Martin Schäfer

Splitting *-ly*'s: Using word embeddings to distinguish derivation and inflection

Abstract: Whether the relation of base and *-ly* forms in seemingly regular pairs like *quick/quickly* and *clever/cleverly* represents an instance of derivation or inflection has been widely debated. This paper starts from the observation that these two cases are actually not parallel, hypothesizing that pairs from the class of SPEED adjectives show features of inflection, but pairs from the class of HUMAN PROPENSITY adjectives show features of derivation. This hypothesis is explored using distributional semantics, employing different embeddings across three studies. Results show that the classes show a clear distributional difference in terms of the contribution of *-ly*, but the difference does not obtain across all studies, and not always as expected, indicating that a binary contrast inflection/derivation is too simplistic.

Keywords: derivation, inflection, semantics, distributional semantics, adjective, adverbial, English

1 Introduction

Whether the relation of bases and *-ly* forms in pairs like *quick/quickly* in English represents an instance of derivation or inflection has been widely debated, with Bauer et al. (2013, 536) "concluding that the evidence is inconclusive". As far as semantics is concerned, the relationship between the two forms is seen as completely regular (a few opaque exceptions aside), with *-ly* carrying no lexical meaning (Plag 2018, 200, Giegerich 2012). This complete regularity is surprising when considering that the items forming the pairs involved come from very different areas, and that the prototypical usage of the base form as attributive modifier and of the *-ly* form as adverbial modifier usually involves the combination with heads whose referents come from very different domains. To illustrate, take the examples in (1), with (1-a) showing pairs from the SPEED class and (1-b) showing pairs from the HUMAN PROPENSITY classes (for these classes, cf. Dixon 1982).

(1) a. quick/quickly, swift/swiftly
b. jealous/jealously, clever/cleverly

Martin Schäfer, Eberhard-Karls-Universität Tübingen, email: post@martinschaefer.info

https://doi.org/10.1515/9783111074917-009

Adjectives from the SPEED class in their attributive usage are typically combined with nouns referring to events, and in their adverbial usage, the verbs they combine with typically also refer to events. In contrast, adjectives from the HUMAN PROPENSITY class are typically combined with nouns referring to human referents, whereas verbs never have human referents. Intuitively, it therefore seems that going from usages of the base form to usages of the *-ly* form involves much less adjustment for the pairs in the SPEED class than for the pairs in the HUMAN PROPENSITY class.

In this paper, I argue that a closer look at the semantics involved suggests that the SPEED pairs show typical properties of inflection and the HUMAN PROPENSITY pairs typical properties of derivation. Three studies explore the extent to which this idea is reflected quantitatively when using distributional semantics measures.

2 Background

To set the stage for the empirical studies using distributional semantics, Section 2.1 outlines the relevant connections between adjective classes and the usage of their members as adverbial modifiers, Section 2.2 introduces distributional semantics and looks at approaches to the distinction between inflection and derivation in distributional semantics, and Section 2.3 outlines the expectations and hypotheses for the following studies.

2.1 Adjective classes and adverbial modification

That the relationship across base/base-*ly* pairs differs depending on the semantics of the adjectives involved is already discussed in Dixon (1982). He distinguishes seven semantic types for the word class 'adjective' in English: DIMENSION, PHYSICAL PROPERTY, COLOR, HUMAN PROPENSITY, AGE, VALUE, and SPEED. Discussing their properties as derived adverbs (under which term he also subsumes the *-ly* forms), he notes that only three of them, HUMAN PROPENSITY, VALUE, and SPEED, have the same, non-metaphorical lexical meaning as the adjective. In terms of their adverbial functions, HUMAN PROPENSITY forms occur as manner and sentence adverbials, and can also modify adjectives, cf. the examples in (2), drawn from the Corpus of Contemporary American English (COCA, Davies 2008–).

(2) a. When Nixon cleverly halted the draft of 18-year olds in the early 70s, that took the backbone out of the anti-war movement [...] [COCA]

b. This team was very cleverly assembled by Checketts. [COCA]
c. The monster itself is unique in design and cleverly ambiguous, making it all the more scary. [COCA]

Cleverly in (2-a) functions as a sentence adverbial, giving rise to paraphrases like *It was clever of Nixon that he halted the draft* .... In (2-b), *cleverly* functions as a manner adverbial (...*was assembled in a clever manner*). The semantic function of *cleverly* in (2-c) seems somewhat similar to that of a manner adverbial, but since no explicit eventive target is given, contextual contextual information is needed for successful interpretation, yielding plausible readings like *designed in a clever way to look ambiguous*.

VALUE forms can also modify adjectives, but otherwise occur only as manner adverbials. And SPEED forms only occur as manner adverbials, cf. (3).

(3) With an easy payment system, even a small business can quickly process payments. [COCA]

This study focuses on just the HUMAN PROPENSITY and the SPEED class, not only because they provide the clearest contrast within the three groups with non-metaphorical lexical meanings already on Dixon's criteria, but because they also show the clearest semantic differences: HUMAN PROPENSITY adjectives are predicates that apply prototypically to humans, that is, physical objects, while SPEED adjectives are prototypical predicates of events (Pustejovsky 1995, Bücking & Maienborn 2019, Schäfer 2021).

This difference places them at opposite ends when it comes to the prototypical functions of the base forms and the *-ly* forms: The prototypical usage of adjectives in their base forms is their usage as attributive modifiers in nominal modification. Prototypical noun referents, in turn, are physical objects. In contrast, the prototypical usage of *-ly* forms is their usage as verb modifiers, and prototypical verb referents are events.

In terms of semantic fit, it is straightforward for the SPEED class to modify verbs, because the adjectives are event predicates to begin with. In contrast, this is not straightforwardly possible for members of the HUMAN PROPENSITY class: there needs to be an intermediate step that allows one to connect them to events. This difference becomes apparent when taking a closer look at a few examples from each class.

Before turning to this difference, a quick reminder that both adjective classes also share an important property since they both are subsective adjectives whose interpretation is always relative to some scale or measure provided either by the linguistic or the extra-linguistic context: a quick answer by a child may take more time than a slow answer of an adult, and clever children are judged as such in

comparison to the relevant age-group (see Schäfer 2018, 79–81 for discussion and literature).

So, where does the difference between the classes show up? Let's start with members of the SPEED class, cf. (4), and features related to them being event predicates.

(4) a. to run quickly/swiftly
b. to answer quickly/swiftly

The combination between verb and *-ly* form gives rise to transparent and regular combinations: the event as a whole or subparts thereof only took a short amount of time. Depending on whether they relate to subparts of events or to events as a whole, or to the position of an event relative to a point in time they may receive different readings, but these readings are closely connected. In all cases, SPEED adjectives are directly connected to events, not to the 'way' or manner of an event. Further, across the usages, they target the same, temporal, dimension (hence the label 'one-dimensional' in Schäfer 2013, 55–57), and are consequentially given uniform analyses in the theoretical literature (cf. Rawlins 2013 and Koev 2017 on *quickly*).

As Schäfer (2021) shows for a subset of speed adjectives considered here (*quick, rapid, slow, speedy* and *swift*), speed adjectives in their prototypical attributive usages combine with nouns that refer to events. This preponderance of attributive combinations with heads referring to events is reflected in the existence of many instances where morphologically related heads occur across both forms, cf. (5).

(5) a. quick glance/to glance quickly
b. swift movement/to move swiftly

Importantly, the events referred to are in both cases the same, only the realization via a nominal vs a verbal construction differs.

Combination of HUMAN PROPENSITY forms with verbs, in contrast, differ in all these respects from the SPEED adjectives: First, they do not give rise to transparent and regular combinations, cf. (6).

(6) a. to run cleverly/jealously
b. to answer cleverly/jealously

It is relatively open which events would count as instances of clever or jealous running, and the same holds for answering events. This holds even more across the two events of running and answering: the actual properties of the events that lead to them being assessed as reflecting cleverness or jealousy might be very different.

Cooccurrence with morphologically related heads is possible, but only for those cases where the head nouns refer to events, cf. (7).

(7) a. jealous taunt/to taunt jealously
b. clever answer/to answer cleverly

But prototypical attributive usages of HUMAN PROPENSITY adjectives combine with head nouns that refer to humans. By their very nature, these do not come with verbal counterparts, cf. (9).

(8) a. jealous husband/?
b. clever girl/ ?

That *-ly* forms from the HUMAN PROPENSITY class are one step removed from the events they modify is also evident in the paraphrases for their usages as sentence and manner adverbials. For the sentential usage, the standard paraphrase *It was ADJ of X that ...* clearly ties the meaning contribution of the *-ly* form to a person. The paraphrases for the manner usage require the introduction of the underspecified way/manner placeholders *in an X manner/The way that ... was ADJ*. Again, what exactly counts as a clever/jealous manner or way is left open, and might, in addition, involve several aspects and dimensions of the event involved. Both adverbial usages are clearly distinct from each other. In fact, in a related language like German, which generally shows much similarity to English in its modification system, the two adverbial usages are typically realized by different forms, with the sentential usage employing the derivational suffix *-erweise*.

These clear differences between the two classes are behind the core idea of this paper: when a member of the SPEED class is used adverbially, it can be directly combined with the verb, no implicit intermediate step is necessary, the process is semantically regular and *-ly* only marks the grammatical function. In other words, it has all the hallmarks of a standard inflectional relationship. In contrast, for members of the HUMAN PROPENSITY class, an intermediate step is needed, the relationships between bases and *-ly* forms are not semantically regular, and *-ly* does not only mark grammatical function: these processes are thus similar to typical cases of derivation.

2.2 Derivation vs. inflection in distributional semantics

This paper explores two ideas to get at the derivation vs. inflection difference in distributional semantics. One rests on the link between derivation and inflection on the one hand and the stability of contrast on the other hand, following Bonami

& Paperno (2018). The second strain exploits different implementations of word embeddings to investigate the issue. Before introducing these two approaches in more detail, the next section provides a short introduction to distributional semantics and studies using it to investigate morphology.

2.2.1 Distributional semantics

Distributional semantics is based on the idea that the context in which a word appears characterizes its meaning. This approach easily lends itself to assessments of the meaning similarity of words; Sahlgren (2006, 21) captures this in his formulation of the distributional hypothesis: "Words with similar distributional properties have similar meanings".

Key to the computational implementation of this idea is the step of encoding the distribution of a word by means of a vector in geometrical space. We can illustrate this with the help of a toy example where we build vector representations for the three *-ly* forms *quickly*, *rudely* and *swiftly* by considering their occurrences in three different contexts: cooccurring with either the verbs *move*, or *decide*, or *follow*. Let's assume we tabulated the cooccurrences as in table 1, where the first cell tells us that *quickly* cooccurred 4 times with *move*.

Tab. 1: Toy example illustrating the first step in creating a simple distributional model: collecting cooccurrence counts for the target words, here the cooccurrences of the three *-ly* forms with the three verbs *follow*, *decide*, and *move*.

	move	decide	follow	total
quickly	4	3	1	8
rudely	2	1	4	7
swiftly	3	4	1	8
total	9	8	6	23

We now take the cooccurrence counts to represent vectors, so that the vector (4,3,1) represents the word *quickly*. Figure 1 illustrates how these vectors can be mapped into geometrical space: the context words, here *move*, *decide*, and *follow*, represent the dimensions of the vector space, and the cooccurrence counts with the *-ly* forms determine the length and directionality of three vectors, each of which encodes the distribution of the corresponding lexeme.

This simple three-dimensional space also allows us to demonstrate the most common measure to assess the semantic similarity between words: the cosine sim-

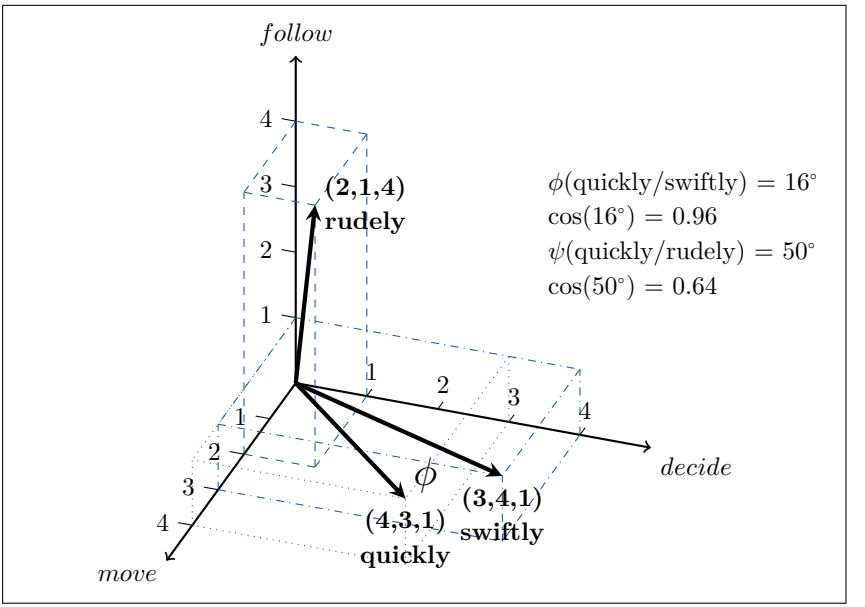

Fig. 1: A three-dimensional space showing the distributional vectors of *quickly*, *swiftly* and *rudely* based on the toy data in table 1. The three dimensions stand for the cooccurrences with the three verbs *move*, *follow*, and *decide*.

ilarity. The cosine similarity is simply the cosine of the angle that holds between any two vectors. In the example, the angle ϕ between the vector for *quickly* and the vector for *swiftly* is 16°, taking the cosine yields the cosine similarity of 0.96, very close to the highest possible value of 1. In contrast, the vectors of *quickly* and *rudely* are far less similar on this measure, the angle ψ of 50° resulting in a cosine similarity of 0.64.

Building on this simple model, it will become clear that this approach can now be fine-tuned in myriad ways. To start with, the context used for the cooccurrence counts can be varied. This parameter, often referred to as context window, can be set to anything from whole documents, paragraphs and sentences to just five, four, three, or even one word either to either side or one side of the target word. Further, the corpora used might be morphologically or syntactically pre-processed (e.g., part-of-speech-tagged, lemmatized, or parsed into dependency trees).

A further parameter is the context words themselves. Already early implementations of distributional semantics were based on a very large number of dimensions, for example the top 10,000 content words of a language, but other numbers are possible. And instead of raw cooccurrence frequencies, some sort of normal-

ization is usually employed. A common method is the transformation of the raw counts into pointwise mutual information or similar measures. For example, the numbers in table 1 can be transformed from their raw frequencies to pointwise mutual information (with logarithm) by using the probabilities of the occurrences of the words involved in the corpus, cf. the formula in (9) (see Turney & Pantel 2010; this normalization is also used in Kotowski & Schäfer 2023 in this volume).

(9) $$pmi_{ij} = log\left(\frac{p_{ij}}{p_i p_j}\right)$$

Using our toy example from above and taking the table to represent our whole corpus, the probability of encountering *quickly* in the context of *move* is 4/23, the overall probability of encountering our target word *quickly* is 8/23, and the probability of encountering the context word *move* is 9/23, leading to a pmi value of 0.35, cf. (10)

(10) $$pmi_{\text{quickly/move}} = log\left(\frac{p_{\text{quickly/move}}}{p_{\text{quickly}} \times p_{\text{move}}}\right) = log\left(\frac{4/23}{(8/23) \times (9/23)}\right) = 0.35$$

A further common manipulation is the application of dimension reduction techniques by which the number of dimensions is reduced to smaller numbers, often 300 dimensions (cf. **L**atent **S**emantic **A**nalysis, Dumais 2004). Since the publication of the word2vec algorithm (Mikolov et al., 2013), word vectors are most often created via machine learning, and many thus trained vector sets are freely available. In a comparative study by Baroni et al. (2014), these vectors, often called embeddings, on average outperform the count models (cf. also there for comparisons of the effects of different parameter settings of many of the other parameters mentioned above). In this paper, three different sets of such pretrained vectors are used, two sets from Levy & Goldberg (2014), and, in Study 3, a set of vectors trained with fastText (Mikolov et al., 2017).

Within distributional semantics, there has been clear focus on word formation, either via derivation or via compounding. For an overview on work on derivation, cf. Boleda (2020) and, in this volume, Kotowski & Schäfer (2023) and Bonami & Guzmán Naranjo (2023). Reddy et al. (2011) is an early study comparing different means of vector composition and their relation to semantic transparency in English compound nouns. Vector composition in this context refers to various ways in which two vectors can be combined into a new vector, with perhaps the simplest way being vector addition: two vectors are combined by simply adding up their numbers. Taking again *quickly* and *swiftly* from our toy example above, vector addition would yield a new vector quickly+swiftly of ((4 + 3),(3 + 4),(1 +1)) = (7,7,2) . A slew of recent studies on compound nouns in English and German combines ideas

from previous work on derivation in distributional semantics (Marelli & Baroni, 2015) with psycholinguistic approaches to conceptual combination (Marelli et al. 2017 and Günther & Marelli 2021 for English and Günther et al. 2020 for German). Inflection itself has not been focused on so much. Mikolov et al. (2013), the study introducing the word2vec algorithm, used a test set with words related by five semantic and nine syntactic relations to validate their approach. The syntactic part contains five different inflectional relations across the whole spectrum available in English (positive forms of adjectives compared to the corresponding comparative and superlative forms, plain forms of verbs compared to their *-ing*, past tense, and 3rd person sg forms, and nouns in the singular to their plural forms), as well as *-ly*, the affix of interest here. The analogy task they used will also be used in this paper, see Section 4. Together, these works convincingly show that both derivational as well as inflectional relationships can be captured with the help of distributional semantics. However, none of these works directly targets the difference between inflection and derivation. This is the focus of Bonami & Paperno (2018), to which we now turn, followed by the discussion of Levy & Goldberg (2014), whose work I will also use to explore possible differences in the behavior of base/*-ly* pairs.

2.2.2 Derivation and inflection in terms of stability of contrast

Bonami & Paperno (2018) provide a very careful operationalization of the differences between inflection and derivation in terms of quantitative aspects accessible through distributional semantics measures. They start out by discussing the list of five criteria from Stump (1998, 14–18), cf. (11), and it is helpful to go through this list again here for the special case of *-ly* forms.

(11) a. change in lexical meaning or part of speech
b. syntactic determination
c. productivity
d. semantic regularity
e. closure

Plag (2018, 200-201), who also goes through these criteria for *-ly*, acknowledges some exceptions (e.g. *hard*/*hardly*), but holds that pairs like *slow/slowly* and *aggressive/aggressively* show no change in lexical meaning. Is there a part of speech change? Plag argues that this could be called into question on the grounds that there are theories taking adjectives and adverbs to belong to the same underlying category. But even if one accepts that there are two distinct wordclasses adjective and adverb, this criterion is still not helpful for the base/*-ly* pairs, because the an-

swer clearly hinges on the overall decision inflection/derivation. Since inflection by definition never leads to a part of speech change, an analysis of the relation between base and *-ly* form as inflection amounts to no change in part of speech, the *-ly* form would count as just a further adjectival form. In contrast, when one adheres to the traditional classification of forms like *slowly* and *aggressively* as belonging to the separate lexical category of adverbs, there is of course a part of speech change. But since this issue, that is, whether the relation between base and *-ly* forms resembles one of inflection or one of derivation, is exactly what we are investigating here, the criterion of part-of-speech-change by itself cannot be used as a diagnostic. For syntactic determination, Plag points to the requirement to use the *-ly*-form in adjectival and verbal modification, and, in contrast, the unacceptability of the *-ly* form in attributive position.

As far as productivity is concerned, he notices only few exceptions (**fastly*, **goodly*), and semantic regularity, discussed by him in terms of semantic transparency, is again given, save exceptions like the above-mentioned *hardly*. Finally, *-ly* closes the word for further derivational processes, assuming that comparative and superlative formation are themselves instances of inflection. While Plag (2018) concludes that most points speak against a classification as a derivational suffix, he notes that this data shows that the distinction is not categorical. Taking a stronger stance, Giegerich (2012) argues that the evidence is solidly in favor of an inflectional analysis. Payne et al. (2010) are proponents of a derivational analysis. While my focus here will be on the semantic aspects, Payne et al. (2010) show that the criteria of syntactic determination, productivity, and closure are in fact not so clear-cut guiding posts to distinguish between inflection and derivation.

Bonami & Paperno (2018) point out that Stump's criteria are "formulated in terms of high-level morphological notions that are not easy to operationalize". Of special interest for the present paper is their point that it is unclear how to decide on what is lexical vs. not-lexical in meaning, and what counts as semantic regularity. They focus on this last point and provide an operationalization of the semantic regularity criterion in terms of stability of contrasts, as given in (12) (= Bonami & Paperno's (2)):

(12) **Stability of contrast**: the morphosyntactic and semantic contrasts between pairs of words related by the same *inflectional* relation are more similar to one another than the contrasts between pairs of words related by the same *derivational* relation.

Bonami & Paperno explored this criterion by looking at sets of triplets of <pivot, inflectionally related form, derivationally related form> in French. They found that overall the contrasts between inflectionally related forms were more stable.

Approaching a single affix, here *-ly*, in terms of derivation and inflection requires a slight adaption of their approach, since I am interested in whether the stability of contrast is different across *ly*-pairs from different semantic classes.

2.2.3 Inflectional and derivational contrasts in terms of topical and functional similarity

The second approach I use to explore the difference between the lexical classes exploits differences between different types of word embeddings, one focusing topical similarity, and one focusing functional similarity. Levy & Goldberg (2014) compare word embeddings trained using a standard bag-of-words SkipGram model on the English Wikipedia as a corpus with embeddings arrived at by using a generalization of that model trained on a dependency-parsed version of the same corpus. They note that "[...], the bag-of-words nature of the contexts in the 'original' SkipGram model yield *broad topical similarities*, while the dependency-based contexts yield more *functional similarities* of a *cohyponym* nature" Levy & Goldberg (2014, 303). They "expect the syntactic contexts to yield more focused embeddings, capturing more functional and less topical similarity." The reason for this expectation is that bag of words embeddings (henceforth: **bow**) ignore the syntactic function of context words, while dependency based embeddings (henceforth: **deps**) take the syntactic function into account.

In their own qualitative evaluation, they compare the five most similar words, the nearest neighbors, of a set of six target words. If there are differences between the embeddings, the **bow** embeddings usually find associates, and **deps** embeddings find words that behave similarly. For example, the top five nearest neighbors for the adjective *object-oriented* are exclusively other adjectives on the **deps** embeddings, while on the two **bow** embeddings tested they contain two nouns each. In a quantitative investigation, Levy & Goldberg used the WordSim353 dataset (Finkelstein et al., 2002), split for similarity and relatedness (Agirre et al., 2009). Among the topmost similar pairs in this dataset are mostly synonyms, e.g. *midday/noon*, *money/cash*, and *journey/voyage*. In other words, items that have the same meanings and can be used in the same syntactic contexts. Examples for highly related pairs in this dataset are *environment/ecology*, *money/bank*, and *law/lawyer*. These pairs are highly associated with each other but clearly differ in meaning. They can often not be used in the same syntactic contexts. **Deps** embeddings have a clear tendency to rank the similar words higher (in terms of the cosine similarity of the pairs) than **bow** embeddings.

How does this relate to the *-ly* pairs of interest here? The link becomes obvious when taking clear cases of inflection and derivation, and the first criterion from

Stump's list, cf. (11), change in lexical meaning or part of speech. Word pairs related by derivation are typically semantically associated, but they are not semantically similar. Prototypically, their meanings and their part of speech changes (*law*/*lawless*), they cannot occur in the same syntactic contexts and are in this way functionally different. In contrast, inflectionally related forms keep the same meaning and are, in terms of semantic similarity, on par with synonyms. That means that on **deps** embeddings, wordpairs related by inflection should be more similar to each other than wordpairs related by derivation in comparison to the same word pairs on **bow** embeddings. Consequently, if the hypothesis is correct that the relation between pairs from the SPEED class resembles inflection and the relation between pairs of the HUMAN PROPENSITY class resembles derivation, the pairs are expected to show marked differences when comparing the similarities across the two types of embeddings.

2.3 Expectations and hypotheses

Because I expect the SPEED pairs to show characteristics typical of inflection and the HUMAN PROPENSITY pairs to show characteristics typical of derivation, I hypothesize that

1. Regardless of the embeddings used, the SPEED pairs show stable contrasts, while the HUMAN PROPENSITY pairs do not show stable contrasts.
2. Any contrast between the two classes will in general be clearer on the **deps** embeddings than on the **bow** embeddings.

These expectations are explored in two studies. Study 1 focuses on similarities and correlations within the classes, and Study 2 focuses on the performance of the pairs in the two classes in an analogy task. These two studies are complemented by a third study that uses distributional vectors containing information about the internal make-up of the words represented by the vectors.

3 Study 1: Within-class similarities and correlations in bow and deps embeddings

The first study explores differences relating to the stability of contrast by comparing the two adjective classes in terms of:

a) their variation in average similarity and
b) their pair internal correlations.

In both cases, differences between the **bow** and **deps** embeddings are additionally considered.

Hypothesizing that HUMAN PROPENSITY pairs behave like derivationally connected forms and SPEED pairs like inflectionally related forms, the expectation is that the latter show stable contrasts in direct comparison. More concretely, I predict:

a) Stability of contrasts across pairs
   On average, the relationship between base and *-ly* forms is more stable for pairs from the SPEED class. The semantic similarities between the members of each pair should be more similar for the SPEED class. Therefore, the variation around the mean similarities should be lower for these pairs than for the pairs from the HUMAN PROPENSITY class.
b) Stability of contrasts for pairs within their classes
   Class internal similarities should be kept intact for the SPEED class, but not for the HUMAN PROP class. For example, if *quick* within its class is most similar to *rapid*, then I expect that similarly *quickly* is within its class most similar to *rapidly*. This expectation follows from the stability of contrasts: since the shift in the semantic space is expected to be similar across the pairs, the relative similarities should remain the same.

Both predicted results are expected regardless of the embedding, but I expect a clear difference between the two embeddings with regard to each other: For hypothesis a), I expect that when directly comparing the similarities between the pairs on the **deps** embeddings with those on the **bow** embeddings, the SPEED pairs are more similar, but the HUMAN PROP pairs are less similar. This expectation comes from the focus on functional similarity of the **deps** embeddings and the complementary focus on topical similarity of the **bow** embeddings. Since the pairs of the SPEED class are expected to show characteristics of inflection, they should be more similar on the **deps** embeddings then on the **bow** embeddings. For hypothesis b), the contrast between the two classes is expected to be more clear on the **deps** embeddings. Again, this follows from the functional vs. topical focus of the two embeddings and the hypothesis that the SPEED class pairs show characteristics of inflection. With their focus on functional similarity, the **deps** embeddings should show even less variation in the rankings across the pairs for the SPEED class, while there should be more variation across the pairs in the HUMAN PROPENSITY class, again in comparison to the **bow** embeddings.

3.1 Study 1: Materials and techniques

To represent the two adjective classes, samples of 11 pairs of SPEED and 11 pairs of HUMAN PROPENSITY base/*-ly* pairs were used, cf. table 2 for the 22 base forms.

Tab. 2: Adjective base forms by class

adjective class	adjectives
SPEED	*brisk, hasty, hurried, prompt, quick, rapid, rhythmical, slow, speedy, sudden, swift*
HUMAN PROP	*clever, cruel, eager, generous, greedy, happy, intelligent, jealous, proud, rude, shrewd*

The sample contains the maximum number of pairs that fall into the SPEED class; Dixon himself just gives *fast*, *quick*, and *slow* as examples and speaks of a few more (*fast* is not used here because it does not form a base/*-ly* pair). Extending the class to 11 was achieved by selecting similar terms using thesauri. Note that this procedure led to the inclusion of items like e.g. *rhythmical* and *sudden*, which at first might strike one as odd members of this class. But considering the underlying semantics, both fit quite well: *rhythmical* is clearly targeting only the temporal dimension of an event, or more specifically, of subevents within a larger event, and *sudden* likewise targets the temporal dimension, this time not of subparts of the event, but of an event in relation to a contextually given point in time. Both of these aspects also occur in usages of *quick*, which served as a prototypical example above: a quick decision often refers to a decision made after a short amount of time relative to a contextual reference point, it is here similar to typical usages of *sudden*, and a quick beat refers to a beat with sequences repeating after short intervals in time, similar to *rhythmical* in targeting repeated subevents. To keep the sample sizes comparable, the HUMAN PROPENSITY pairs were also capped at 11 (of these, seven are already listed by Dixon).

Two sets of pre-trained vectors from Levy & Goldberg (2014)[1] were used (cf. their paper for further details):

– **bow**: 183,870 embeddings trained on the English Wikipedia using a 5-word window and the original word2vec skip-gram implementation, with the negative sampling parameter set to 15, and 300 dimensions.

1 See https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/.

– **deps:** 174,015 embeddings trained using a modified skip-gram algorithm, trained on the dependency-parsed version of the same corpus, again with 300 dimensions.

Levy & Goldberg (2014) also provide embeddings using a 2-word window. As their own study shows that the results from using these fall in-between the 5-word window and the dependency trained vectors in their qualitative and quantitative assessments, they are not considered here, as I am interested in maximal differentiating behavior.

To assess the similarity of two vectors, cosine similarity is used. All calculations of cosine similarities and vector manipulations were done with Python, all statistical analysis was done with R. All code used in producing the analyses here and in Studies 2 and 3 is available here: https://doi.org/10.6084/m9.figshare.19093496.v2.

3.2 Study 1: Results

3.2.1 Hypothesis a): contrasts across pairs

Table 3 shows the average similarities along with the standard deviations between the pairs in the SPEED and HUMAN PROPENSITY classes.

Tab. 3: Average pairwise similarities in the two classes of adjectives by embedding

adjective class	BOW			DEPS		
	mean	median	SD	mean	median	SD
SPEED	0.45	0.44	0.143	0.47	0.48	0.108
HUMAN PROP	0.48	0.47	0.072	0.40	0.39	0.052

The standard deviation in the HUMAN PROP class similarities is lower than the standard deviation for the SPEED class. This holds across both embeddings. The differences in variation are significant in both cases (F = .26, p = .04274 and F = .23, p = .02981, respectively). This finding is not in line with hypothesis a), which predicted less variation within the SPEED class. When we turn to the changes in average differences resulting from using the different embeddings, the results are in line with the prediction: On the **deps** embeddings, the SPEED pairs are on average more similar than on the **bow** embeddings (0.47 vs. 0.45). For the HUMAN PROPEN-

Tab. 4: Beta regression model for the cosine similarities between the pairs on both embeddings, with the pairs as random effects and the HUMAN PROPENSITY class and the **bow** vector space as reference levels. (Parametric coefficients with logit link function, adjusted R-squared: 0.749)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.08235	0.12457	-0.661	0.508585
adjective class	-0.11512	0.17661	-0.652	0.514485
vector space	-0.30877	0.08866	-3.483	0.000497
adjective class:vector space	0.38317	0.12594	3.043	0.002346

SITY pairs, the **deps** embeddings show less similarity than the **bow** embeddings (0.40 vs. 0.48).
When modeling the pairwise similarities with a beta regression with the pairs as random effects, a significant interaction of vector space and class emerges, cf. the model in table 4 and the interaction plot in Figure 2.

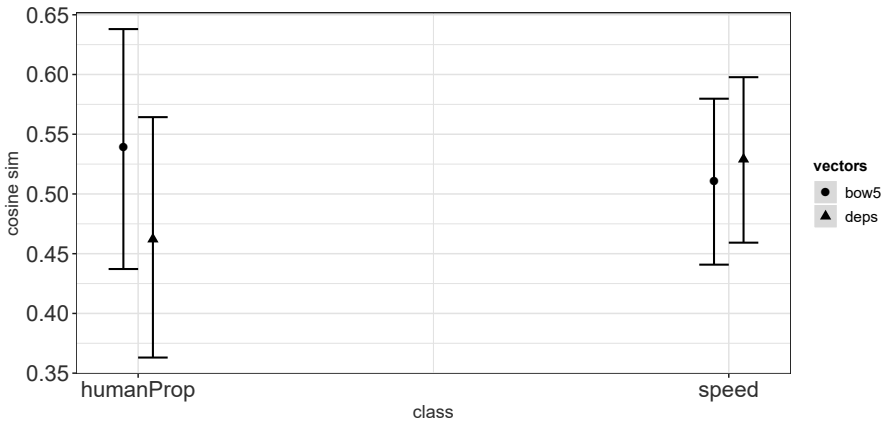

Fig. 2: Interaction of vector space and adjective class when modeling the cosine similarities between base and *-ly* forms

The figure clearly shows the differentiated effect of vector space on the two semantic classes: for the HUMAN PROPENSITY pairs, the **deps** embeddings show significantly less similarity than the **bow** embeddings. In contrast, for the SPEED pairs, the **deps** embeddings show slightly but not significantly higher similarities than the **bow** embeddings.

3.2.2 Hypothesis b): The internal correlations

To what extent are class internal similarities kept intact across forms in the two classes of interest? Table 5 exemplifies the data of interest, using the pair *quick/quickly* and cosine similarities from the **bow** embeddings. The first two columns show the cosine similarities between *quick* and all the other bases from the SPEED pairs. The third and fourth columns shows the cosine similarities between *quickly* and the *-ly* forms of the other speed pairs.

Tab. 5: Cosine similarities of the **bow** embeddings of *quick* and *quickly* and the other bases and *-ly* forms in the SPEED class. Items are ranked by the similarity of *quick* to the other base forms.

***quick* to**	**similarity**	***quickly* to**	**similarity**
slow	0.54	*slowly*	0.69
hasty	0.45	*hastily*	0.41
prompt	0.42	*promptly*	0.62
rapid	0.42	*rapidly*	0.70
sudden	0.39	*suddenly*	0.51
swift	0.39	*swiftly*	0.75
hurried	0.39	*hurriedly*	0.42
brisk	0.38	*briskly*	0.35
rhythmical	0.26	*rhythmically*	0.25
speedy	0.17	*speedily*	0.25

The table is ordered by the cosine similarity between *quick* and the other base forms in the SPEED class, with the second column showing the corresponding similarity values between *quickly* and the *-ly* forms. The measure of interest is the correlation between the two rankings, in this case, Pearson's **r** = 0.726, with p = 0.018. This correlation was also calculated for the same pair on the **deps** embeddings, and the same two calculations were performed for all other pairs from the SPEED class and from the HUMAN PROPENSITY class. Table 6 summarizes the data across all pairs and both embeddings by showing the mean correlations and the standard deviations for both classes on both embeddings.

As the table shows, the mean correlations on the **deps** vectors are across the board smaller. When comparing the two adjective classes, the mean correlations are in both cases higher for the SPEED vectors, with a mean difference of .11 on the **bow** embeddings and of 0.03 on the **deps** embeddings. When modeling the correlations with a beta regression, only the effect of vector spaces as such emerges as significant, cf. table 7.

Tab. 6: Summary of the correlations between the base and *-ly* similarities to the other class members per pair.

	bow		deps	
class	**mean cor**	**sd**	**mean cor**	**sd**
SPEED	0.70	0.16	0.26	0.23
HUMAN PROPENSITY	0.59	0.26	0.23	0.35

Tab. 7: Beta regression model for the class-wise correlations across pairs on both embeddings, with pairs as random effects and the **bow** vector space as reference level. (Coefficients (mean model with logit link), adjusted R-squared: 0.732)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.5517	0.1466	10.586	< 2e-16
vector space	-1.0446	0.1346	-7.761	8.42e-15

In other words, there is no significant difference due to the classes, but the **bow** embeddings show significantly higher average correlations in comparison to the **deps** embeddings.

3.3 Study 1: Discussion

Looking at the two hypotheses, we observe mixed results: Against the expectation of hypothesis a), the HUMAN PROPENSITY class shows less variation in the pairwise similarities on either embedding, although this difference is only significant for the **bow** embeddings. The expectation that on the DEPS embeddings the SPEED pairs should be more and the HUMAN PROPENSITY less similar than on the **bow** embeddings is confirmed: the factors vector space and class significantly interact in predicting the similarities, showing a non-significant higher similarity for SPEED pairs on the **deps** embedding and significantly lower similarities for the HUMAN PROPENSITY pairs on the **deps** embedding.

The class internal correlations considered in hypothesis b) go in the expected directions for both embeddings, that is, on average, the correlation is higher for the pairs from the SPEED class than for the pairs from the HUMAN PROPENSITY class. But class does not emerge as a significant factor when modeling the data. The effect of vector space again emerges as significant here: **bow** embeddings lead to higher correlations across the board. This is unexpected, as I hypothesized that the focus on functional similarity should yield comparatively higher correlations for the SPEED class pairs on the **deps** embeddings.

Trying to understand the many unexpected patterns in the results of Study 1 necessitates a closer look at the data. I will first discuss an interesting pattern in the data underlying the correlations reported, and then explore the behavior of the embeddings further by looking at nearest neighbors across the spectrum of base/*-ly* pairs.

The correlations of interest for hypothesis b) were based on pairwise similarities between each base form and all other base forms within a class, and the corresponding pairwise similarities of the *-ly* forms. Looking not at the correlations but instead zooming in on the average similarities between the base forms and the *-ly* forms reveals a very interesting pattern, cf. table 8.

Tab. 8: Average similarities per form within their classes, on both embeddings.

		average similarities	
class	**vector space**	**base forms**	***-ly* forms**
SPEED	**bow** embeddings	0.29	0.38
HUMAN PROPENSITY	**bow** embeddings	0.41	0.31
SPEED	**deps** embeddings	0.50	0.75
HUMAN PROPENSITY	**deps** embeddings	0.88	0.42

Strikingly, a uniform split by adjective class emerges on both embeddings: For the SPEED class, the average similarities are higher for the *-ly* forms, on the **bow** embedding rising from 0.29 to 0.38, on the **deps** embedding from 0.5 to 0.75. In contrast, for the HUMAN PROPENSITY class, the average similarities are higher for the non-*ly* forms, with 0.41 to 0.31 on the **bow** embeddings, and 0.88 to 0.42 on the **deps** embeddings. One attractive speculation is that the average similarities across the forms are higher whenever the class members are in their semantically natural environment: the event predicates of the SPEED class when functioning adverbially, and the object predicates of the HUMAN PROPENSITY class in all non-*ly* environments. Note, though, that this cannot be decisively concluded from the data. As mentioned in Section 2.1, SPEED adjectives in their attributive usages also prototypically modify nouns that refer to events.

To arrive at a better understanding of the behavior of the embeddings with regard to the target words, and to get an idea of what lies behind the wide variation in the results across the pairs, table 9 shows the nearest neighbors for a representative selection of the items. The table shows the top five words that are most similar in terms of cosine similarities to three pairs from each adjective class. The three pairs are chosen to represent the spectrum of pairwise similarities, using the **bow** embeddings as a reference point. That is, *swift/swiftly* and *eager/eagerly*

show the lowest similarity, *brisk/briskly* and *happy/happily* represent the median, and *rhythmical/rhythmically* and *generous/generously* show the highest similarity. The table does not show very clear patterns. While for *swift* on the **bow** embeddings proper names obscure the picture (*Taylor/Gulliver*), the other sets of nearest neighbors also show considerable variance. For example, the vectors for the top five nearest neighbors for *brisk* in the **bow** embedding seem split into two groups, the two other speed adjectives *fast* and *slow* and *frenetic/leisurely/lively*. The nearest neighbors for *briskly* on the same embedding have an even greater semantic spectrum, including *crazily* and *noisily*. When comparing the SPEED and *human propensity* pairs, the latter seem overall more coherent in that they mostly contain other items from this class. Perhaps this is partially an effect of the small size of the SPEED class. The difference in focus on topical similarity vs. functional similarity across the two vector spaces becomes clearest when looking at the words most similar to the *-ly* forms: the **deps** vectors only pick out other *-ly* forms, whereas the **bow** embeddings include other forms, for example just the base forms. Finally, across the board the top five items are very different both across base and *-ly* forms as well as across embeddings. They also vary in the numbers of antonyms included in the top five. The next study explores whether switching to a different paradigm, the analogy task, brings out the differences between the two classes in a more consistent way.

4 Study 2: An analogy task

In the second study, I approach the difference between the two classes by using an analogy task. The task is first described in Mikolov et al. (2013), who queried the relationship between words by using the question in (13) (the comparison of *aggressive/aggressively* to *rapid/rapidly* is one out of the nine pairs they use to test word-to-word syntactic relationships).

(13) "What is the word that is similar to *aggressive* in the same sense as *rapidly* is similar to *rapid*?"
= Which word d is similar to word c in the same sense as word b is to word a?

The question can be answered via word embeddings in two simple steps:

1. A probe vector is calculated by subtracting the vector for a from the vector for b and adding the vector for c, e.g. $\text{vector}_{\text{rapidly}} - \text{vector}_{\text{rapid}} + \text{vector}_{\text{aggressive}}$
2. Cosine similarity is used to rank all the word vectors in the vector space in terms of their similarity to the probe vector.

Tab. 9: Top five nearest neighbors to six selected target word pairs on either embedding. The first three pairs are from the SPEED class, with low, median, and high pairwise similarity. The last three pairs are from the HUMAN PROPENSITY class, again with low, median, and high pairwise similarity.

	bases		***-ly* forms**	
target pair	**bow**	**deps**	**bow**	**deps**
SPEED class				
swift/swiftly	taylor	tardy	quickly	promptly
	farley	quick	promptly	forcefully
	rutter	spry	slowly	amicably
	cooke	hick	rapidly	gracefully
	gulliver	brisk	gradually	expeditiously
brisk/briskly	frenetic	languid	zipping	noisily
	leisurely	hectic	crazily	fluidly
	lively	desultory	noisily	uneventfully
	fast	adroit	gently	painlessly
	slow	frenetic	leisurely	erratically
rhythmical/ rhythmically	declamatory	polyrhythmic	melodically	sonically
	melismatic	dance-like	harmonically	melodically
	syncopated	gestural	syncopated	harmonically
	contrapuntal	melismatic	rhythmical	cognitively
	polyrhythmic	bell-like	polyrhythmic	aerodynamically
HUMAN PROPENSITY class				
eager/eagerly	anxious	anxious	anxiously	earnestly
	willing	willing	enthusiastically	fervently
	reluctant	hesitant	warmly	anxiously
	wanting	loath	joyfully	strenuously
	unwilling	reluctant	gratefully	vociferously
happy/happily	glad	ecstatic	gladly	unhappily
	pleased	glad	unhappily	silently
	happier	anxious	joyfully	discreetly
	thankful	thankful	cheerfully	peacefully
	thrilled	hesitant	quietly	quietly
generous/ generously	gracious	gracious	generous	ably
	generously	amiable	graciously	enthusiastically
	frugal	judicious	liberally	sympathetically
	compassionate	agreeable	handsomely	handsomely
	handsome	courteous	gifts	cordially

3. The word vector that is most similar to the probe vector, that is, the vector that is on the first rank, is the answer to the question in (13).

Recall the example of vector addition using the toy vectors in Section 2.2.1: the procedure here is not more complicated, involving only subtraction and addition of the corresponding vector components.

Thinking about the implementation of the analogy task in terms of the role of the *-ly*, the first step itself consists essentially of the isolation of the meaning contribution of *-ly*: subtracting the vector of *rapid* from the vector of *rapidly* yields a vector that stands for the contribution of *-ly*. Adding the vector of a different adjectival base, here *aggressive*, yields the probe vector. As my interest here is not in testing a specific pair, but to explore the behavior of the contribution of *-ly* averaged across the two sets of vectors drawn from the SPEED and HUMAN PROPENSITY classes, I will explore the effect of this approach at class-level. That is, I will first isolate the contribution of *-ly* in each of the pairs of interest here, and then use the resulting vector presentations to compare the meaning contribution of *-ly* averaged over each of the two classes as well as over both classes, that is, all pairs.

Recall the finding presented in the discussion of Study 1 that on both embeddings the *-ly* forms of the SPEED pairs are more similar to each other then the corresponding base forms while the opposite holds for the HUMAN PROPENSITY pairs. Based on that finding, I expect that a single meaning contribution of *-ly* can only be successful in the analogy task for the SPEED pairs. A second expectation, deriving from the tables of the top five nearest neighbors considered in the discussion of Study 1, is that this will work better on the **deps** embeddings, which consistently had *-ly* forms as nearest neighbors of the six *-ly* forms of the target pairs. In addition, the average vectors themselves allow again to assess the stability of contrast aspect from hypothesis a) in Study 1.

4.1 Study 2: Materials and techniques

Study 2 uses the same set of pairs and the same set of word embeddings as Study 1. The probes for the analogy task were calculated as follows (always separately for both the **bow** and the **deps** embeddings):

1. To isolate the contribution of *-ly* across a pair, the vector for the base form was subtracted from the vector for the *-ly* form. This yields 22 *-ly* vectors, or, in the terminology of Bonami & Paperno (2018), 22 vector offsets, one for each pair. These 22 vectors were used to calculate two average vectors. One average vector was calculated using the 11 vectors from the SPEED class, the second av-

erage vector was calculated using the 11 vectors from the HUMAN PROPENSITY class.

2. From the 22 individual vectors, two average *-ly* vectors were calculated for every base, using a leave-one-out approach. One 'all-*ly*' vector, encompassing the individual *-ly* vectors from both classes, and one class-specific vector, depending on the class of the base: either 'SPEED-ly' or 'HUMAN PROPENSITY *-ly*':
   a) all-*ly*: a vector for class-unspecific *-ly*, created by averaging across all individual *-ly* vectors except the *-ly* vector from the target pair itself. I.e., for *quick/quickly*, the all-*ly* vector is calculated by averaging over the 10 other SPEED class *-ly* vectors and all the HUMAN PROPENSITY *-ly* vectors.
   b) SPEED-*ly*: a class specific vector calculated by averaging across the 10 other *-ly* vectors from the SPEED class.
   c) HUMAN PROPENSITY-*ly*: a class specific vector calculated by averaging across the 10 other *-ly* vectors from the HUMAN PROPENSITY class.
3. For every base, the two average *-ly* vectors resulting from the previous step were used to calculate two probes: the general probe using the all-*ly* vector, and the class-specific *-ly* probe using the class-specific *-ly* vectors. This was done by simply adding the corresponding *-ly* vector to the vector of the base form.

The resulting vectors are explored in three different ways:

1. by comparing the variation of the individual *-ly* vectors,
2. by comparing the ranks in the analogy task, and
3. by comparing the similarities in the analogy task.

The first is the most straightforward assessment: the variance of the similarities between the individual vector offsets, and the average vector offset within a class is compared. This method is adapted from Bonami & Paperno (2018), who use it to investigate the different relations in their triplets of one pivot form and one derivationally and one inflectionally related other form. They hypothesize that when comparing inflectional and derivational relations this way, the vector offsets from the inflectional relations should show less dispersion around the average vector offset, and thus less variance, in line with the idea that inflection is connected to stable contrasts. Derivational relations, in comparison, should show more dispersion and hence more variance. Here, of course, there is no pivot, and the forms to be compared are the members of the two different semantic classes. As described above, the individual vector offsets are the 22 vectors arrived at by subtracting the vector of the base of a pair from the *-ly* form of that pair. The average vector offset for a class is the average of each 11 *-ly* vectors of a class which are simply added together and then divided by 11.

The second and third are two aspects of the same task: when using the leave-one-out-vectors to construct analogy probes, I want to know how well these probes work in identifying the expected target. This is assessed in two different ways: first, by looking at the rank of the target among the nearest neighbors of the probe, and second, by looking at the cosine similarity of the target to the probe. For both measures, the probes using the all-*ly* vectors are compared to the class-specific *-ly* vectors. I am using both ranks and absolute similarities because these two aspects are in principle independent of each other and therefore both of interest. The comparison of the ranks is the standard way to assess the success of the analogy task: only if the target is the nearest neighbor of the probe is the analogy task correctly answered. Whether this success comes about with a relative high or relative low cosine similarity between the probe and the target depends on the overall distribution of items in the vector space. The absolute similarities are therefore independently of interest.

As before, for all three comparisons both vector spaces are explored. In line with the hypotheses explored in Study 1 and the remarks at the end of the previous section, there should be a more marked contrast between the two classes on the **deps** embeddings. The SPEED class should show less variation for its *-ly* vectors and better performance on the analogy task in comparison to the results from the **bow** embeddings.

4.2 Study 2: Results

4.2.1 Variance in the vector offsets

Recall that the variance in the vector offsets is simply the variance in the cosine similarities between the 11 individual *-ly* vectors per class (SPEED or HUMAN PROPENSITY) and the average *-ly* vector per the respective class. For the **bow** vectors, the variance in the cosine similarities is bigger for the SPEED class (0.018) than for the HUMAN PROPENSITY class (0.004). This difference is significant ($F = 4.9641$, $p = 0.01838$). For the **deps** vectors, the SPEED class shows a variance of 0.0030, compared to a variance of 0.0025 in the HUMAN PROPENSITY class. This very small difference is not significant (F 1.1908, $p = 0.7878$). This finding is not in line with expectations, as for both embeddings less variation, that is, less dispersion, was expected for the members of the SPEED class. Note also that across the two vector spaces the results are unexpected: only on the **bow** embeddings is there a significant difference in the variances (in an unexpected directions), whereas the **deps** embeddings show no difference.

4.2.2 Ranks in the analogy task

When the target *-ly* form is the nearest neighbor of the probe, the task is considered a success. For example, after calculating a probe for *swiftly* by adding an average *-ly* vector to the base *swift*, the task is successful if the vector for *swiftly* is placed on the first rank in the set of all vectors in the vector space ranked by cosine similarity to the probe. Mikolov et al. (2013) implicitly acknowledge that this task is hard, because they discard the words used to construct the probes when looking for the closest nearest neighbor. That is, for them, if the base form *swift* is on the first rank and the target *swiftly* on the second rank, it would still count as a success. Here, I will count as successes only true first ranks, but I will report on the top five ranks. This allows to assess how close the targets were to success, that is, to the first rank, even if the first rank itself is missed.

To conveniently assess the performance of the probes across the different classes, average vectors, and vector spaces, the results are always presented by using the same types of graphs. The ranks are shown on the y-axis, with the first rank on top, going down to the fifth rank and finally collecting all targets outside the top five range in the category 'out'. The *-ly* targets are represented on the x-axis, first the targets from the SPEED class, second the targets from the HUMAN PROPENSITY class. Both are ordered alphabetically, and the dots showing their place in the ranking are color-coded for the two classes: blue for the SPEED class and red for the HUMAN PROPENSITY class.

Figure 3 shows the ranking of the *-ly* targets when using the **bow** probes and the *ly*-vector generalized across all 22 pairs except the respective probe pair.

The task is not successful for any of the pairs, but the overall performance for the SPEED probes is better. For the SPEED class, the majority of the targets, six, are on the second rank, the vector for *suddenly* is on the third rank, and the probe for *slowly* on the 5th rank. The vectors for *promptly*, *speedily*, and *swiftly* are outside of the top five range. In contrast, only two of the HUMAN PROPENSITY targets are on the second rank, one is occupying the third rank and another the fourth rank. Seven out of eleven are outside the top five. In all cases, that is, for both the SPEED as well as the HUMAN PROPENSITY targets, it is always the base form that occupies the first rank. In other words, the contribution of the all-*ly* vector does not manage to move the probe vector far enough from the vector of the adjectival base, indicating that the contributions of *-ly* across all 21 other items are too different from each other and the target *-ly* to yield a specific enough average vector.

When instead using the probes calculated using the class-wise averaged *-ly* vectors, the rankings clearly improve for the SPEED class, but also for the HUMAN PROPENSITY class, cf figure 4.

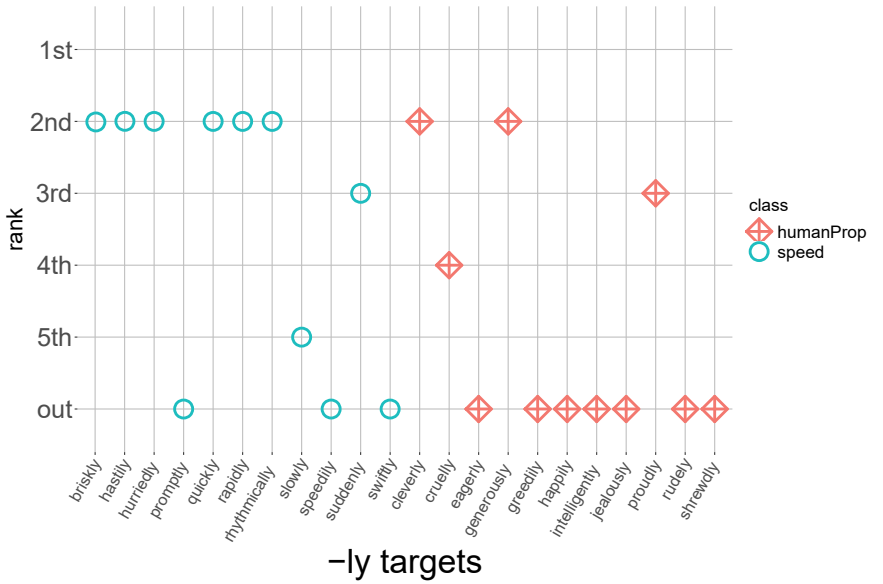

Fig. 3: Rankings of the *-ly* targets in terms of their cosine similarity to the probes. Using the **bow** embeddings, with the *-ly* vector averaged across all other pairs using the leave-one-out technique.

For the SPEED class, all three vectors that were outside the top five using the all-*ly* probes are now within the top five, with the vector for *promptly* in second place, the vector for *swiftly* in third place and the *speedily* vector in fifth place. Only the *briskly* vector falls in ranks, from second to fourth. For the HUMAN PROPENSITY class, three are in second place, with the *shrewdly* vector previously outside the top-five. The *intelligently* vector also moves from outside the top five to rank three. At the same time, the vector for *cruelly*, on rank four with the all-*ly* probe, is now also outside of the top five, so that the number of items in this class outside of the top five ranks only decreases by one. Just as before, when looking at the items ranked first, these are always the base forms.

Figure 5 shows the result when averaging *-ly* across both classes using the **deps** embeddings. The result is clearly worse than the results obtained with the **bow** vectors. Within the **deps** embeddings, the results for the SPEED class are better, but for both classes, the majority of targets, seven and nine, respectively, are ranked outside the top five. For the SPEED class, there are two second ranks, and two fifth ranks. For the HUMAN PROPENSITY class, the *rudely* vector is ranked second, and the *cruelly* vector is ranked fourth.

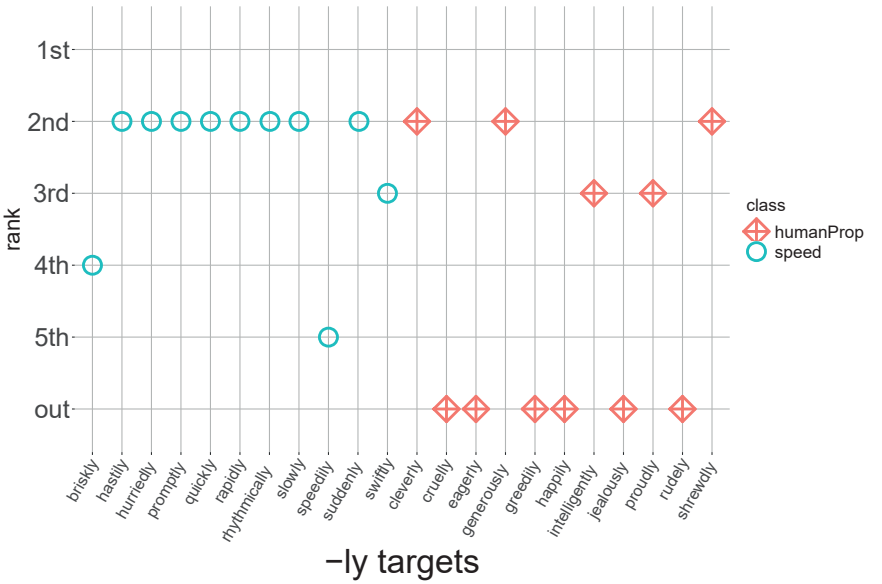

Fig. 4: Rankings of the -*ly* targets in terms of their cosine similarity to the probes. Using the **bow** embeddings, with the -*ly* vector averaged by class.

That these probes behave notably different from the **bow** probes is already apparent in that the vectors for *speedily* and *rudely*, two of the overall three vectors ranked second, are ranked outside of the top-five on the **bow** embeddings. Another notable contrast when looking in more detail at the nearest neighbors of the probes emerges when considering the words whose vectors are on rank one in this task, cf. the second column in table 10. For both classes, only six first-ranked items each correspond to the base forms, and five first-ranked items each are -*ly* forms, although not the target forms. Of these, some forms in the SPEED class seem notably off, e.g. *laboriously* as closest to the *quickly*-probe.

Figure 6 shows the results for the class-specific -*ly* vectors on the **deps** embeddings. This constitutes a notable improvement, again with an advantage for the SPEED pairs. The vector for *briskly* is successfully predicted, and the vectors for *speedily* and *swiftly* are ranked second, the one for *rapidly* fourth and the vector for *hastily* fifth. For the HUMAN PROPENSITY class, the improvements are limited to the vectors already ranked within the top five. The *rudely* vector is also a success, as expected placed on the first rank. The *cruelly* vector is now on the second rank, but all other vectors remain outside of the top five. Intriguingly, the two successful items performed relatively badly on the class-specific **bow** embeddings, with the *briskly*-probe ranked fourth and the *rudely*-probe outside of the top five.

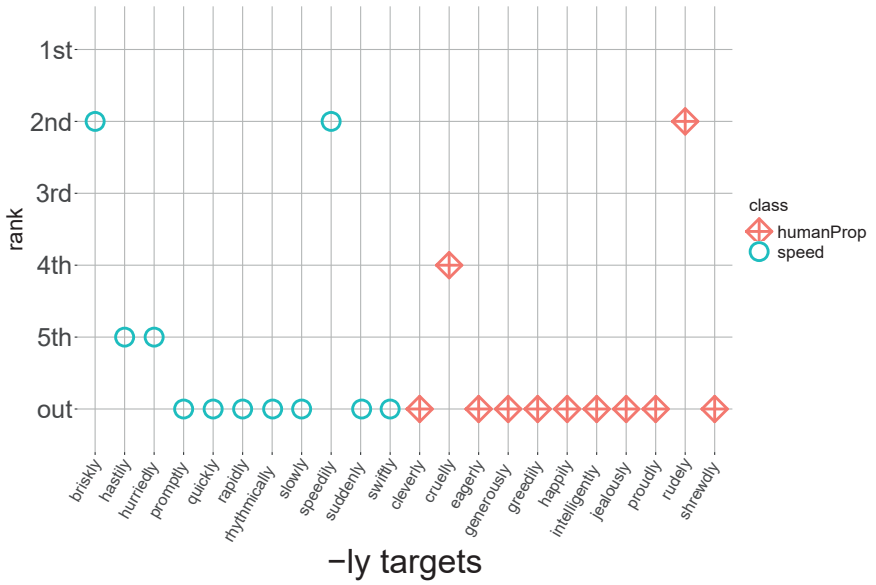

Fig. 5: Rankings of the *-ly* targets in terms of their cosine similarity to the probes using the **deps** embeddings. *-ly* vector averaged across all pairs using the leave-one-out technique.

Just as for the probes using the all-*ly* vector, the actual items on the first rank in this task are not just the corresponding base forms, in contrast to the results from the **bow** embeddings. All first ranked items for the class-specific *-ly* probes are shown in the third column in table 10. For the SPEED class, besides the single success in the form *briskly*, there are the same six base forms that already showed up for the all-*ly* probes. The remaining four first ranked items are all *-ly* forms, three times *swiftly*, and again *self-consciously* for *rhythmical*. While these are not the expected words, they are clearly closer in meaning than the first ranked items for the all-*ly* probes as *swiftly* is itself part of the SPEED class. For the HUMAN PROPENSITY class, all first ranked items are now *-ly* forms, mostly coming from the HUMAN PROPENSITY class.

4.2.3 The similarities in the analogy task

Table 11 shows the average cosine similarities between probes and targets for both types of *-ly* vectors, all-*ly* and class-specific *-ly*, on both embeddings.

For the **bow** embeddings, the similarity between probe and target for the items in the SPEED class is higher when using the class-specific *-ly* vectors than

Tab. 10: Analogy task: most similar words to the probe for the generalized *-ly* vectors and the separated *-ly* vectors under the **deps** embedding, using the leave-one-out approach

target relation	depsAll	depsSep
brisk-briskly	erratically	briskly
hasty-hastily	rudely	swiftly
hurried-hurriedly	hurried	hurried
prompt-promptly	prompt	prompt
quick-quickly	laboriously	swiftly
rapid-rapidly	rapid	rapid
rhythmical-rhythmically	self-consciously	self-consciously
slow-slowly	slow	slow
speedy-speedily	speedy	speedy
sudden-suddenly	irreversibly	swiftly
swift-swiftly	swift	swift
clever-cleverly	tactfully	concretely
cruel-cruelly	cruel	jealously
eager-eagerly	eager	royally
generous-generously	cordially	cordially
greedy-greedily	tactfully	tactfully
happy-happily	happy	cheerfully
intelligent-intelligently	maturely	flexibly
jealous-jealously	jealous	impulsively
proud-proudly	proud	graciously
rude-rudely	rude	rudely
shrewd-shrewdly	ably	ably

Tab. 11: Similarities in the analogy task: descriptive overview

		bow		deps	
class	lyVector	sim	sd	sim	sd
speed	all-*ly*	0.562	0.0868	0.744	0.0701
speed	speed-*ly*	0.590	0.0726	0.742	0.0577
human propensity	all-*ly*	0.561	0.0749	0.735	0.0459
human propensity	human propensity *-ly*	0.559	0.0776	0.768	0.0392

when using the all-*ly* vectors. In contrast, for the items from the HUMAN PROPENSITY class, the all-*ly* result in very slightly higher similarities between probes and targets in comparison to the class-specific *-ly* vectors. For the **deps** embeddings, it is exactly the other way around: For the SPEED class, the all-*ly* vectors produces slightly higher similarities than the class-specific *-ly* vectors. For the HUMAN PROPENSITY class, the class-specific vectors produce higher similarities between

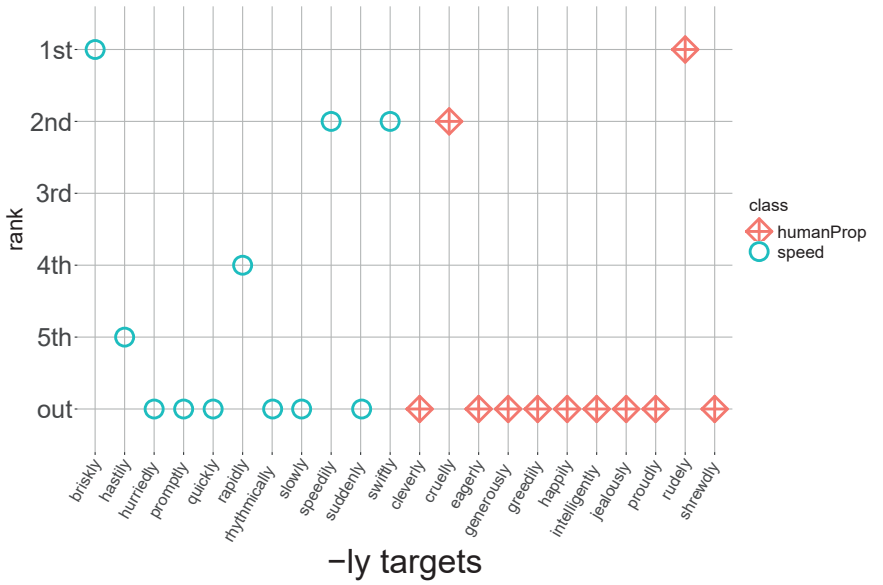

Fig. 6: Rankings of the *-ly* targets in terms of their cosine similarity to the probes using the **deps** embeddings. *-ly* vector averaged by class using the leave-one-out method.

probes and targets. Across the board, the similarities on the **deps** embeddings are notably higher than the similarities on the **bow** embeddings.

When modeling the similarities for each vector space with beta regression models, we see for both embeddings significant interactions of semantic classes and *-ly* vectors, cf. the models in tables 12 and 13 and the interaction plots in figure 7.

Tab. 12: Beta regression model for the cosine similarities probe-target and the bow embeddings, with pairs as random effects, and the HUMAN PROPENSITY class and the all-*ly* vectors as reference levels. (Parametric coefficients with logit link, R-squared adjusted: 0.732)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.251095	0.097536	2.574	0.0100
adjective class	0.001197	0.137940	0.009	0.9931
-ly vector	-0.008422	0.038421	-0.219	0.8265
adjective class:*-ly* vector	0.128437	0.054484	2.357	0.0184

Tab. 13: Beta regression model for the cosine similarities probe-target and the **deps** embeddings, with pairs as random effects and the HUMAN PROPENSITY class and the all-*ly* vectors as reference levels. (Parametric coefficients with logit link, R-squared adjusted: 0.902)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.02944	0.08896	11.572	< 2e-16
adjective class	0.06169	0.12592	0.490	0.624225
-ly vector	0.17790	0.03872	4.594	4.34e-06
adjective class:*-ly* vector	-0.18993	0.05455	-3.482	0.000498

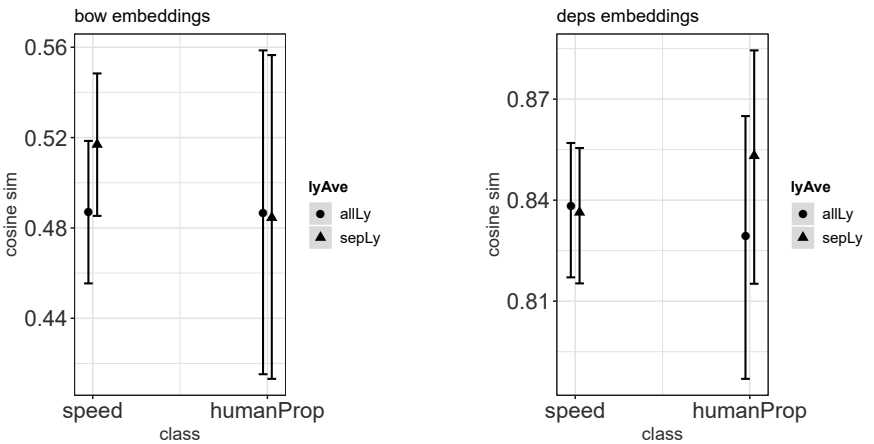

Fig. 7: Interaction plots for average vectors and adjective class on the **bow** embeddings (left hand side) and on the **deps** embeddings (right hand side).

The two interaction plots clearly show the difference between the results in the two vector spaces, with the prediction of the two average *-ly* vectors for each class almost completely overlapping for the HUMAN PROPENSITY class on the **bow** embeddings but for the SPEED class on the DEPS embeddings. Also of interest is the clear difference in the confidence bands between the two semantic classes on both embeddings: the confidence bands for the SPEED class are consistently much smaller than those for the HUMAN PROPENSITY class.

4.3 Study 2: Discussion

Study 2 brought differentiated results that are again only partially in line with expectations. Against expectations, the offset vectors using the **bow5** embeddings

showed less variation for the HUMAN PROPENSITY class, with no difference on the **deps** embeddings.

For the rankings in the analogy task, the **bow** embeddings turned out to be more successful than the **deps** embeddings: While the task was not successful in its final step with the highest ranked targets only on the second ranks, the results show that the embeddings are able to pick up a difference at the class level, with the difference in the expected direction: probes from the SPEED class performed better overall, and there was more improvement in the rankings for this class in comparison to the HUMAN PROPENSITY class when switching to class-specific *-ly* vectors. The pairs in the SPEED class seem to have more in common so that a vector averaged over these forms is more meaningful than using the same method for the HUMAN PROPENSITY class. For the **deps** vectors, the results were comparatively weak, although SPEED targets were overall ranked higher and again improved more on the class-specific vectors.

When looking at the absolute cosine similarities for the targets on the **bow** embedding, the **bow** embeddings behave mostly as expected: only for the SPEED class is there a clear difference when comparing the results using the all-*ly* vectors to the results from using the class-specific *-ly* vector. As the model shows, the interaction between class and type of vector is significant. The clear difference in the confidence bands between the semantic classes in the interaction plot also shows very clearly that the model is more confident when it comes to predicting the behavior of the items from the SPEED class. Comparing the absolute similarities with the rankings, it is notable that when using the all-*ly* vectors there is no significant difference between the two semantic classes in terms of the absolute similarities of the targets to the probes, while we saw that the targets from the SPEED class are ranked higher than those from the HUMAN PROPENSITY class. This underscores the usefulness of employing both rankings and absolute values together, suggesting that the items from the two semantic classes clearly occupy very differently structured areas in the semantic space.

The pattern for the absolute cosine similarities on the **deps** embeddings is exactly opposite, and therefore clearly against expectations. The only commonality with the model for the **bow** results are the tighter confidence bands for the items from the SPEED class. It is not clear to me how meaningful these results are, especially since this set-up on the whole did not perform well as far as the rankings are concerned. At the same time the rankings still showed an advantage of the SPEED class items when using the class-specific *-ly* probes, and a better performance of the SPEED class items in general. This does not line up with the absolute similarities. Overall, as we saw from investigating the first ranked items for this class, the **deps** embeddings show more unexpected, and at the moment unexplainable, behavior.

5 Study 3: Analogy and *-ly* aware embeddings

The results so far have been mixed; the only result that went mostly as expected were the ranking and cosine similarities for the **bow** embeddings in the analogy task. With one intriguing twist: in all cases, the first rank in the analogy task on the **bow** embeddings are occupied by the base forms. Intuitively, this is highly unexpected: the target is very obviously a *-ly* form, so non-*ly* forms should not be selected.

The embeddings used so far don't represent information concerning the forms of the words themselves. What happens when we switch do embeddings that do exactly this?

Mikolov et al. (2017) implement one way of including word internal structure in embeddings, in their terminology subword information. The words are broken down into character-ngrams (with characters simply the orthographic characters), which are in turn represented as vectors, the sum of which is added to the standard word vector. Study 3 employs these enriched embeddings in the analogy task.

Note that, conceptually, including word internal structure in the embeddings is a clear step away from the original distributional slogan "You shall know a word by the company it keeps!" from Firth (1957, 11). Crucially, when using the embeddings as stand-in for meanings, it lets form aspects of the word contribute to its meaning. I will come back to this issue in the discussion.

5.1 Study 3: Material and techniques

Study 3 uses the exact same materials and techniques for the analogy task as Study 2, with the only difference being the embeddings used. These are now the pretrained fasttext (= **fast**) embeddings including subword information, wiki-news-300d-1M-subword.vec.zip, from https://fasttext.cc/docs/en/english-vectors.html. These vectors have been trained on a corpus of 16 billion tokens, using Wikipedia 2017, the UMBC webbase corpus and the statmt.org news dataset. For details of the training, cf. Mikolov et al. (2017). As the vector space contains considerably more embeddings (1 million vs. 183,870 (**bow**) and 174,015 (**deps**), only the embeddings for the most frequent 175,000 words were used to search for the nearest neighbors in order to keep the results comparable.

5.2 Study 3: Results

5.2.1 Rankings in the analogy task

When using the general *-ly* probes, two of the SPEED targets and three of the HUMAN PROPENSITY targets are correctly identified. Seven of both the SPEED targets and the HUMAN PROPENSITY targets are ranked second. Finally, two SPEED targets and one HUMAN PROPENSITY target are ranked third. All non-target first ranks are occupied by the respective non-*ly* forms.

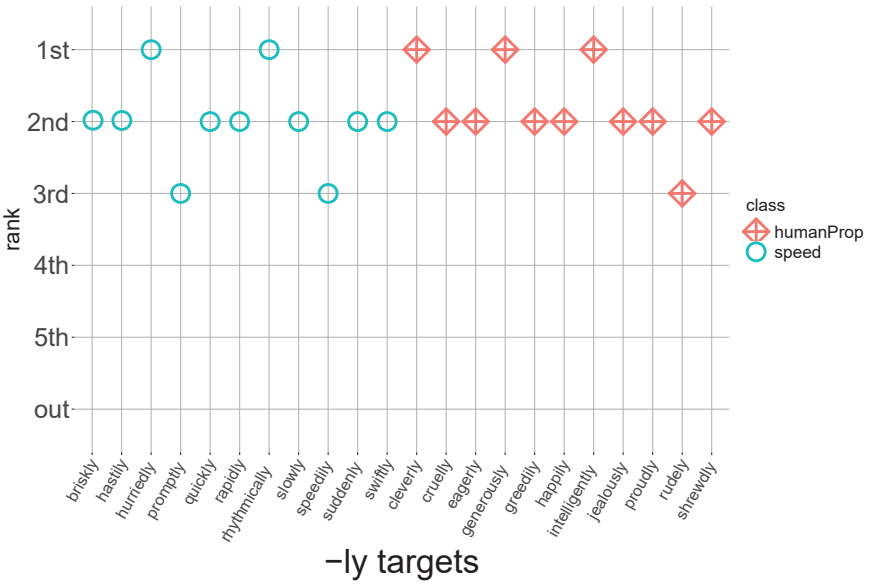

Fig. 8: Rankings of the *-ly* targets in terms of their cosine similarity to the probes using the **fast** embeddings, with the *-ly* leave-one-out vector averaged across both classes.

Switching to the class-specific *-ly*-vectors results in a slight improvement for the SPEED vectors, where the target vector for *hastily* moves from second to first rank. The ranks for the HUMAN PROPENSITY targets remain unchanged.

Just as for the all-*ly* probes, the other first ranks are occupied by the respective base forms.

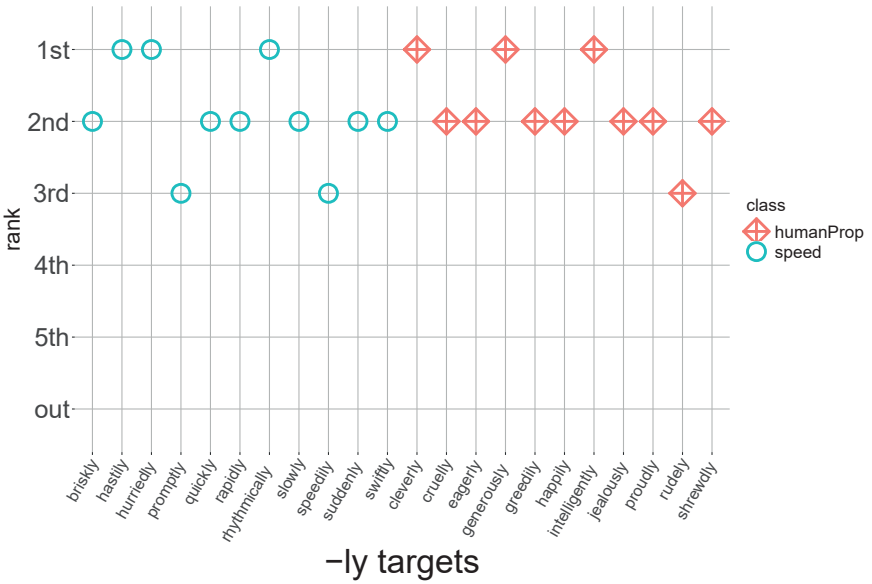

Fig. 9: Rankings of the -*ly* targets in terms of their cosine similarity to the probes using the **fast** embeddings, with the -*ly* leave-one-out vector averaged by class.

5.2.2 Similarities in the analogy task

The average similarities of the targets to the probes are shown in table 14.

Tab. 14: Average cosine similarities between probe and target vectors on the **fast** embeddings.

class	-*ly* vector	cosine similarity	standard deviation
speed	all-*ly*	0.755	0.0709
speed	speed--*ly*	0.750	0.0859
human propensity	all-*ly*	0.786	0.0484
human propensity	human propensity-*ly*	0.800	0.0445

For the SPEED class, the cosine similarities between targets and probes are minimally more similar when using the all-*ly* vector than when using the class-specific speed -*ly* vector. In contrast, the opposite effect, slightly larger this time, can be observed for the HUMAN PROPENSITY vectors. Modeling the similarities with beta regression yields no significant effects neither for the interaction between class and *ly*-vectors nor for adjective class or average vector used individually.

5.3 Study 3: Discussion

The aim of Study 3 was to explore whether including subword information might result in more targets being successfully identified in the analogy task. This was indeed the case, and the vector space gave overall the best results across all pairs and for both classes. However, using this vector space also eroded all differences observed so far for the two different classes and for the effect of a switch of all-*ly* vectors to class specific vectors: neither made a significant difference.

This leveling of all differences brings us back to the nature of these embeddings. As mentioned above, the very fact that the **fast** embeddings use subword information makes them conceptually very different from the classic idea behind distributional semantics. That *swift* and *swiftly* share patterns in their orthographic form is not a feature of their distribution, but simply a feature of their form. And this form feature is not linked to any differences in lexical semantics of the adjectival base forms, therefore it is not surprising that encoding subword information in the embeddings weakens any effect of semantic class membership encoded in the embeddings via the distribution of the words. In effect, using the base/-*ly* pairs with the **fast** embeddings, the analogy task is no longer merely about semantic or syntactic similarity, but about these similarities in conjunction with form overlap.

6 General discussion

The three studies presented here explored the hypothesis that the relations between base and -*ly* forms for SPEED adjectives behave like instances of inflection whereas the relations between bases and -*ly* forms of HUMAN PROPENSITY adjectives behave like instances of derivation in terms of their distributional semantics. Two specific aspects were investigated: a) is there evidence for SPEED adjectives to show more stable contrasts than HUMAN PROPENSITY adjectives and b) is there evidence for a vector space more sensitive to functional similarity to show clearer differences for these two classes than a vector space focusing on topical similarities.

The results from Study 1 and Study 2 only yield partial evidence for this. In particular, assessing the difference between the two classes by looking at variance, either in the similarities across the pairs in Study 1, or in terms of the dispersion of the offset vectors in Study 2, has shown that the HUMAN PROPENSITY class shows less variance (except for the offset-dispersion on the **deps** embeddings, which showed no difference). Support for the hypothesis that SPEED pairs

behave more like inflection came via the second aspect of comparison in Study 1: the HUMAN PROPENSITY pairs are less similar and SPEED pairs more similar when comparing the **deps** embeddings to the **bow** embeddings. Not significant, but in the expected direction was the final finding from Study 1: class internal correlations are higher for the SPEED class. Another chance finding from Study 1 was the overall higher mean similarity of *-ly* forms from the SPEED class to each other in comparison to the base forms, with the opposite pattern obtaining for the HUMAN PROPENSITY class: this shows that the difference in semantic class has reflexes in the distribution, although these particular reflexes cannot straightforwardly be related to the general issue of inflection versus derivation.

The analogy task in its original form, that is, focusing on the ranks of the targets within the nearest neighbors of the probes, shows the clearest pattern in line with the original hypothesis: switching from the non-specific all-*ly* vectors to the class specific vectors comes with a very clear improvement for the targets from the SPEED class, notably on the **bow** embeddings. For the **bow** embeddings, this finding goes together with the effect of the switch for the absolute similarities of the targets, while yielding opposite results for the **deps** embeddings. The latter result, however, is perhaps not so relevant given that this vector space performed so poorly on this task overall.

Study 3 explored whether the failure of the best performing embeddings in the analogy task in study 2 to produce any successes could be overcome by using the **fast** embeddings, which also encode subword information. This was expected to help answer the analogy task successfully, as except for the *-ly* ending itself and some minor adjustments (hasty/hastily etc.) the words of a pair completely overlap orthographically. This expectation turned out to be correct. This vector space overall performed best on the analogy task, but the comparison of the results from the all-*ly* vectors to those of the class-specific vectors showed virtually no effect. In a way, this result links back very nicely to the starting point of this paper: *-ly* forms are interesting, because the theoretical literature cannot decide whether they should be treated as inflectional or derivational. The clear semantic differences between the members of pairs of the SPEED class on the one hand and the HUMAN PROPENSITY class on the other hand outlined in section 2.1 are not immediately obvious and require close semantic analysis. They are obscured by the pairs sharing the same affix to link their forms, and the fact that they both occur in the prototypical adjectival and adverbial patterns.

These semantic differences are at least partially accessible via the classic distributional paradigm, that is, when using the distribution of a word to characterize its meaning. They are again hidden when a distributional approach adds form-based word-internal information, like the subword information in the **fast** embeddings used here.

Were the number of items to small to show a quantitative difference? It is clear that purely distributional approaches often use far higher numbers of items. For example, the work by Bonami & Paperno (2018) on French is based on 100 triples each containing the pivot form and a inflectionally and a derivationally related form, whereas I just compared 22 pairs. I believe that the very fact that clear distributional reflexes were found shows that the set of items was not too small. At the same time, widening the approach to encompass larger numbers is certainly desirable, but was bound here by the few overall numbers of pairs from the SPEED class. But 11 pairs is already a considerable improvement when compared to typical theoretical works in semantics, where the focus is usually on at most a handful of words (compare many of the works cited in Section 2.1). A further desirable extension of the methods used in this paper suggests itself when looking at Schäfer (2020). There, I only looked at four pairs, *quick/quickly*, *slow/slowly*, *wise/wisely* and *lucky/luckily*, but I compared their vectors across six distinct usages, three using the base form, and three using the *-ly* form. This approach allows to tease apart usage details necessarily glossed over in the current study and might lead to more consistent results across tasks.

A final issue concerns the internal consistency of the two classes compared here: just as for the number of pairs considered overall, the very fact that clear differences could be found without taking further steps shows that the classes were consistent to a sufficient degree. But there are two steps that should be taken in future research to put these results on a more solid footing. On the one hand, this investigation would ideally be complemented by a purely quantitative approach that explores whether the distribution of pairs by themselves would also establish these same classes, or whether perhaps other classifications might capture the internal semantic structure of the system of English adjectives more adequately. On the other hand, it would also be useful to validate this paper's methodology on datasets that show similar characteristics. One such dataset of interest are the *-ing* nominalizations discussed in this volume by Lieber (2023), where she finds a clear difference of the effect of *-ing* depending on whether the base verbs are durative or not. If the approach presented here is on the right track, this difference should also show up in a distributional analysis.

7 Conclusion

In this paper, I argued that a closer look at the semantics involved across base/*-ly* pairs such as *quick/quickly* on the one hand and *clever/cleverly* on the other hand suggests that only the SPEED pairs are instances of inflection whereas the HU-

MAN PROPENSITY pairs are instances of derivation. The overall mixed results of the three distributional semantics studies show that there clearly are differences in the semantic contribution of *-ly* across the two semantic classes. But these differences are more complex than hypothesized. This suggests that a binary distinction between inflectionally and derivationally related forms does not do justice to the data, and gives yet more evidence why the relation encoded across base/*-ly* pairs remains a challenge. This challenge is made even more difficult by the finding that embeddings including word-internal information perform best on the analogy task but are insensitive to any semantic distinction between the SPEED and HUMAN PROPENSITY class. Given that these two specific adjective classes were selected because they show the clearest and most systematic contrasts according to theoretic accounts, these results also show that there is still a long way to go when it comes to understanding the relationship between lexical classes established in theoretical works and possible reflections of these classes in distributional semantics.

Acknowledgements

Many thanks to Sven Kotowski and two anonymous reviewers for helpful comments on earlier versions of the paper and to Dominic Schmitz for a similarly helpful beta regression starter file in R. Work on this paper was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 833 – Project ID 75650358.

Bibliography

Agirre, Enekoand, Enrique Alfonseca, Keith B. Hall, Jana Kravalova, Marius Pasca & Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 19–27.

Baroni, Marco, Georgiana Dinu & Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference* 1. 238–247.

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

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*, 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.

Bücking, Sebastian & Claudia Maienborn. 2019. Coercion by modification. The adaptive capacities of event-sensitive adnominal modifiers. *Semantics and Pragmatics* 12(9). 1–39. 10.3765/sp.12.9.

Davies, Mark. 2008–. The corpus of contemporary American English: 450 million words, 1990–present. Available online at http://corpus.byu.edu/coca/.

Dixon, Robert M. W. 1982. *'where have all the adjectives gone?' and other essays in semantics and syntax*. Berlin: De Gruyter.

Dumais, Susan T. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology* 38(1). 188–230. 10.1002/aris.1440380105. http://dx.doi.org/10.1002/aris.1440380105.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman & Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst.* 20(1). 116–131. 10.1145/503104.503110. https://doi.org/10.1145/503104.503110.

Firth, John R. 1957. *Papers in linguistics: 1934 –1951*. Oxford: Oxford University Press.

Giegerich, Heinz J. 2012. The morphology of *-ly* and the categorial status of 'adverbs' in English. *English Language and Linguistics* 16(3). 341–359. 10.1017/S1360674312000147.

Günther, Fritz & Marco Marelli. 2021. Caoss and transcendence: Modeling role-dependent constituent meanings in compounds. *Morphology* 1–24. 10.1007/s11525-021-09386-6. https://doi.org/10.1007/s11525-021-09386-6.

Günther, Fritz, Marco Marelli & Jens Bölte. 2020. Semantic transparency effects in German compounds: A large dataset and multiple-task investigation. *Behavior Research Methods* 52(3). 1208–1224. 10.3758/s13428-019-01311-4.

Koev, Todor. 2017. Adverbs of change, aspect, and underspecification. In Dan Burgdorf, Jacob Collard, Sireemas Maspong & Brynhildur Stefánsdóttir (eds.), *Proceedings of the 27th semantics and linguistic theory conference*, 22–42. Washington, DC: Linguistic Society of America. Https://doi.org/10.3765/salt.v27i0.4123.

Kotowski, Sven & Martin Schäfer. 2023. Quantifying semantic relatedness across base verbs and derivatives. english *out*-prefixation. In Sven Kotowski & Ingo Plag (eds.), *The semantics of derivational morphology. Theory, methods, evidence*, 177–217. Berlin: De Gruyter.

Levy, Omer & Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, 302–308. Baltimore, MN: Association for Computational Linguistics. 10.3115/v1/P14-2050. https://www.aclweb.org/anthology/P14-2050.

Lieber, Rochelle. 2023. Ghost aspect and double plurality. on the aspectual semantics of eventive conversion and -ing nominalizations in English. In Sven Kotowski & Ingo Plag (eds.), *The semantics of derivational morphology. Theory, methods, evidence*, 15–35. Berlin: De Gruyter.

Marelli, Marco & Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review* 122(3). 485–515. 10.1037/a0039267.

Marelli, Marco, Christina L. Gagné & Thomas L. Spalding. 2017. Compounding as abstract operation in semantic space: Investigating relational effects through

a large-scale, data-driven computational model. *Cognition* 166. 207 – 224. https://doi.org/10.1016/j.cognition.2017.05.026. http://www.sciencedirect.com/science/article/pii/S0010027717301440.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *ArXiv e-prints* .

Mikolov, Tomás, Edouard Grave, Piotr Bojanowski, Christian Puhrsch & Armand Joulin. 2017. Advances in pre-training distributed word representations. *CoRR* abs/1712.09405. http://arxiv.org/abs/1712.09405.

Payne, John, Rodney Huddleston & Geoffrey K. Pullum. 2010. The distribution and category status of adjectives and adverbs. *Word Structure* 3(1). 31–81.

Plag, Ingo. 2018. *Word-formation in english* Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press 2nd edn.

Pustejovsky, James. 1995. *The generative lexicon*. Cambridge, MA: The MIT Press.

Rawlins, Kyle. 2013. On adverbs of (space and) time. In Boban Arsenijević, Berit Gehrke & Rafael Marín (eds.), *Studies in the composition and decomposition of event predicates*, 153–193. Dordrecht: Springer Netherlands. 10.1007/978-94-007-5983-1_7. https://doi.org/10.1007/978-94-007-5983-1_7.

Reddy, Siva, Diana McCarthy & Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th international conference on natural language processing*, 210–218. Chiang Mai, Thailand: AFNLP.

Sahlgren, Magnus. 2006. *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Stockholm: Stockholm University dissertation.

Schäfer, Martin. 2013. *Positions and interpretations: German adverbial adjectives at the syntax-semantics interface*. Berlin: De Gruyter.

Schäfer, Martin. 2018. *The semantic transparency of English compound nouns*. Berlin: Language Science Press. 10.5281/zenodo.1134595.

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. http://www.cssp.cnrs.fr/eiss13/.

Schäfer, Martin. 2021. From quick to quick-to-infinitival: on what is lexeme specific across paradigmatic and syntagmatic distributions. *English Language and Linguistics* 25(2). 347–377. 10.1017/S1360674320000167.

Stump, Gregory T. 1998. Inflection. In A. Spencer & A. Zwicky (eds.), *The handbook of morphology*, 13–43. London: Blackwell.

Turney, P. D. & P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37. 141–188.