

# Modeling Semantic Distance in the Pattern Dictionary of English Verbs

Silvie Cinková, Zdeněk Hlávka

Charles University,  
Faculty of Mathematics and Physics  
Ke Karlovu 3  
121 16 Praha 2

cinkova@ufal.mff.cuni.cz, hlavka@karlin.mff.cuni.cz

**Abstract.** We explore human judgments on how well individual patterns of 29 target verbs from the Pattern Dictionary of English Verbs describe their random KWICs. We focus on cases where more than one pattern is judged as highly appropriate for a given KWIC and seek to estimate the effect of event participants (arguments) being denotatively similar in two patterns, considering all pair combinations in a given lemma. We compare this effect to the effect of several contextual features of the KWICs, the effect of paired PDEV implicatures implying each other, and the effect of belonging to a given lemma. We show that the lemma effect is still stronger than any feature going across lemmas we have examined so far, so that each verb appears to be a little universe in its own right.

## 1 Introduction

Since many verbs are perceived as highly polysemous, their senses are both difficult to determine when building a lexicon entry and to distinguish in context when performing Word Sense Disambiguation (WSD). An alternative to verb senses is *usage patterns* coined by Hanks in the *Pattern Dictionary of English Verbs* (PDEV, [1], Fig. 1). Previous studies ([2], [3]) have shown that PDEV represents a valuable lexical resource for WSD, in that annotators reach good interannotator agreement despite the semantically fine-grained microstructure of PDEV.

Recently, we created a data set annotated with graded decisions (VPSGradeUp, cf. Section 2.2) from PDEV to investigate features suspected of blurring distinctions between the patterns [4]. We have been preliminarily considering features related to the KWICs independently of the lexicon design, such as *finiteness*, *argument opacity*, and *factuality*<sup>1</sup> of the target verb on the one hand, and those related to the lexicographical design of PDEV, such as *textual entailment* between PDEV *implicatures* within a lemma or *denotative*

<sup>1</sup> For explanation of terms used in this section, kindly refer to Section 2.

*similarity of the verb arguments* on the other hand.

This paper focuses on the denotative similarity of the verb arguments. We have attempted to approach it in a quantitative way by modeling it as *semantic distance* between the *corresponding syntactic slots* in pattern definitions in a PDEV lemma (henceforth *colempats*), as comprehensively described in Section 6. We compare all colempats pairwise, examining their scores in the graded decision annotation (see Section 2.2) with respect to how much they compete to become the most appropriate pattern, as well as the semantic distance between their subjects, objects, and adverbials. To quantify the comparisons, we have introduced a measure of *rivalry* for each pair of colempats. Rivalry increases, the more *appropriate* both colempats are considered for a given KWIC and the more similar their *appropriateness* scores are (see Sections 4.1 and 4.2). We have observed a significant association between high *rivalry* in paired colempats and their corresponding arguments being labeled with denotatively similar semantic labels (henceforth *semlabels*, see Section 2.1 ).

## 2 Related Work and Important Terms

### 2.1 Pattern Dictionary of English Verbs (PDEV)

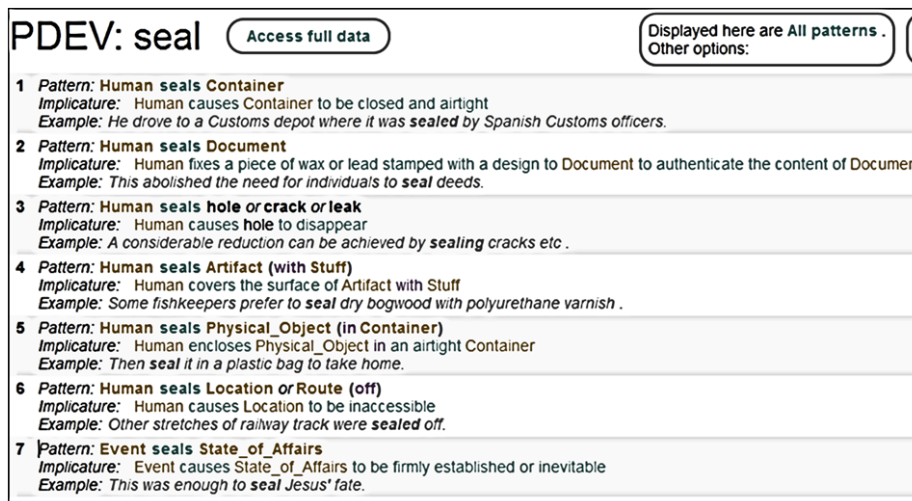
This section not only gives a brief description of PDEV, but it also introduces key terms that will be used throughout this paper.

PDEV's core idea is that a verb has no meaning in isolation; instead, it has a *meaning potential*, whose diverse components and their combinations are activated by contexts. To capture the meaning potential of a verb, the PDEV lexicographer manually clusters random KWICs into a set of prototypical usage patterns, considering the semantic and morphosyntactic similarity alike (Fig. 1).

Each PDEV *pattern* contains a *pattern definition* and an *implicature* to explain or paraphrase its meaning. Both are shaped as finite clause templates where important syntactic slots are populated with *semantic type* labels, alternatively with a set of collocates (*lexical set*), and, complementary to both, optional *semantic roles*. This paper merges them all under the umbrella term *semlabels*.

## 2.2 Graded Decisions: an Alternative WSD Setup

The graded-decision data set used in this paper draws on Erk et al. [5], who experimented with the WSD setup: instead of assigning a single sense to a given context of the key word, the annotators indicated on a Likert scale<sup>2</sup> how well each sense matched a given KWIC, allowing for ties. The data set has two subsets: WSim with graded decisions on matching relations between WordNet synsets and KWICs of 11 selected key word lemmas; and USim with graded decisions on how well two different words in two different KWICs paraphrase each other. Both displayed very good annotator correlation, suggesting graded decisions be a sensible alternative to the traditional WSD setup.



**PDEV: seal** [Access full data](#) Displayed here are All patterns .  
Other options:

- 1 Pattern: Human seals Container**  
*Implicature:* Human causes Container to be closed and airtight  
*Example:* He drove to a Customs depot where it was **sealed** by Spanish Customs officers.
- 2 Pattern: Human seals Document**  
*Implicature:* Human fixes a piece of wax or lead stamped with a design to Document to authenticate the content of Document  
*Example:* This abolished the need for individuals to **seal** deeds.
- 3 Pattern: Human seals hole or crack or leak**  
*Implicature:* Human causes hole to disappear  
*Example:* A considerable reduction can be achieved by **sealing** cracks etc .
- 4 Pattern: Human seals Artifact (with Stuff)**  
*Implicature:* Human covers the surface of Artifact with Stuff  
*Example:* Some fishkeepers prefer to **seal** dry bogwood with polyurethane varnish .
- 5 Pattern: Human seals Physical\_Object (in Container)**  
*Implicature:* Human encloses Physical\_Object in an airtight Container  
*Example:* Then **seal** it in a plastic bag to take home.
- 6 Pattern: Human seals Location or Route (off)**  
*Implicature:* Human causes Location to be inaccessible  
*Example:* Other stretches of railway track were **sealed** off.
- 7 Pattern: Event seals State\_of\_Affairs**  
*Implicature:* Event causes State\_of\_Affairs to be firmly established or inevitable  
*Example:* This was enough to **seal** Jesus' fate.

Fig. 1. PDEV entry

## 2.3 Verb Finiteness

Finiteness is a morphosyntactic category associated with verbs. Virtually all verbs appear in finite as well as infinite forms when used in context. A finite verb form is such a verb form that expresses person and number. Languages differ in whether these categories are expressed morphologically (e.g. by affixes or stem vowel changes) or syntactically (obligatorily complemented with a noun/pronoun expressing these categories explicitly). Finite forms are typically all indicative and conditional forms, as well as some imperative forms, e.g.

<sup>2</sup> Likert scale is a psychometric scale used in opinion surveys. It enables the respondents to scale their agreement/disagreement with a given opinion.

*reads, are reading, (they) read, čtu, gehst, allons!*. Infinite forms are infinitives (*to read, to have read, to be heard, to have been heard*) and participles along with gerunds and *supines* (*reading, known, deleted, försvunnit*). The grammars of many languages know diverse other finite as well as infinite verb forms. Infinite forms typically allow more argument omissions than finite forms: *to go to town* vs. *\*went to town*. This suggests that descriptions of events rendered by infinite verb forms may be more vague, and, in terms of annotation, more prone to match several different patterns/senses at the same time. Verb finiteness is easy to determine, and therefore it was only annotated by one annotator in our data set.

#### **2.4 Argument Opacity**

Argument opacity typically, but not necessarily, relates to verb finiteness. By argument opacity we mean how many arguments relevant for disambiguation of the target verb are either omitted in the context (e.g. subject in infinitive) or ambiguous or vague. Ambiguous and vague arguments are often arguments expressed by personal pronouns that refer to entities mentioned distantly from the target verb, sometimes even not directly, but by longer chains of pronouns (so-called *coreference* or *anaphora chains*), or arguments expressed by indefinite or negative pronouns. Some examples of opaque verb contexts:

*The Greater London Council was ABOLISHED in 1986. (Who abolished it?)*

*The company's ability to adapt to new opportunities and capitalize on them depends on its capacity to share information and involve everyone in the organization in a systemwide search for ways to improve, ADJUST, adapt, and upgrade. (Who exactly adjusts what?)*

This feature, too, was annotated by a single annotator in our data set. The categories were “opaque subject”, “opaque object”, “opaque arguments” (i.e.; more slots at the same time, including adverbials) or empty field.

#### **2.5 Textual Entailment between PDEV Implicatures**

Textual Entailment, mentioned in the Introduction, is the key notion of Recognizing Textual Entailment (RTE), a computational-linguistic

discipline coined by Dagan et al. [6]. The task of RTE is to determine, “given two text fragments, whether the meaning of one text can be inferred (entailed) from another text. More concretely, the applied notion of textual entailment is defined as a directional relationship between pairs of text expressions, denoted by T the entailing ‘text’ and by H the entailed ‘hypothesis’. We say that T entails H if, typically, a human reading T would infer that H is most probably true”. So, for instance, the text *Norway’s most famous painting, ‘The Scream’ by Edvard Munch, was recovered yesterday, almost three months after it was stolen from an Oslo museum* entails the hypothesis *Edvard Munch painted ‘The Scream’* [6]. We have pursued a double RTE annotation of pairs of PDEV implicatures, measured the interannotator agreement, and investigated its effect on pattern distinctions [7].

## 2.6 Word Embeddings

To compute the semantic distance between semlabels, we used *word embeddings*. Word embeddings are vector representations of the individual words in a corpus. Each vector dimension represents a word from that corpus. The more similarly words are distributed in the corpus, the more similar the directions of their respective vectors are; that is, the mutual semantic similarity of words is quantified by the (cosine) similarity of their vector representations. Cosine similarity renders the correlation coefficient between these two vectors, ranging from -1 to 1. *Word embeddings* is, loosely speaking, a quantitative expression of the Firthian “knowing the word by the company it keeps”. The embeddings used in this paper are based on Word2Vec [8], a neural network trained to reconstruct linguistic contexts of words. We use its implementation for R, *text2vec* [9].

## 3 Graded Decisions on Verb Usage Patterns

### 3.1 VPS-GradeUp

The VPS-GradeUp data set draws on Erk’s experiments with paraphrases (USim). It consists of both graded-decision and classic-WSD annotation of 29 randomly selected PDEV lemmas: *seal, sail, distinguish, adjust, cancel, need, approve, conceive, act, pack, embrace, see, abolish, advance cure, plan, manage, execute, answer, bid, point, cultivate, praise, talk, urge, last, hire, prescribe, and murder*.

Each lemma comes with 50 KWICs processed by three annotators in parallel.

In the graded-decision part, the annotators judged for each pattern how well it described a given KWIC, on a Likert scale. In the WSD part, each KWIC was assigned one best-matching pattern. The entire data set contains WSD judgments on 1,450 KWICs, corresponding to 11,400 graded decisions. A more detailed description of VPS-GradeUp is given by Baisa et al. [4]. Fig. 2 presents the most essential annotation elements.

| ID of graded judgment:<br>lemma-sentence-pattern<br>ID number | graded judgments<br>on Likert scale<br>as a 1-7 range |      |       | WSD judgments<br>- ID's of the best<br>matching patterns |        |        | concordance - the target verb is in capital letters                         |
|---|---|------|-------|--|--------|--------|---|
|   | LikA  | LikE | LikSC | WSDA   | WSD EK | WSD SC |   |
| X   | LikA LikE LikSC WSDA WSD EK WSD SC KWIC               |      |       |  |        |        |   |
| abolish:Sent_1.1:Pattern_1                                    | 7   | 2    | 4     | 2  | 3      | 3      | interstate and foreign travel in aid of racketeering activities and         |
| abolish:Sent_1.1:Pattern_2                                    | 5   | 5    | 4     | 2  | 3      | 3      | conspiracy . -- Los Angeles Times Mitterrand warns of rapid                 |
| abolish:Sent_1.1:Pattern_3                                    | 7   | 7    | 6     | 2  | 3      | 3      | change . By Anna Tomforde and Michael Farr PRESIDENT                        |
| abolish:Sent_2.1:Pattern_1                                    | 4   | 7    | 4     | 3  | 1      | 3      | Francois Mitterrand , the first head of state of the wartime Allies to      |
| abolish:Sent_2.1:Pattern_2                                    | 3   | 2    | 4     | 3  | 1      | 3      | visit East Germany , said yesterday that the existence of two               |
| abolish:Sent_2.1:Pattern_3                                    | 7   | 6    | 6     | 3  | 1      | 3      | sovereign German states could not be ' ABOLISHED ' at a                     |
| abolish:Sent_3.1:Pattern_1                                    | 7   | 6    | 2     | 1  | 2      | 2      | stroke ' . Reflecting French anxiety over German reunification ,            |
| abolish:Sent_3.1:Pattern_2                                    | 5   | 7    | 7     | 1  | 2      | 2      | Mr Mitterrand said the two Germanys were jointly responsible for            |
| abolish:Sent_3.1:Pattern_3                                    | 5   | 6    | 2     | 1  | 2      | 2      | stability in Europe . ' German unity depends first of all on the            |
| abolish:Sent_4.1:Pattern_1                                    | 3   | 6    | 2     | 1  | 1      | 1      | unclassified German people &hellip; and France will not be against it . But |
| abolish:Sent_4.1:Pattern_2                                    | 1   | 6    | 2     | 1  | 1      | 1      | unclassified Germany must be aware that it has to keep the balance in       |
| abolish:Sent_4.1:Pattern_3                                    | 1   | 6    | 2     | 1  | 1      | 1      | unclassified Europe , ' Mr Mitterrand said                                  |
| abolish:Sent_5.1:Pattern_1                                    | 7   | 7    | 6     | 1  | 1      | 1      | There can be unestimated value in the same sense that there can             |
| abolish:Sent_5.1:Pattern_2                                    | 5   | 2    | 2     | 1  | 1      | 1      | the international community . In 1983 a Government initiative to            |
| abolish:Sent_5.1:Pattern_3                                    | 3   | 2    | 6     | 1  | 1      | 1      | the international community . In 1983 a Government initiative to            |

Fig. 2. Human judgments in Graded-Decisions and WSD tasks. Each line contains one graded judgment by all annotators. The WSD judgments repeat as many times as there are patterns to judge for each KWIC.

## 4 Measures of Appropriateness and Rivalry

### 4.1 Appropriateness

The appropriateness of a pattern for a given KWIC line is based on the triple of annotation judgments and conflates their sum and standard deviation in this formula:

$$Appropriateness = \sum (x) - \frac{sd(x)}{3.5}$$

The function returns values in the range of 3 to 21. The 3.5 coefficient is roughly the maximum standard deviation (sd) possible with three judgments ranging from 1 to 7. (The x value must be a natural number ranging from 1 to 7 and the sum must be the sum of exactly 3 such x.) Appropriateness reflects both the mean and the dispersion of the judgments, unlike mere mean or median.

#### 4.2 Rivalry

To compare the competition between PDEV patterns in pairs, we have introduced *rivalry*. Rivalry always concerns the appropriateness rates for a pair of patterns of one lemma (*colempts*, see Section 1), being computed for all pairs. Rivalry increases with the *appropriateness* of each *colempt* (Section 4.1) and with decreasing difference between the appropriateness values in the given *colempt* pair: the higher the rivalry, the more the two patterns compete for becoming selected as the best match in the WSD annotation. The rivalry function is simple:

$$Rivalry = \max(appr_{pair}) - \max(\max(appr_{pair}) - \min(appr_{pair})) = \min(appr_{pair})$$

Under  $appr_{pair}$  we understand the two computed appropriateness values of patterns in a *colempt* pair:  $\max(appr_{pair})$  and  $\min(appr_{pair})$ . They represent the higher and the lower appropriateness, respectively. Hence, rivalry is defined as the difference between the higher appropriateness value and the difference between that and the lower appropriateness value, which boils down to the lower appropriateness.

It is to be emphasized that rivalry is always computed *on a given KWIC*. Hence we cannot tell e.g. the rivalry between *abandon\_1* and *abandon\_3* in general, but we get one rivalry value of this *colempt* pair for each of the 50 KWICs.

Measuring rivalry is interesting, even though we have not yet abstracted from individual KWICs; it enables us to quickly and consistently identify cases of pattern overlap for further analysis of both the design of the patterns and of contextual features in the KWICs affected.

## 5 Lemma-wise Cluster Analysis of Patterns and Concordances according to Appropriateness

The appropriateness function allows us a first visual overview of the individual lemmas regarding how well and how many colempts matched each individual KWIC. Using a standard clustering algorithm, we created a heatmap diagram for each lemma. Fig. 4 shows heatmaps for six selected lemmas. The heatmaps revealed striking differences between the individual lemmas, leaving the impression that each verb behaves in its own way. Nonetheless, we hazarded a coarse division into three groups:

Unproblematic lemmas, such as *murder* and *hire*, represent verbs whose PDEV patterns are distinct. For *murder*, the data contained only usages covered by Pattern 1. Two KWICs were not covered well by any pattern, although Pattern 1 marginally matched in the second last KWIC (white). In case of *hire*, the PDEV patterns were also distinct, with all three even occurring.

Lemmas with competing patterns. The lemmas *approve* and *last* reveal high pattern rivalry – an indication that something either in the entry design or in the contexts makes sharp distinction impossible and is potentially harmful for the WSD-interannotator agreement outcome. We carried out a manual analysis of these cases, which yielded a pool of possible features that increase rivalry, on one of which we focus in this paper.

Lemmas with many KWICs uncaptured by the current PDEV patterns. For instance, the lemmas *pack* and *seal* contain a non-trivial proportion of KWICs in which all patterns display low appropriateness by the current PDEV patterns. This suggests a problem in the entry. Two preliminary explanations are that either the KWICs contain unknown usages and that the entry ought to be complemented with new pattern(s), or that a number of patterns contain a relevant interpretation aspect, but not the ones perceived as important.

## 6 Semantic Distance between Semlabels in PDEV

### 6.1 Selectional Preferences of Verb Uses Modeled by Synslots and Semlabels

The manual inspection of clusters (Section 5) revealed that KWICs often contain objects that happen to match two semantic types at the same time, due to regular polysemy [10] or semantic coercion [11]. These semantic types can be the only aspect in which two pattern



definitions differ. When, moreover, the implicatures of these two pattern definitions *entail* one another (see Section 2.5), the two patterns (colempats) become intuitively hard to distinguish, as is repeatedly illustrated by high rivalry scores of such colempat pairs. In our sample, this happened most prominently in colempats *cancel\_2/cancel\_6* and *approve\_2/approve\_1*. (see Fig.3).

|   |   |
|---|---|
| 1 | Pattern: Human or Institution approves <b>Plan or System or Rule</b><br>Implicature: Human = Authority Figure or Institution states formally that they agree with a Plan or proposed System or Rule +<br>Example: social services committee approved a plan to 'externalise' services worth £12 million |
| 2 | Pattern: Human or Institution approves <b>Document</b><br>Implicature: Human = Authority Figure or Institution states formally that they agree with and accept the contents of Document<br>Example: delegates approved an executive consultation paper on the block vote and reform of the conference.  |

**Fig. 3. PDEV entry of the verb *approve*. Both the two colempats are transitive, and their subjects are populated with the semantic types Human and Institution. Their pattern definitions differ only in the semantic types of their direct objects (Plan/System/Rule vs. Document). Whenever a KWIC contains a noun such as *contract*, *agreement*, and *treaty* as its direct object, these two colempats become indistinguishable, since Plan/System/Rule are more often than not semantically associated with Document, and the implicatures entail each other, too.**

Therefore, when considering factors potentially increasing rivalry, the similarity of selectional preferences is among the most prominent suspects. We introduce a measure of *semantic distance*, which operationalizes the similarity of selectional preferences to quantify its effect on colempat rivalry.

The concept of *semantic distance between corresponding synslots in a colempat pair* is based on the following understanding of the PDEV entry, as already sketched out in Section 2.1: each PDEV entry consists of patterns. Patterns consist of pattern definitions and implicatures. Each pattern definition consists of a main predicate constituted by the target verb (entry lemma) and a number of arguments (syntactic slots, *synslots*, with different syntactic functions), which are populated by *semilabels*. The lemma-pattern combination (e.g. *abandon\_1*) is called *lempat*. Lempats with the same lemma are called *colempats*. A *colempat pair* is a pair thereof. Verb arguments are called *synslots*. Semilabels populating each synslot along with the syntactic functions of the synslots represent the selectional preferences of the target verb in a given pattern. Selectional preferences in verbs describe “knowledge about possible and plausible fillers for a predicate’s argument positions” ([12], p. 723). Any clustering of verb occurrences according to their semantic similarity is bound to rely on the selectional preferences, and PDEV entries capture them explicitly. Based on the Distributional Hypothesis ([13]), we can assume that the more similar

the selectional preferences of two colempats, the more semantically related the colempats are. Consequently, more similar colempats should display higher rivalry than less similar colempats.

## 6.2 Semlabels in Synslots and Corresponding Synslot Pairs

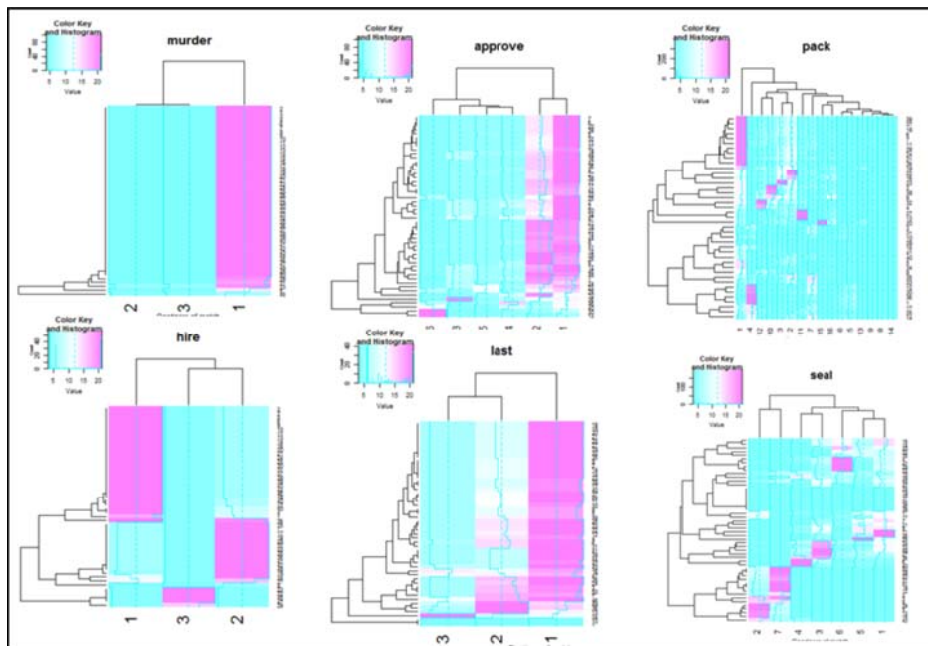
When modeling the selectional preferences of the target verb in colempats by semlabels in synslots, the corresponding synslots were easily extracted by their syntactic labels provided in PDEV. The actual problem was the *semantic distance* between a pair of corresponding synslots in itself: synslots are often populated with more than one semlabel (cf. Object in Pattern 6, Fig. 1), whereas we needed a one-number summary. First we computed a semantic-distance score for *each possible pair of semlabels* (more in Section 6.3) and mapped these scores on observed pairs of semlabels populating the pairs of corresponding synslots. For instance, in a pair of colempats where the subject of Colempat A was populated by HUMAN and ANIMAL, while that of Colempat B by INSTITUTION, LAND, and ANIMAL, we used the semantic-distance scores for the following semlabel pairs: HUMAN-INSTITUTION, HUMAN-LAND, HUMAN-ANIMAL, ANIMAL-INSTITUTION, ANIMAL-LAND, and ANIMAL-ANIMAL. From these 6 scores, we needed a one-number summary to render the semantic distance between the entire corresponding synslots, which we computed as described in Section 6.4.

## 6.3 Computing the Semantic Distance between Individual Semlabels

To determine the semantic distance between the individual semlabels used in PDEV, we exploited the Distributional Hypothesis ([13]) by transforming a corpus of pattern definitions and implicatures from the entire PDEV into a vector space with each token represented by a vector, whose dimensions reflected the co-occurrence with other tokens in the space. To render tokens as vectors, we used *tex2vec* ([9]), an implementation of word embeddings [8] for R. Although *text2vec* offers a large general-language vector space, we preferred to train our own on the text of PDEV patterns, since we were interested in the distributional similarity of the words (e.g. HUMAN) *used as PDEV labels* rather than in their regular usage.

Before training the corpus, we cleaned the PDEV data to capture all semlabels and minimize their variants. Markup variation was less a problem in the approx. 200 semantic types listed in the PDEV

documentation than in lexical sets, semantic roles, and diverse grammatical markers. First we cleared the text of punctuation, converted all tokens to lower case, and erased numerical indices (e.g. HUMAN 1). We also had the corpus lemmatized using MorphoDiTa [14]. Then we extracted all multi-word semlabels (e.g. HUMAN GROUP) and collapsed their strings into one token with underscores



**Fig. 4.** Heatmaps of selected lemmas. The lines and columns represent the KWICs and the PDEV patterns, respectively. The color appropriateness scores range from turquoise (low appropriateness) to purple (high appropriateness). Each line is indexed with a unique KWIC ID on the right and is therefore easily tracked back in the corpus for a more detailed analysis.

(obtaining e.g. *human\_group*). This cleaned and lemmatized corpus was put into *text2vec* to obtain vector representations for each token.

When computing the vector space of PDEV with *text2vec*, we set a minimum frequency threshold to 5 and limited the number of vector dimensions to 50. Having obtained a 50-dimension vector for each token, we computed their pairwise similarities with cosine. Cosine ranges between -1 and 1 and expresses the angle between the two vectors. The more the angle differs from 0, the more dissimilar the two vectors are. We transformed the results into a range between 0 and 1 by adding 1 to each cosine value and dividing the result by 2. The resulting structure was a distance matrix for almost all pairs of semlabels (for

exceptions see Section 6.4).

At this point, we had a mutual-similarity score for nearly each possible pair of PDEV semlabels, and thus we could establish the mutual similarity of all semlabel pairs within a corresponding synslot pair (cf. the example in Section 6.2 – we would have 6 cosine similarity scores to characterize the relation between the subjects of the Colempat A & Colempat B pair).

#### 6.4 Determining the Semantic Distance between Entire Synslots

At that point, we needed a one-number summary for each pair of corresponding synslots in a pair of colempats to be derived from the cosine similarities of individual semlabels. We defined it as the *Hausdorff distance* between the first corresponding synslot and the second corresponding synslot. The Hausdorff distance is commonly used to model the distance between two subsets of a metric space, in our case the two synslots, if we pretend that the semlabels that each of them contains are points in the metric space. To compute the Hausdorff distance between Colempat A and Colempat B, we took one semlabel of Colempat A after another; for each we computed its distance to each semlabel in Colempat B and saved the shortest distance. From these distances we took the maximum. For this computation we adapted the *HD* function from the *polydect* R package [15] to immediately process distances between points across the subsets instead of deriving them from coordinates of each point as the original function would have done. Before we had to solve three minor issues:

1. Transformation of the individual cosine similarities to render *distance* rather than *similarity* to conform to the concept of Hausdorff distance; that is, have high score mean *dissimilarity*;
2. Model cases of only one colempat having the given synslot;
3. Dealing with semlabels for which we had not obtained a cosine similarity score.

Transformation of similarity into distance. We subtracted the cosine similarity value from 1 to get a number that would *decrease* with growing similarity instead of increasing. The resulting structure was a distance matrix for semlabel pairs, required by the adapted *HD* function.

Synslot mismatch between colempats. In many pairs, the inventory of

synslots was different for each colempat; e.g. one colempat had a direct object while the other not. We modeled the syntactic mismatch as a *tenfold of the maximum distance observed in the matrix*. This is an approximation of the common annotator experience that differences in argument structure usually help distinguish patterns at the first sight, while distinctions based on semlabels in corresponding synslots are often blurred – especially when each synslot is populated by several semlabels and some of them are present in both, or when they are very general (e.g. PHYSICAL OBJECT), or if it is difficult to find for them a reasonable hyperonym (one-word and not too general).

Missing cosine similarity/distance for a pair of semlabels. The automatic cleaning of the PDEV corpus before training the vector space by *text2vec* (Section 6.3) captured virtually all PDEV semantic types, but even so the vector space missed many members of lexical sets and semantic roles (see Section 2.1), because they were too rare to pass the frequency threshold. We set the cosine similarity/distance of pairs consisting of one or two unmatched semlabels to the *mean cosine similarity/distance observed in the data set*.

We computed the Hausdorff distances for each syntactic function of synslots separately. As a last step, we computed their sum for each colempat pair. This number rendered the semantic distance between the entire colempats.

## **7 Association between Rivalry and Semantic Distance of Corresponding Syntactic Slots**

As a first approximation, we have run the Pearson correlation test with rivalry and the semantic distance between colempats (i.e. the sum of the semantic distances of all observed synslots in the colempat pair). The test detected a statistically significant negative correlation (p-value  $\approx 0.005$ ). This conforms to our intuition: the larger the semantic distance, the smaller the rivalry between patterns. However, the effect is very small (95% confidence interval between -0.002 and -0.004).

We also examined possible correlations between rivalry and the semantic distance in the individual syntactic functions of synslot pairs (i.e. subjects, objects, etc.), but the results were not statistically significant for any of them.

## 8 Discussion and Future Work

We have observed a statistically significant but tiny effect of the semantic distance of colempats on pattern rivalry – a far smaller one than we had expected. The annotator intuition certainly suggests that, when considering a verb as an event with participants, the cognitive characteristics of event participants are important for the perception of the event itself, and we may believe that we are able to compare how similar or different the typical participants of two events are, but there is no straightforward way to operationalize this perception. We decided to rely on *text2vec* as a robust state-of-the-art device for lexical semantic analysis, which goes beyond simple distributional similarity yielded by binary distance matrices, but we naturally do not know what exactly happens behind the scenes of the neural network. Then, even if a vector representation had been the most adequate approach, we might have lost much of the actual effect size by observing the *vectors of labels* in the pattern definitions instead of *vectors of the actual words* populating the KWICs in the corresponding synslots. *Labels* were easily extracted from PDEV, while the extraction of *words* populating relevant synslots would have been much more complex, involving coreference resolution, dealing with coordinated elements, extensive manual checks, etc. Also, our approach to mismatches in the argument structure between patterns has been rather crude. In fact, our approach assigns the tenfold of the largest semantic distance to all cases where the Hausdorff distance could not be computed, which includes a synslot present in one but absent in the other colempat as well as a synslot missing in both and thus clearly blurs an important difference which could have added to the effect.

Apart from the semantic distance, we have been preliminarily examining other features suspect of increasing rivalry, such as the explicit presence/absence of relevant arguments (*argument opacity*) and *finiteness* of the target verb in the KWICs, along with a very preliminary manual annotation of *mutual implications of paired pattern implicatures* (only 2 annotators, 0.45 Cohen's  $\kappa$ ). A statistically significant linear model predicting rivalry (Table 1) finds all these predictors significant, but the semantic distance focused in this paper turns out weakest. Interestingly, verb finiteness (promising more explicit contexts) does not help distinguish between patterns but in fact increases rivalry. Considering the argument opacity, the model grows slightly more powerful when the opacity of subjects and objects is

singled out, with opaque object being the most rivalry increasing predictor from the opacity family (coeff. 1.42). The most effective rivalry increaser turns out to be implicatures implying each other, raising each rivalry unit by 2.55 (to the extent we can believe averaged triple human judgments on entailment). We have also been considering the factuality of the target predicates (for which we have used verb finiteness here as a primitive proxy), but a pilot annotation has yielded poor interannotator agreement, making results based on such data highly speculative.

All the aforementioned predictors are apparently not general enough to beat the effects of individual lemmas, as Table 2 reveals: most lemmas are significant, have high coefficients, and increase the predictive power of the model (cf. R-squared in both): despite efforts to find universal features, each verb appears to remain a little universe in its own right.

| Call:           |  |  |  |  |  |
|-----------------|--|--|--|--|--|
| lm(formula = ri |  |  |  |  |  |
| + z_args.opa    |  |  |  |  |  |
| Residuals:      |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |
|                 |  |  |  |  |  |

**Table 1:** Linear model predicting rivalry from semantic distance, verb finiteness, argument opacity and mutual implications between paired implicatures





## Acknowledgements

This work was supported by the Czech Science Foundation Grant No. GA 15-20031S and by the LINDAT/CLARIN project No. LM2015071 of the MEYS CR. All computation has been performed in R [16].

## References

- [1] Hanks, P. (2000). Pattern Dictionary of English Verbs [<http://pdev.org.uk>]. UK.
- [2] Cinková, S., Holub, M., Rambousek, A. and Smejkalová, L. (2012). A database of semantic clusters of verb usages. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association, Istanbul, Turkey. p. 3176–3183.
- [3] Cinkova, S., Krejčová, E., Vernerová, A. and Baisa, V. (2016). Graded and Word-Sense-Disambiguation Decisions in Corpus Pattern Analysis: a Pilot Study. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, France. p. 23-28.
- [4] Baisa, V., Cinková, S., Krejčová, E. and Vernerová, A. (2016). VPS-GradeUp: Graded Decisions on Usage Patterns. *LREC 2016 Proceedings*, Portorož, Slovenia.
- [5] Erk, K., McCarthy, D. and Gaylord, N. (2009). Investigations on Word Senses and Word Usages. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Association for Computational Linguistics, Suntec, Singapore. p. 10–18.
- [6] Dagan, I., Dolan, B., Magnini, B. and Roth, D. (2010). Recognizing textual entailment: Rational, evaluation and approaches - Erratum. *Natural Language Engineering*, **16**, 105.
- [7] Cinková, S. and Vernerová, A. (to appear). Are Annotators' Word-Sense-Disambiguation Decisions Affected by Textual Entailment between Lexicon Glosses? *ITAT 2017*.
- [8] Mikolov, T., tau Yih, W. and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *HLT-NAACL*, The Association for Computational Linguistics. p. 746–51.

- [9] Selivanov, D. (2016). text2vec: Modern Text Mining Framework for R [<https://CRAN.R-project.org/package=text2vec>].
- [10] Martínez Alonso, H., Sandford Pedersen, B. and Bel, N. (2013). Annotation of regular polysemy and underspecification. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Sofia, Bulgaria. p. 725–730.
- [11] Pustejovsky, J., Rumshisky, A., Moszkowicz, J. and Batiukova, O. (2009). Gml: Annotating argument selection and coercion. *IWSD-8: Eighth International Conference on Computational Semantics*,.
- [12] Erk, K., Padó, S. and Padó, U. (2010). A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, **36**, 723–63. [https://doi.org/10.1162/coli\\_a\\_00017](https://doi.org/10.1162/coli_a_00017)
- [13] Harris, Z.S. (1970). *Papers in structural and transformational linguistics*. Reidel.
- [14] Straka, M., Hajič, J. and Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In: Calzolari N, Choukri K, Declerck T, Grobelnik M, Maegaard B, Mariani J, et al., editors. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association, Paris, France. p. 4290–4297.
- [15] Zhihua, S. (2008). R Polydect. [<https://github.com/cran/polydect/blob/master/R/HD.R>, 22-01-2017].
- [16] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.