12-a) was coded as LOUDNESS, while the item in (12-b) was coded as 'unspecified'. However, like example (12-a), very many OUTSING-examples refer to football supporters competing in some form of singing contest whose primary goal is to sing louder than, and thereby defeat, a rival group of supporters at some football ground. Arguably, it is such pieces of world knowledge that lead to the inference

6 We do not have any numbers available for VerbNet's DRIVE class. However, it seems likely that OUTFLY's behavior finds a similar explanation via FLY's membership in this class. As the DRIVE class's members typically specify the means or vehicle used in motion events, they arguably suggest quality-judgements of the operation of a vehicle.

that examples such as (12-b) are also based on LOUDNESS and thus follow the pattern of the specified majority dimensions of the lemma.

At the same time, for quite a number of examples the available linguistic context does not allow for resolving dimensional ambiguity, as in (12-c). Recall from section 2.1 that two competing theories on *out*-'s semantics analyze the comparative component as either the macro-event itself (Ahn 2022; Tolskaya 2014) or as embedded as a sub-event in a causative macro-event (Kotowski 2021; McIntyre 2003). On the fairly uncontroversial assumption that dimensions are indispensable in scalar constructions (see Solt 2015), the existence of a fair amount of unclear examples at least casts doubt on comparison as the only semantic component of *out*- as a word formation process. Albeit tentatively, we regard this finding as an argument in favor of causative analyses that rely less on the prominence of comparison.

We will now move on to Study 3 and the investigation of distributional similarity measures.

5 Study 3: Distributional similarities – *out*- across VerbNet classes

5.1 Rationale

In Study 3, we operationalize semantic relatedness using distributional similarity measures and apply these to both the *out*-lemmas used in Study 2 as well as to their bases. The objective is to tease apart the relative contributions of the base and of the word formation process to the semantics of *out*-derivatives. To this end, we look at distributional similarities between bases, between derivatives, and between bases and derivatives.

Distributional semantic approaches have been used successfully in investigating a variety of morphological phenomena. A number of studies tries to model derivation by isolating a distributional representation of an affix and combining it with the vector representation of the base. For example, Marelli & Baroni (2015) model affixes as matrices that are multiplied with base vectors, while Padó et al. (2016) and Kisselew et al. (2015) test various models on their predictive power depending on the base forms' properties. A similar strain of research explores whether different types of affixation show different distributional reflexes. For example, Bonami & Paperno (2018) investigate distributional differences between inflected and derived forms, Varvara (2017) looks at different German event nominalizations, while Bonami & Guzmán Naranjo (2023) explore distributional evi-

dence for paradigmatic processes of derivational categories. Finally, some studies have looked at more fine-grained aspects of derivation. Lapesa et al. (2017) show that when using valence as a tertium comparationis, the same German derivational suffixes have different effects depending on lexical properties of the bases, while Lapesa et al. (2018) use distributional semantics to disambiguate the meanings of new English *-ment* derivations. Similarly, Wauquier (2020) uses the vectors of different semantic and morphological classes of words to investigate the uniformity of these classes, and Schäfer (2023) investigates the alleged derivation–inflection dichotomy by looking at the distribution of *-ly*-adverbs based on different semantic base classes.

Our approach differs from the first strain since we are not interested in modeling derivation as composition or subtraction. It differs from the second strain in that we only consider a single prefix and its base-dependent internal variation. There are conceptual similarities with the last strain, as we are interested in specific aspects of a single word formation process, and in the possible influence of different semantic classes of bases.

The first measure we are interested in concerns similarities between bases and derivatives. We compare pairwise similarities, i.e. similarities between base and derivative (e.g. between the lemmas RUN and OUTRUN), with similarities across pairs, i.e. similarities between non-pairs of bases and derivatives in the data (e.g. between RUN and OUTSWIM). If properties of the base are preserved in *out*-derivatives, as strongly suggested by Studies 1 and 2 above, we expect base-derivative pairs to be more similar to each other than non-pairs are to each other. This measure can also be taken as a proof of concept for Study 3, as the assumptions reflect that, lexicalized forms aside, we expect the semantics of base words to feature in the corresponding derivative semantics:

Research question 3a (RQ3a) Do distributional measures show a higher degree of similarity between base-derivative pairs than between non-pairs?

The argument structural and event structural properties and the comparison-adding nature of *out*-prefixation clearly suggest a generally rather homogeneous semantic make-up of *out*-derivatives (see Section 2.1). We therefore also investigate similarities that hold between derivatives (e.g. between OUTRUN and OUTSWIM) and how these derivative-derivative similarities compare with both base-base similarities (e.g. between RUN and SWIM) and with similarities of base-derivative pairs:

RQ3b Do derivative-derivative similarities differ from base-base similarities and base-derivative similarities?

5.2 Methodology

Most *out*-derivatives have low token frequencies in corpora and are therefore typically not included in pretrained collections of distributional vectors. For example, the semantic space that produces the best empirical results in Baroni et al. (2014) only contains vectors for three of the *out*-lemmas from Study 2: OUTLIVE, OUTRUN, and OUTSTAY. We therefore decided to calculate our own vectors, opting for a classic distributional semantics approach strictly based on co-occurrence counts.

5.2.1 Corpus

We used a combination of two corpora for our measures. To create our vectors, we used the ukWaC corpus, a web-derived 2 billion word corpus of English (see Baroni et al. 2009). The version we used is part-of-speech-tagged and lemmatized with TreeTagger.[7] As the absolute frequencies of most of our target items are still low in ukWaC, we expanded the corpus by adding all sentences containing the *out*-lemmas from the iWeb corpus (14 billion words).

In order to be able to use the iWeb data as an expansion of the tagged ukWaC corpus, we first extracted all sentences containing any wordform of our target lemmata. We then tagged and lemmatized the sentences using Python's NLTK library (see Bird et al. 2009) and adjusted the tags to the tree tagger conventions used in the tagged ukWaC.[8]

Table 5 provides the absolute numbers of occurrences of *out*-formations that we extracted from iWeb.

7 https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

8 One reviewer remarks that ukWac is a corpus of British English, while iWeb is a multi-variety corpus. While we acknowledge this difference, we are not aware of any differences in the usage of *out*- depending on the variety of English. We therefore cannot comment on whether this may bear on our results.

Tab. 5: *out*-derivatives in iWeb: absolute number of occurrences.

EXIST-verb	occurrences	PERFORMANCE-verb	occurrences	RUN-verb	occurrences
outlive	12389	outdance	51	outfly	134
outstay	830	outrap	21	outrun	11291
outsurvive	46	outsing	166	outsprint	545
outwait	134	outwrite	62	outswim	104

5.2.2 Creation of distributional vectors

We created vectors strictly based on co-occurrence counts, that is, no machine learning was involved. We proceeded as follows:

1. Co-occurrence counts were collected for the top 10,000 content words (nouns, verbs, adjectives, and adverbs) in ukWaC, resulting in raw vectors with 10,000 dimensions for our target lemmata.
2. For further parameter setting, we explored different window sizes (sentence, and ranges of 2 to 4 words to the left and right of the target word) and transformations (pointwise mutual information and log-likelihood, both with and without logarithm) by comparing the correlation of the resulting cosine similarity measures between pairs of verbs with the human similarity and association scores in the SimLex-999 dataset described in Hill et al. (2014).
3. We chose the combination of window size and transformation that correlated best with both the similarity and the association scores: a three-word window, and a transformation of the raw counts to pointwise mutual information (with logarithm). This setting has also been shown to perform best for adjective similarities in the SimLex-999 dataset (see Schäfer 2020).

Keeping our procedure as simple and transparent as possible, we did not reduce the dimensionality of the resulting vectors (note also that Baroni et al. 2014a found that within the count models they considered those using no compression at all worked best).

5.2.3 Accomodating differences in overall frequencies

The similarities we are interested in are between vectors that are based on hugely diverging numbers of absolute occurrences. These differences are tied to major differences in the expected range of co-occurring other lexical items. Take our least frequent lemma, OUTRAP, as an example: it occurs 21 times in total. Given that

we used a three-word window to the left and right of our target words, the theoretically maximally possible number of different content words that could occur here are only 6x21=126 content words, necessarily resulting in a very sparse vector. Even its base, RAP, the base with the lowest frequency, occurs only 1589 times, yielding maximally 6*1589 = 9534 content words, still less than the dimensions of our vectors. This is very different from our most frequent base, WRITE, which occurs 837,319 times. In order to address this issue, we used vectors created from smaller samples of co-occurrences.

Specifically, per lemma, we drew 10 random samples of 100 contexts each from the set of all contexts of that lemma. We then created 10 sets of vectors based on these 10 random samples for each of these lemmas. The size of 100 contexts per sample makes the resulting vectors very similar in terms of the expected overall sparsity to the vectors of the rare items. Using 10 random samples each safeguards against the possibility that a single randomly drawn set of 100 contexts might in fact not be very representative of the contexts of the target verbs. Four verb lemmas occur in less than 100 sentences and were therefore not downsampled (i.e. OUTWRITE (62 tokens), OUTRAP (21), OUTSURVIVE (46), OUTDANCE (51)).

To calculate cosine similarities, we used average values across all ten samples. That is, the similarity between the rare item OUTWRITE and the most frequent base WRITE is calculated by taking the mean of the ten cosine similarities between the vector for OUTWRITE on the one hand and each of the 10 vectors based on the 10 downsampled contexts randomly drawn from all WRITE sentences on the other hand. Thus, we report the average of 10 similarity values. In contrast, the similarity between RUN and OUTRUN, both occurring more than 100 times, is based on the average similarity between 10 vectors each, created from 10 random samples of 100 contexts for both of these lemmas. In this case, we report the average of 100 similarity values. Note that one result of our approach is, relatively speaking, low similarity values. For example, the mean similarity across all the base-derivative pairs with *out-* is at just 0.06. For validation of our approach, we compare these values against the values of the full data, i.e. pre-downsampling, and against values from word embeddings in the next section.

5.2.4 Comparison to word embeddings

Most current distributional studies use machine-learned word embeddings (cf. the two other studies using distributional semantics in this volume, Bonami & Guzmán Naranjo 2023 and Schäfer 2023). While Baroni et al. (2014a) argue against using count vectors, subsequent studies comparing the two appproaches paint a more nuanced picture, see Varvara et al. (2021). Most importantly, by using count

vectors, we keep our approach transparent and lightweight. It is transparent, because neither machine-learning nor dimension reduction is involved, and we know exactly how every single value in the vectors was created. It is lightweight, because our method only requires the sentences with our target verbs and the already available frequency list for all words in the corpus. Furthermore, we believe that our method of accounting for frequency effects by both some transformation (in our case, pointwise mutual information with logarithm) and our downsampling procedure can lead to new insights into how frequency differences are best dealt with.

For comparability to word embedding approaches, we compared our vectors in their full (pre-downsampling) and downsampled versions against the best performing vector space in Baroni et al. (2014a), available via Baroni et al. (2014b).[9]

The Baroni vectorspace is not lemmatized. For our comparison, we used the base form vectors. Since only three of our *out*-derivatives in their base forms are included, we compared not only against the pairs used in Study 3 but also against the additional pairs used in Study 4.

For the resulting 13 base–derivative pairs, we calculated three sets of cosine similarities, using i) the Baroni vectors, ii) our pre-downsampling vectors, and iii) our downsampled vectors. Table 6 summarizes the range of the similarities observed in the three vectorspaces.

Tab. 6: Overview of the distribution of cosine similarities based on three different sets of vectors.

vectors	Min	Max	Mean	Median
a. Baroni	0.05	0.51	0.22	0.20
b. Pre-downsampling	-0.06	0.50	0.17	0.15
c. Downsampled	0.03	0.15	0.06	0.05

Mean and median are relatively close to each other in all three datasets, with the mean always slightly higher, reflecting a right skew in all three datasets. Both the pre-downsampling and the Baroni vectors are overall similar to a normal distribution, while the downsampled vectors are clearly different (Shapiro-Wilk normality test, W = 0.76, p-value = 0.003). The cosine similarities based on the downsampled vectors occupy a much smaller range than both the full and the Baroni-based similarities. This should be kept in mind when interpreting the results. Also, the mean

9 Many thanks to Ingo Plag for not only suggesting this comparison but also for running it on a previous version of our data.

and median values based on the Baroni vectors are highest within this set, but still relatively low when compared to e.g. Lazaridou et al. (2013, 1522), who find mean cosine similarity values of 0.47 for base-derivative pairs in English. In other words, the absolute value of cosine similarities is not meaningfully comparable across methods and corpora. This contrasts with the correlations between the similarities across items, which allow us a direct comparison of the three vectorsets. We calculated the correlations between the cosine similarities from the three different sets, and also between these values and absolute frequencies of the bases and derivatives in our corpus. Only three correlations turn out to be significant. The downsampled vectors are positively correlated with the pre-downsampling vectors (Pearson's r=0.66, p = .015), and the pre-downsampling vectors are negatively correlated with the absolute number of occurrences of the bases (r=-0.67, p = .011). Most importantly, there is a very strong and highly significant positive correlation between the similarities based on the downsampled vectors and the similarities based on the Baroni vector space (r = 0.88, p = 6.675e-05).

We take this as clear evidence for the success of our downsampling method to eliminate the influence of differences in absolute frequency, and for the overall validity of our approach.

5.2.5 Statistical analysis

As described in Section 5.1, we compare cosine similarities involving four conditions: To answer RQ3a, we compared the cosine similarities between the conditions base-derivative pairwise and base-derivative across pairs. To answer RQ3b, we compared the derivative-derivative similarities to a) the base-base similarities and b) the base-derivative pairwise similarities. We used the Shapiro-Wilk test of normality on all our grouped similarity values. As not all of them are normally distributed, we used the non-parametric two-sample Wilcoxon test to compare the similarities across conditions. Note that this does not create a problem of multiple comparisons: For any constellation we are interested in, we consistently only make one comparison, namely comparison of cosine similarities.

For all comparisons, we are primarily interested in generalizing over the measures for all twelve lemmas we investigated, although we are using the same lemmas from the same three VerbNet classes that we used in Study 2. However, these classes are represented by only four lemmas each, which makes generalizations over specific classes difficult. While we will focus on all items, we will also report both the numbers and the significant patterns per verb class.

5.3 Results

The similarity values of interest are reported in Table 7 as two columns each for all items and per VerbNet class, the first column showing the mean cosine similarities, the second one the standard deviations within the similarities. The first two rows of pairings show the average similarities between bases and derivatives. The first row, base-derivative pairwise, only considers base-derivative pairs (e.g. SPRINT-OUTSPRINT) and quantifies the extent to which the meaning of the base is related to that of the derived form. The second row, base-derivative across pairs, quantifies the similarities of all bases and derivatives in the data except for base-derivative pairs (e.g. for RUN-OUTSPRINT). The last two rows show the base-base and derivative-derivative similarity measures.

Tab. 7: Cosine similarity measures across 12 *out*-lemmas from VerbNet's RUN, EXIST, and PERFORMANCE classes.

	All items		RUN		EXIST		PERFORMANCE	
Pairings	**SIM**	**SD**	**SIM**	**SD**	**SIM**	**SD**	**SIM**	**SD**
base-base	0.03	0.01	0.04	0.01	0.04	0.01	0.05	0.03
base-derivative across pairs	0.04	0.01	0.05	0.01	0.04	0.01	0.03	0.02
base-derivative pairwise	0.06	0.02	0.08	0.02	0.05	0.00	0.05	0.02
derivative-derivative	0.07	0.03	0.10	0.02	0.06	0.01	0.06	0.02

Figure 3 represents the results graphically by indicating via brackets the three comparisons between the conditions of interest here. The brackets are labeled with the p-values of the respective comparison.

As shown in the left panel, across all items base-derivative pairwise similarities are significantly higher than base-derivative across pairs (w=200, p = 1.907e-05). Although the measures for the individual VerbNet classes pull in the same direction, the difference is significant only for the RUN class (W = 44, p = .01319).

Across all items, derivative-derivative similarities are significantly more similar to each other than the bases are to each other (w = 536, p = 7.976e-14). These differences are also significant for the RUN and EXIST classes (W = 0, p = .002 and W = 3, p = .015, respectively).

Finally, across all items and within each individual class, derivative-derivative similarities are descriptively slightly higher than the similarities of base-derivative pairs, both in their means as well as their medians. However, this difference does not reach significance, neither across classes nor within any class.

5.4 Discussion

Our RQ3a asked whether base-derivative pairs are more similar to each other than the non-pairs. In line with our expectations, this turned out to be the case, and is reflected in higher cosine similarities for the base-derivative pairs in comparison to base-derivative pairless similarities.

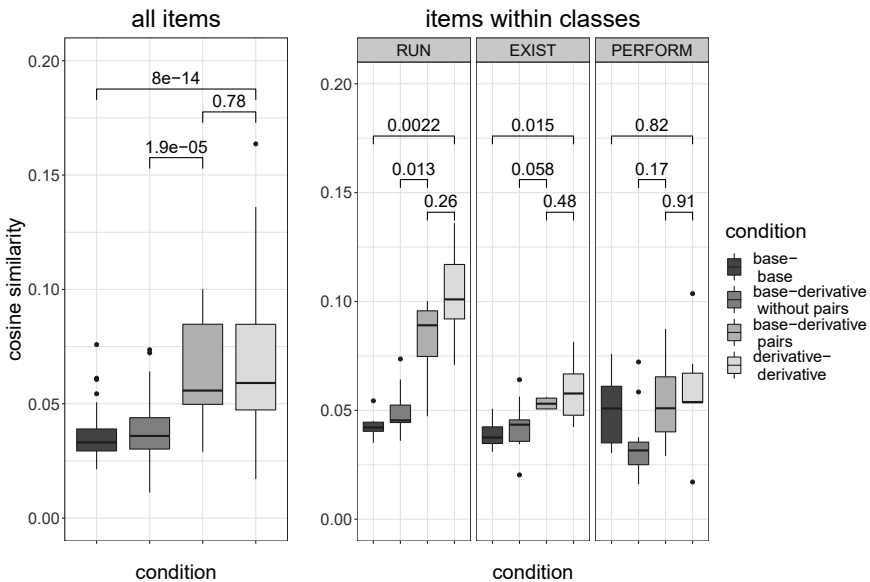

Fig. 3: Cosine similarities across all items and by semantic class in the four different conditions. The p-values for the target comparisons are given across the braces for the respective comparisons.

Turning to RQ3b, exploring differences between derivative-derivative, base-base, and base-derivative similarities, we find the overall highest similarities between derivatives. The derivatives are more similar to each other than the bases are to each other. Furthermore, descriptively, derivatives are more similar to each other than bases are to (their) derivatives. On the assumption that relatively high derivative-derivative similarities indicate a high degree of semantic coherence of the products of the word-formation process, this is strong support for the view that *out*-prefixation itself imposes a uniform semantics. Importantly, these findings are not evidence that derivatives do not inherit any features from their bases outside (sets of) suggested scalar dimensions (pace Ahn 2022). The way we set up our investigation does not allow any such conjecture.

One caveat is in place concerning the interpretation of our findings in terms of a relatively high degree of semantic coherence between derivatives: Recall that we did not distinguish between different senses of bases in either study presented here. This also holds for all vector representations of bases in Study 3. Thus, while we used random example sentences for all bases, their vectors represent generalizations over all senses. We cannot exclude the possibility that the word-formation process is selective with respect to the senses of a particular base form. In other words, a given *out*-lemma may inherit only a subset of the senses its base lemma encodes. This possibility is in line with post-hoc comparisons of the standard deviations of the similarity measures in Table 7 and the average cross-listing scores for the lemmas of the three classes (see Section 3.3). The PERFORMANCE-class has the highest cross-listing score (3.5), and the pertinent sample shows consistently high standard deviations. One reason for these consistently high deviations, relative to the other two classes, could lie in the a priori higher number of different senses the derivatives are based on.

As we saw in Studies 1 and 2, the members from different VerbNet classes differ in their preferred scalar dimensions when serving as bases to *out-*. The low number of lemmas per VerbNet class in Study 3 makes claims on class-specific distributional characteristics difficult. While the patterns that we find hold, at least descriptively, per individual verb class, they also hold upon comparing derivatives across classes, e.g. OUTRUN-OUTLIVE. In other words, the distributional similarities in this study do not point to a privileged status of the verb classes under consideration, but to a pronounced effect of the word formation process.

In summary, similarity measures show that *out*-prefixation has a relatively large distributional effect, in that *out*-derivatives show a relatively high degree of uniformity. While Studies 1 and 2 have shown that *out*-formations can be distinguished by base verb classes, Study 3 has shown that all three classes show similar distributional characteristics. However, it is unclear from this investigation whether this behavior is peculiar to *out*-verbs. The last study will address this question by gathering distributional data on four further English prefixes, including the spatial prefix *out-*.

6 Study 4: Distributional similarities – comparative *out*- contrasted with other prefixes

6.1 Rationale

In order to establish whether the relatively high similarity among derivatives is peculiar to comparative *out*-prefixation, we extended our investigation of distributional similarity measures to four further prefixes. We compare the patterns observed for comparative *out*- i) to a second prefix with bases from the same verb class (i.e. spatial *over*- and the RUN-class), ii) to spatial *out*-, and iii) to two different prefixes with bases from a different VerbNet class (reversative *un*-, iterative *re*-, and the TAPE-class). We chose these prefixes for comparison as they all show clearly weaker argument structural and event structural effects than comparative *out*- and thus presumably are more similar to their respective bases distributionally. Recall from Section 2.1 that comparative *out*-derivatives systematically license direct object arguments that are not licensed by their respective base verbs and systematically differ in event structure from their bases (for example ??*Mary sang John* vs. *Mary outsang John*).

As illustrated by means of the (base) verb *to fly* in (13) (both from iWeb), spatial *over*- (see Lieber 2004, ch.4) shows a pattern of preposition incorporation. The realization as a prefixed verb with a direct object in (13-b) is semantically identical to a base form with a PP-object as in (13-a), with the internal PP-argument corresponding to the direct object (see Wunderlich 2012). This constitutes a weaker form of argument structural effect than the one we observe for comparative *out*-. Overflying-events are still flying-events, and both of these events allow for the same type of locative argument, i.e., roughly, physical objects with some dimensional extension in space.

(13) a. Today, cotton growers hire companies with airplanes to **fly over the field**…
b. For airports with no AWOS or ASOS, **overfly the field** at or above pattern altitude and check the windsock.

Spatial *out*- behaves in similar ways to spatial *over*- with respect to argument structural effects. The examples in (14-a,b) with the (base) verb *to haul*, partly repeated here from section 2.1, illustrate that *out*- as a spatial prefix has a similar distribution as the synonymous spatial particle *out*, and licenses the same kinds of argument. Moreover, as shown in (14-c), *haul* also behaves similarly without any locative argument and all verbs in (14) denote hauling-events (all from iWeb):

(14) a. "haulback" means the cable used to **outhaul the rigging** or grapple when yarding...
b. They used a cable system to **haul out the ore**.
c. another driver has a job as a second truck had **to haul the extra freight**.

Reversative *un-* (cf. Bauer et al. 2013, 371ff.) and iterative *re-* (cf. Bauer et al. 2013, 419f.; Lieber 2004) do not display any clear effects on argument structure and license the same argument types as their respective bases. This is shown for both prefixes and the (base) verb *to seal* in (15) (all from iWeb). In terms of semantic type, resealing-events are clearly sealing-events, while the negation semantics of reversative *un-* always changes the base's event type.

(15) a. I recently did use white teflon tape to **seal a fuel pressure tester joint** [...]
b. [...] **unsealing a pressure tube** with very high reliability is necessary each time a fuel bundle is added or taken away.
c. Its only an hours work to **reseal a swing cylinder** if everything goes ok [...]

In this study, we thus compare different deverbal prefixed verbs, which differ regarding the strength of argument structural effects as well as the change in semantic type of the respective bases. To our knowledge, this kind of comparison has not been addressed in work on distributional semantics before. Primarily, we are interested in the following question:

Research question 4 (RQ4) Compared to comparative *out-*, are the presumably weaker semantic contributions and the relative lack of induced argument structural effects of the prefixes *over–*, spatial *out-*, *un-*, and *re–* reflected distributionally in lower derivative-derivative similarities?

6.2 Methodology

We used the same methodology as in Study 3. We extracted additional corpus attestations for the four additional prefixes, choosing four base lemmas from the same verb class for each prefix except for spatial *out-*. The base lemmas are listed by prefix and base class in (16).

(16) a. *over-* + RUN class:
OVERDRIVE, OVERFLY, OVERRUN, OVERSTEP
b. spatial *out-* (various VerbNet classes):
OUTCROSS, OUTGAS, OUTLOAD, OUTSTREAM
c. *re-/un-* + TAPE class:
(i) REFASTEN, RELOCK, RESEAL, REWIND
(ii) UNFASTEN, UNLOCK, UNSEAL, UNWIND

As in Study 3, all vectors for the base verbs were created from the ukWaC sentences, while all vectors for the derivatives were created from the iWeb sentences. The absolute occurrences of forms with the additional prefixes in iWeb are given in Table 8.

Tab. 8: Derivatives with the additional prefixes in iWeb: absolute number of occurrences.

***over*-verb**	**occurrences**	***out*-verb**	**occur.**	***re*-verb**	**occur.**	***un*-verb**	**occur.**
overdrive	37629	outcross	1047	refasten	412	unfasten	2092
overfly	1495	outgas	2586	relock	1517	unlock	302674
overrun	35212	outload	64	reseal	7732	unseal	8591
overstep	6740	outstream	759	rewind	25894	unwind	54349

The number of overall occurrences varies considerably, with *outload* occurring just 64 times. For all other derivatives, as well as for all bases, the same downsampling to 100 items as in Study 3 was applied. Recall that comparing our approach to the word embeddings from Baroni et al. (2014) already included these pairs (see Section 5.2.4), and showed that downsampling successfully eliminates effects of absolute frequencies.

6.3 Results

Table 9 contrasts the results for spatial *over-* (left two columns) with those for comparative *out-* from Study 3 (right two columns), both with bases from the RUN-class.

Figure 4 shows the corresponding box plots, again with the p-values for the three target comparisons as bracket labels (with the graph for comparative *out-* again corresponding to the one presented in Study 3).

Spatial *over-* shows the same pattern as comparative *out-* with respect to base-derivative pairs being more similar than base-derivative non-pairs (W=42, p

Tab. 9: Average cosine similarites for the verbs from the RUN class, contrasting spatial *over-* and comparative *out-*.

	RUN (+ *over*)		RUN (+ *out*)	
	SIM	**SD**	**SIM**	**SD**
base-base	0.04	.004	0.04	0.01
base-derivative across pairs	0.04	.005	0.05	0.01
base-derivative pairwise	0.06	0.03	0.08	0.02
derivative-derivative	0.05	0.01	0.10	0.02

= .02967). The two prefixes are also similar in that for *over-*, derivative-derivative similarities are also higher than base-base similarities (W=1, p-value = 0.004). Noticeably, in both cases the difference between the conditions for comparative *out-* is more pronounced, with the similarities for the two *out-* conditions clearly higher than those of the corresponding *over-* conditions. In addition, while mean (and median) similarities for comparative *out-* are highest between derivatives (higher than for base-derivative pairs), for spatial *over-* the mean of base-derivative pairs are most similar, followed by the derivative-derivative ones, with the medians showing the same pattern as *out-*. Neither difference is significant.

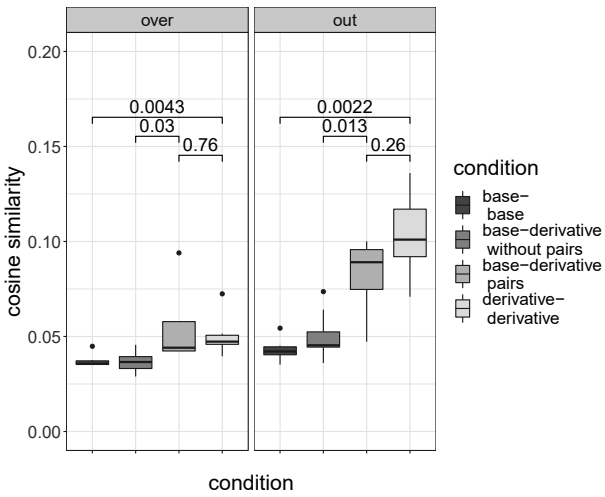

Fig. 4: Cosine similarites across the four different conditions for both *over-* and *out-* and the RUN class. The p-values for the target comparisons are given across the braces for the respective comparisons.

Table 10 contrasts the results for spatial *out-* with the across-classes results for comparative *out-* from Study 3 (see the 'all items' columns in Table 7). Figure 5 shows the corresponding figures with the p-values of the target comparisons.

Tab. 10: Average cosine similarites for spatial *out-* contrasted with comparative *out-*.

	spatial *out-*		**comparative *out-***	
Pairings	**SIM**	**SD**	**SIM**	**SD**
base-base	0.03	0.01	0.04	0.01
base-derivative across pairs	0.03	0.01	0.04	0.01
base-derivative pairwise	0.05	0.03	0.06	0.02
derivative-derivative	0.04	0.01	0.07	0.03

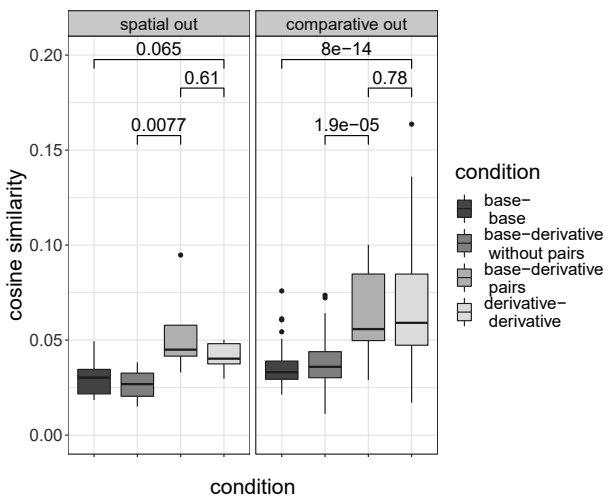

Fig. 5: Cosine similarites across the four different conditions for both spatial and comparative *out-*. The p-values for the target comparisons are given across the braces for the respective comparisons.

Similar to all results thus far, similarities of base-derivative pairs of spatial *out-* are higher than similarities of base-derivative non-pairs. This difference is significant (W = 45, p-value = 0.007692). Again, the base-base similarities are lower than the derivative-derivative similarities. This difference is not significant (W = 6, p-value = 0.06494). Importantly, just as for spatial *over-*, mean (and, in contrast to

spatial *over-*, also median) pairwise base-derivative similarities are higher than derivative-derivative similarities for spatial *out-*. The difference is not significant (W = 15, p-value = 0.6095). As shown before, for comparative *out-*, this difference, while also not significant, goes in the opposite direction both across all classes and for each class individually.

The similarites for *un-* and *re-* with the same bases from the TAPE-class are shown in Table 11.

Tab. 11: Average cosine similarities for the bases from the TAPE class and their *un-* and *re-* derivatives.

	TAPE (+ *un*)		TAPE (+ *re*)	
	SIM	SD	SIM	SD
base-base	0.05	0.01	0.05	0.01
base-derivative across pairs	0.05	0.01	0.05	0.01
base-derivative pairwise	0.08	0.05	0.10	0.04
derivative-derivative	0.06	0.01	0.09	0.02

The prefixes *un-* and *re-* show the by now familiar patterns. Pairwise base-derivative measures are on average higher than pairless ones. However, the difference is not significant, neither for *un-* nor for *re-*, (both W = 40, p-value = .05824). The derivative-derivative similarities are higher than the base-base similarities. This is significant only for *re-* (W =1, p-value = .004). Importantly, the patterns again diverge from comparative *out-* in that base-derivative pairs show the highest similarities (both on means and medians), followed by derivative-derivative similarities. Just as before, this difference is not significant for either prefix.

6.4 Discussion

Based on our results, it is a general property of English deverbal verbs with a given prefix that they are more similar to each other than their respective bases are to each other. Likewise, base-derivative pairs are more similar than base-derivative non-pairs for all morphological processes. These properties are expected given the idea that any derivational process is characterized by a semantic core, and while they show up only partly significantly, they hold descriptively across the board. Regarding RQ4, an interesting difference between the prefixes emerges elsewhere: for comparative *out-*, derivatives are more similar to each other than

base-derivative pairs. In this respect, the distributional behavior of *out-* clearly stands out among the prefixes we investigated.

This finding reinforces our interpretation that the word-formation process is by itself a highly pronounced semantic contributor to comparative *out*-derivatives, and arguably a more influential one than their respective bases. The most plausible reasons for the differences between prefixes are the change of semantic type and concurring argument structure alternation induced by comparative *out*-prefixation. In case this interpretation of the results of Study 4 is on the right track, it offers support for how we interpreted the findings of Study 3: the similarity differences we find between base-derivative pairs and across derivatives in *out*-prefixation follow from the prefix's applicative force, in particular with respect to licensed object arguments. At least relative to prefixes with different properties, this can be regarded as a peculiarity of comparative *out-*.

More generally, our results are of interest to distributional studies that deal with argument structure alternations. Recall that we are using distributional count models in this study (see section 5.2.4), and therefore cannot directly comment on prediction models. However, our findings that the one category that induces strong argument structural effects stands out is partly in line with Padó et al. (2016), who show that it is difficult to predict derivative semantics from the base for German derivational processes with argument structural effects. It is unclear to us, however, if there are other derivational processes in English with properties that are more similar to comparative *out-*'s and that would allow to further isolate argument structure as a core contributor to derivative similarity.

Finally, as remarked by one reviewer, a possibly confounding effect for our comparisons may be found in the differences of relative frequencies (see e.g. Hay 2001) and different degrees of lexicalization of the *out*-derivatives vis-à-vis the frequencies of the twelve additional prefixed lemmas. We chose our data largely on semantic grounds, and they do not allow for any conclusions in this respect. We can thus not exclude the possibility that relative frequency plays a role for the interpretation of Study 4.

7 Conclusion

In this paper, we have taken a decidedly quantitative approach for the investigation of a fundamental problem of derivational semantics, namely the semantic relatedness between different components of a word formation process. We have used English comparative *out*-prefixation as a testing ground and looked at the potential of bases to predict derivative semantics, at the relative uniformity of

bases and derivatives, at similarities both across bases and across derivatives, and at comparisons with a number of further prefixes with diverging characteristics. Most generally, our results back up assumptions of comparative *out-*'s relative semantic richness as a word formation process, as well as assumptions of semantic relatedness being gradient in nature (cf. Spencer 2010a,b, 2013).

Studies 1 and 2 have shown that the resolution of systematic underspecification of *out-*'s comparative semantics is, largely, predicted by both base lemmas and by the semantic class of base lemmas, and that predictive power correlates with measures of the ambiguity of the members of these classes. We are not aware of studies that investigate such a form of underspecification for morphological processes. The findings are in line, however, with studies that show that the semantics of bases are crucial for constraining polysemy in derivation (e.g. Aronoff & Cho 2001; Kawaletz 2021; Plag et al. 2018).

Using distributional measures, Study 3 has revealed quantitative reflexes of two separate features of *out*-prefixation. First, derivatives and their bases show relative high degrees of similarity. Second, and more importantly, similarities between derivatives are far more pronounced. This constellation is peculiar to *out-*, as comparisons with four other prefixes have shown in the final study. Likely, this difference can be attributed to differences in the semantic structures that different word-formation processes systematically add to their respective bases. While our findings on differences of semantic coherence between the products of word formation processes are in line with other studies (see e.g. Wauquier 2020), we are not aware of studies that quantitatively reveal differences between base-derivative and derivative-derivative similarities of this kind.

A number of open issues require further investigation. As shown by, for example, Marelli & Baroni (2015) and Padó et al. (2016), different distributional measures are better suited for different word-formation processes. The role of lexicalization for frequent patterns, including fixed multi-word expressions (such as *to outstay one's welcome*), is likely to interfere with general ideas of compositionally derived vector representations in derivational semantics. Moreover, in all of our studies, we have generalized over base forms and, deliberately, not controlled for different senses. Possibly, sense selection for bases may allow for more accurate similarity measures between bases and derivatives from a distributional perspective. We leave these open issues to future research.

Acknowledgements

The research reported on in this article has been funded by the German Research Foundation (DFG project PL 151/11-1 'Semantics of derivational morphology' at Heinrich-Heine-Universität Düsseldorf and SFB 833 'The Construction of Meaning: The Dynamics and Adaptivity of Linguistic Structures' at Eberhard Karls Universität Tübingen). We thank Ingo Plag and two anonymous reviewers for helpful comments on earlier versions of the paper, and Ann-Sophie Haan for annotation work.

Bibliography

Ahn, Byron. 2022. Mapping *OUT*- argument structure. *Syntax* (to appear).

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Aronoff, Mark & Sungeon Cho. 2001. The semantics of *-ship*. *Linguistic Inquiry* 32. 167–173.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3). 209–226.

Baroni, Marco, Georgiana Dinu & Germán Kruszewski. 2014a. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014* 238–247.

Baroni, Marco, Georgiana Dinu & Germán Kruszewski. 2014b. Don't count, predict! Semantic vectors [Data set]. https://doi.org/10.5281/zenodo.2635544.

Bauer, Laurie, Rochelle Lieber & Ingo Plag. 2013. *The Oxford reference guide to English morphology*. Oxford: Oxford University Press.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media Inc.

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

Bonami, Olivier & Matías Guzmán Naranjo. 2023. Distributional evidence for derivational paradigms. In Sven Kotowski & Ingo Plag (eds.), *The semantics of derivational morphology. Theory, methods, evidence* Linguistische Arbeiten, 219–259. Berlin: De Gruyter.

Bonami, Olivier & Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio* 17(2). 173–195.

Davies, Mark. 2018. The 14 billion word iWeb Corpus: Available online at https://corpus.byu.edu/iWeb/.

Haspelmath, Martin & Andrea D. Sims. 2013. *Understanding morphology. 2nd ed.* London and New York, NY: Routledge Taylor & Francis Group.

Hay, Jennifer. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 39(6). 1041–1070.

Herweg, Michael. 2020. Motion verb constructions, frames, and profiling. *Questions and Answers in Linguistics* 6(1). 1–30.

Hill, Felix, Roi Reichart & Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* 41(4). 665–695.

Kawaletz, Lea. 2021. *The semantics of English -ment nominalizations*. Düsseldorf: Heinrich-Heine-University Dissertation.

Kennedy, Christopher & Beth Levin. 2008. Measure of change: The adjectival core of degree achievements. In Louise McNally & Christopher Kennedy (eds.), *Adjectives and adverbs*, 156–182. Oxford: Oxford University Press.

Kennedy, Christopher & Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* (81). 345–381.

Kipper, Karin, Anna Korhonen, Neville Ryant & Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation* 42(1). 21–40.

Kisselew, Max, Sebastian Padó, Alexis Palmer & Jan Šnajder. 2015. Obtaining a better understanding of distributional models of German derivational morphology. In *Proceedings of the 11th international conference on computational semantics*, 58–63. London, UK: Association for Computational Linguistics. https://aclanthology.org/W15-0108.

Koefoed, Geert & Jaap van Marle. 2000. Productivity. In Geert E. Booij, Armin Burkhardt, Gerold Ungeheuer, Herbert Ernst Wiegand, Hugo Steger & Klaus Brinker (eds.), *1. Halbband: Ein internationales Handbuch zur Flexion und Wortbildung*, 303–311. Berlin: De Gruyter.

Kotowski, Sven. 2021. The semantics of English *out*-prefixation: A corpus-based investigation. *English Language and Linguistics* 25(1). 61–89.

Lapesa, Gabriella, Lea Kawaletz, Ingo Plag, Marios Andreou, Max Kisselew & Sebastian Padó. 2018. Disambiguation of newly derived nominalizations in context: A distributional semantics approach. *Word Structure* 11(3). 277–312.

Lapesa, Gabriella, Sebastian Padó, Tillmann Pross & Antje Rossdeutscher. 2017. Are doggies really nicer than dogs? The impact of morphological derivation on emotional valence in German. In *IWCS 2017 — 12th international conference on computational semantics — short papers*, https://aclanthology.org/W17-6922.

Lazaridou, Angeliki, Marco Marelli, Roberto Zamparelli & Marco Baroni. 2013. Compositional-ly 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)*, 1517–1526. Sofia, Bulgaria: Association for Computational Linguistics. https://aclanthology.org/P13-1149.

Lieber, Rochelle. 2004. *Morphology and lexical semantics*. Cambridge: Cambridge University Press.

Lieber, Rochelle. 2016. *English nouns: The ecology of nominalization*. Cambridge: Cambridge University Press.

Maguire, Phil, Barry Devereux, Fintan Costello & Arthur Cater. 2007. A reanalysis of the carin theory of conceptual combination. *Journal of Experimental Psychology. Learning, Memory & Cognition* 33(4). 811 – 821.

Maienborn, Claudia. 2019. Flexible Bedeutungszuordnung im Lexikon: Polysemie – Unterbestimmtheit – Uminterpretation. In Yasuhiro Fujinawa & Jiro Inaba (eds.), *Wie entsteht Bedeutung? Semantik zwischen Grammatik, Kognition und Kontext*, 11–44. München: Iudicium Verlag.

Marchand, Hans. 1969. *The categories and types of present-day English word-formation. 2nd ed.* München: C.H. Beck.

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

McHugh, Marry L. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica* 276–282.

McIntyre, Andrew. 2003. Preverbs, argument linking and verb semantics. In Geert Booij & Jaap Marle (eds.), *Yearbook of morphology 2003*, 119–144. Dordrecht and London: Kluwer Academic.

Mititelu, Verginica Barbu, Gianina Iordăchioaia, Svetlozara Leseva & Ivelina Stoyanova. 2023. The meaning of zero nouns and zero verbs. In Sven Kotowski & Ingo Plag (eds.), *The semantics of derivational morphology. Theory, methods, evidence*, 63–102. Berlin: De Gruyter.

Padó, Sebastian, Aurélie Herbelot, Max Kisselew & Jan Šnajder. 2016. Predictability of distributional semantics in derivational word formation. In *Proceedings of COLING*, 1285–1296. Osaka, Japan: The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1122.

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

Plag, Ingo, Marios Andreou & Lea Kawaletz. 2018. A frame-semantic approach to polysemy in affixation. In Olivier Bonami, Gilles Boyé, Georgette Dal, Hélène Giraudo & Fiammetta Namer (eds.), *The lexeme in descriptive and theoretical morphology*, 546–568. Berlin: Language Science Press.

Plag, Ingo, Lea Kawaletz, Sabine Arndt-Lappe & Rochelle Lieber. 2023. Analogical modeling of derivational semantics. Two case studies. In Sven Kotowski & Ingo Plag (eds.), *The semantics of derivational morphology. Theory, methods, evidence* Linguistische Arbeiten, 103–141. Berlin: De Gruyter.

Rainer, Franz, Wolfgang U. Dressler, Francesco Gardani & Hans Christian Luschützky. 2014. Morphology and meaning: An overview. In Franz Rainer, Francesco Gardani, Hans Christian Luschützky & Wolfgang U. Dressler (eds.), *Morphology and meaning*, 3–46. Amsterdam and Philadelphia: John Benjamins.

Rappaport Hovav, Malka & Beth Levin. 1992. *-er*-nominals: Implications for a theory of argument structure. In Tim Stowell & Timothy Stowell (eds.), *Syntax and the lexicon*, 127–153. New York, NY: Academic Press.

Rappaport Hovav, Malka & Beth Levin. 2001. An event structure account of English resultatives. *Language* 77(4). 766–797.

Schröder, Anne. 2011. *On the productivity of verbal prefixation in English: Synchronic and diachronic perspectives*. Tübingen: Narr.

Schäfer, Martin. 2020. Distributional profiling and the semantics of modifier classes. In Christopher Pinon & Laurent Roussarie (eds.), *Empirical issues in syntax and semantics 13*, 139–166. Paris: CSSP.

Schäfer, Martin. 2023. Splitting *-ly*s: Using word embeddings to distinguish derivation and inflection. In Sven Kotowski & Ingo Plag (eds.), *The semantics of derivational morphology. Theory, methods, evidence* Linguistische Arbeiten, 261–296. Berlin: De Gruyter.

Solt, Stephanie. 2015. Measurement scales in natural language. *Language and Linguistics Compass* 9(1). 14–32.

Spencer, Andrew. 2010a. Factorizing lexical relatedness. In Susan Olsen (ed.), *New impulses in word-formation*, 133–172. Hamburg: Helmut Buske Verlag.

Spencer, Andrew. 2010b. Lexical relatedness and the lexical entry – a formal unification. In Stefan Müller (ed.), *Proceedings of the 17th international conference on head-driven phrase structure grammar*, 322–340. Stanford: CSLI Publications.

Spencer, Andrew. 2013. *Lexical relatedness*. Oxford: Oxford University Press.

Talmy, Leonard. 2000. *Toward a cognitive semantics. Vol.II: Typology and process in concept structuring*. Cambridge, MA: MIT Press.

Tolskaya, Inna. 2014. *Verbal prefixes: Selection and interpretation*. Tromsø: University in Tromsø Dissertation.

Varvara, Rossella. 2017. *Verbs as nouns: Empirical investigations on event-denoting nominalizations*. Trento: University of Trento dissertation.

Varvara, Rossella, Gabriella Lapesa & Sebastian Padó. 2021. Grounding semantic transparency in context. *Morphology* 31. 1–38.

Wauquier, Marine. 2020. *Confrontation des procédés dérivationnels et des catégories sémantiques dans les modèles distributionnels*. Toulouse: University of Toulouse Dissertation.

Wunderlich, Dieter. 2012. 84. Operations on argument structure. In Claudia Maienborn, Klaus von Heusinger & Paul Portner (eds.), *Semantics* Handbücher zur Sprach- und Kommunikationswissenschaft, 2224–2259. Berlin: De Gruyter.

Zwarts, Joost. 2008. Aspects of a typology of direction. In Susan D. Rothstein (ed.), *Theoretical and crosslinguistic approaches to the semantics of aspect*, 79–105. Amsterdam: John Benjamins.