J. Hlaváčová (Ed.): ITAT 2017 Proceedings, pp. 5–14
CEUR Workshop Proceedings Vol. 1885, ISSN 1613-0073, © 2017 S. Cinková, A. Vernerová

Are Annotators' Word-Sense-Disambiguation Decisions Affected by Textual Entailment between Lexicon Glosses?

Silvie Cinková, Anna Vernerová

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25
118 00 Praha 1
Czech Republic
ufal.mff.cuni.cz
{cinkova,vernerova}@ufal.mff.cuni.cz

Abstract: We describe an annotation experiment combining topics from lexicography and Word Sense Disambiguation. It involves a lexicon (*Pattern Dictionary of English Verbs, PDEV*), an existing data set (*VPS-GradeUp*), and an unpublished data set (*RTE in PDEV Implicatures*). The aim of the experiment was twofold: a pilot annotation of Recognizing Textual Entailment (RTE) on PDEV implicatures (lexicon glosses) on the one hand, and, on the other hand, an analysis of the effect of Textual Entailment between lexicon glosses on annotators' Word-Sense-Disambiguation decisions, compared to other predictors, such as finiteness of the target verb, the explicit presence of its relevant arguments, and the semantic distance between corresponding syntactic arguments in two different patterns (dictionary senses).

1 Introduction

A substantial proportion of 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). To tackle the polysemy of verbs, diverse lexicon designs and annotation procedures have been deployed. One alternative way to classic verb senses (e.g. ***to blush*** *- to redden, as from embarrasment or shame*[1]) is *usage patterns* coined in the *Pattern Dictionary of English Verbs (PDEV)* [9], which will be explained in Section 2.2. Previous studies [3], [4] 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. This paper focuses on cases challenging the interannotator agreement in WSD and considers the contribution of *textual entailment* (Section 2.3) to interannotator confusion.

We draw on a data set based on PDEV and annotated with graded decisions (cf. Section 2.4) to investigate features suspected of blurring distinctions between the patterns [1]. We have been preliminarily considering features related to language usage independently of the lexicon design, such as finiteness and argument opacity of the target verb on the one hand, and those related to the lexicographical design of PDEV, such as semantic relations between implicatures within a lemma or denotative similarity of the verb arguments, on the other hand (see Section 3 for definitions and examples).

This paper focuses on a feature related to PDEV's design (see Section 2.3), namely on textual entailment between implicatures in pairs of patterns of the same lemma entry (henceforth *colempats*, see Section 3.1 for definition and more detail).

We pairwise compare all colempats, examining their scores in the graded decision annotation with respect to how much they compete to become the most appropriate pattern, as well as the scores of presence of textual entailment between their implicatures. To quantify the comparisons, we have introduced a measure of *rivalry* for each pair. The more the rivalry increases, the more appropriate both colempats are considered for a given KWIC[2] and the more similar their appropriateness scores are (see Section 3.2).

We confirm a significant positive association between rivalry in paired colempats and textual entailment between their implicatures.

2 Related Work

2.1 Word Sense Disambiguation

Word Sense Disambiguation (WSD)[18] is a traditional machine-learning task in NLP. It draws on the assumption that each word can be described by a set of word senses in a reference lexicon and hence each occurrence of a word in a given context can be assigned a word sense. Bulks of texts have been manually annotated with word senses to provide training data. Nevertheless, the extensive experience from many such projects has revealed that even humans themselves do not do particularly well interpreting word meaning in terms of lexicon senses, despite specialized lexicons designed entirely for this task: the English WordNet [8], PropBank [14], and OntoNotes

[1] http://www.dictionary.com/browse/blush

[2] *KWIC = key word in context: a corpus line containing a match to a particular corpus query*

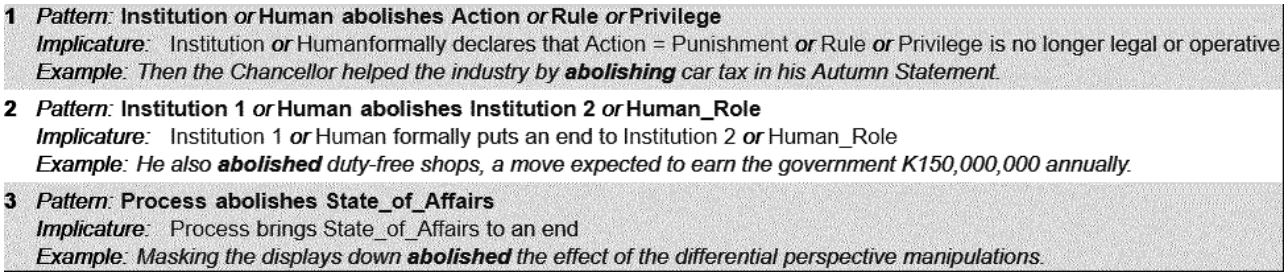

Figure 1: PDEV entry with three patterns

Word Senses [12], to name but a few. Although the annotators, usually language experts, have neither comprehension problems nor are they unfamiliar with using lexicons, their interannotator agreement has been notoriously low. This in turn makes the training data unreliable as well as the evaluation of WSD systems harder.

Attempts have been made to increase the interannotator agreement by testing each entry on annotators while designing the lexicon [12], as well as word senses were clustered post hoc on the other hand (e.g. [17]), but even lexicographers have been skeptical about lexicons with hardwired word senses for NLP([13, 15]).

2.2 Pattern Dictionary of English Verbs (PDEV)

The reasoning behind PDEV is that a verb has no meaning in isolation; instead of word senses, it has a *meaning potential*, whose diverse components and their combinations are activated by different contexts. To capture the meaning potential of a verb, the PDEV lexicographer manually clusters random KWICs into a set of prototypical usage patterns, considering both their semantic and morphosyntactic similarity. Each PDEV ***pattern*** contains a ***pattern definition*** (a finite clause template where important syntactic slots are labeled with ***semantic types***) and an ***implicature*** to explain or paraphrase its meaning, which also is a finite clause (Fig. 1). The PDEV implicature corresponds to *gloss* or *definition* in traditional dictionaries.

The ***semantic types*** (e.g. *Human, Institution, Rule, Process, State_of_Affairs*) are the most typical syntactic slot fillers, although the slots can also contain a set of collocates (a *lexical set*) and *semantic roles* complementary to semantic types. The semantic types come from an approximately 250-item shallow ontology associated with PDEV and drawing on the Brandeis Semantic Ontology (BSO), [19]. The notion of semantic types, lexical sets, and semantic roles (altogether dubbed ***semlabels***) is, in this paper, particularly relevant for Section 3.5.

2.3 Recognizing Textual Entailment (RTE)

Recognizing Textual Entailment (RTE) is a computational-linguistic discipline coined by Dagan et al. [5]. 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'* [5].

2.4 Graded Decisions on Verb Usage Patterns: VPS-GradeUp

The VPS-GradeUp data set draws on Erk's experiments with paraphrases (USim)[7]. VPS-GradeUp 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[3] in parallel.

In the graded-decision part, the annotators judged each pattern for how well it described a given KWIC, on a Likert scale[4]. 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 (50 sentences $\times$ 29 lemmas $\times$ sum of patterns). A more detailed description of VPS-GradeUp is given by Baisa et al.[1].

Fig. 2 shows a VPS-GradeUp sample of three KWICs of the verb *abolish* (see Fig. 1 to refer to the lexicon entry). Columns 1, 2, and 3 identify the pattern ID, lemma, and sentence ID, respectively. Columns 4-6 and 7-9 contain the graded and WSD decisions by the three annotators, respectively. Column 10 contains the annotated KWIC, which for Sentence 1 reads: *President Mitterrand said yesterday that the existence of two sovereign German states*

[3] linguists, professional but non-native English speakers

[4] Likert scale is a psychometric scale used in opinion surveys. It enables the respondents to scale their agreement/disagreement with a given opinion.

Pattern ID	Lemma	SentID	GD1	GD2	GD3	WSD1	WSD2	WSD3	KWIC
1	abolish	1.1	7	2	4	2	3	3	intersta
2	abolish	1.1	5	5	4	2	3	3	intersta
3	abolish	1.1	7	7	6	2	3	3	intersta
1	abolish	2.1	4	7	4	3	1	3	more fo
2	abolish	2.1	3	2	4	3	1	3	more fo
3	abolish	2.1	7	6	6	3	1	3	more fo
1	abolish	3.1	7	6	2	1	2	2	it is mis
2	abolish	3.1	5	7	7	1	2	2	it is mis
3	abolish	3.1	5	6	2	1	2	2	it is mis

Figure 2: A VPS-GradeUp annotation sample

could not be 'ABOLISHED at a stroke'. On the third table row, Pattern 3 was judged as maximally appropriate by Annotator 1 and 2; Annotator 3 gave one point less. In the WSD part, Annotator 1 voted for Pattern 2, while Annotators 2 and 3 preferred Pattern 3.

3 Important Concepts

3.1 Lempats and Colempats

To begin with, we introduce the concept of ***lempats*** and ***colempats***. The lemma-pattern combination, as represented by Columns 1 and 2 in Fig. 2, is called ***lempat***. All lempats sharing a common lemma are called ***colempats***. That is, the table presents the colempats *abolish_1*, *abolish_2*, and *abolish_3* and their three annotator judgments on the sentences 1.1, 2.1, and 3.1. A pair of patterns, such as *abolish_3* and *cancel_1*, are also two lempats which we could compare, but they are not *co*lempats, because each belongs to a different lemma (*abolish* vs. *cancel*).

Fig.2, Columns 4-6, shows that, on Sentence 1.1, the annotators disagree in their WSD judgments (Annotator 2 and 3 voted for Pattern 3, but Annotator 1 preferred Pattern 2). This is probably caused by the fact that Annotators 1 and 2 had also regarded Pattern 2 as somewhat appropriate (Row 2). Interestingly, Annotator 1 even considered Pattern 1 maximally appropriate for the given KWIC, unlike the others, but eventually did neither vote for this pattern nor for Pattern 3. As with all manual annotations, human error cannot be a priori dismissed, but even the oddest judgments mostly turn out to come with a plausible explanation.

How do then the graded decisions map on the WSD judgments, if they do at all? To perform quantitative observations of how much two patterns compete in the WSD annotation, we needed a measure of ***appropriateness*** of a given pattern for a given KWIC across all annotators (see Section 3.2), along with yet another measure to tell which two patterns were the most serious competitors (***rivalry***, see Section 3.3).

Having a lempat, we need to measure its ***appropriateness*** for a given KWIC. To be able to examine the mapping between the graded-decisions and the WSD annotation, we observe ***rivalry*** within each possible pair of colempats for a given KWIC.

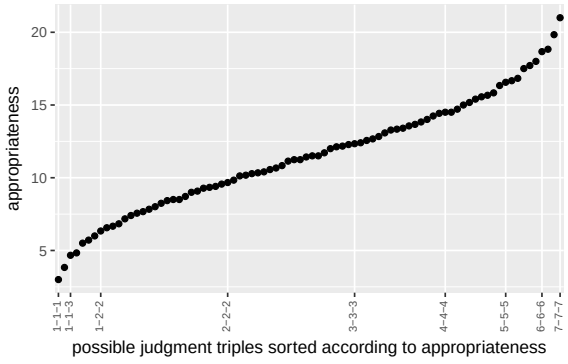

Figure 3: Shape of the appropriateness function

3.2 Appropriateness

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

$$Appropriateness = \sum x - sd(x)/3.5$$

The function returns values ranging from 3 to 21. These are all possible sums of judgments by three annotators on 7-point Likert scales: as a minimum and maximum, a pattern can obtain 1 and 7 from each annotator, respectively. The 3.5 coefficient is roughly the maximum standard deviation (sd) possible with three judgments ranging from 1 to 7. Compared to mean or median, appropriateness discounts triples with higher dispersion. We made no effort to generalize this measure beyond the specific setup of this particular experiment with 7-point Likert scales and three annotators, and therefore 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.

Fig. 3 shows the shape of the curve. The x-axis contains all possible combinations of 1–7 triples with replacement, sorted in ascending order according to their corresponding appropriateness value. The curve is designed to reflect the opinion strength by steepness: the extreme positions indicate stronger opinions than central scale positions. Therefore the dispersion of the judgments affects appropriateness more strongly at both ends of the scale than around its center.

3.3 Rivalry

To compare the competition between PDEV pairs of patterns, we have introduced rivalry. Rivalry always concerns the appropriateness rates for a pair of patterns of one lemma (colempats), being computed for all pairs. Rivalry increases with the appropriateness of each colempat and with decreasing difference between the appropriateness values in the given colempat 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:
$Rivalry = \max(appr_{pair}) - (\max(appr_{pair}) - \min(appr_{pair})) = \min(appr_{pair})$.

Under $appr_{pair}$ we understand the two computed appropriateness values of patterns in a colempat pair: $\max(appr)$ and $\min(appr)$. 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. The idea behind rivalry is that, given the nature of the WSD annotation task, we are interested in colempats competing at the positive rather than at the negative end of the scale.

It is to be emphasized that rivalry is always computed *on a given KWIC*. Hence we cannot immediately tell e.g. the rivalry between *abandon_1* and *abandon_3* in general, but we get one rivalry value of this colempat 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 identify cases of pattern overlap for further analysis of both the design of the patterns and of contextual features in the KWICs affected.

3.4 Corresponding Synslots

As Fig. 1 shows, the syntactic slot fillers of the target verb in the pattern definition are described by semantic labels (henceforth ***semlabels***). Each syntactic slot (henceforth ***synslot***) also has a syntactic function in the clause: *subject, object, adverbial*, or *complement*. When observing synslots across a pair of colempats, we check whether a synslot with a particular syntactic function (e.g. object) is present in both colempats in the pair. When this is the case, these two synslots are called *corresponding synslots*.

3.5 Semantic Distance between Corresponding Synslots

In a past experiment, we measured how the rivalry is impacted by the extent to which the sets of synslot fillers in a colempat pair are cognitively similar. We observed a statistically significant (yet weak) positive association. The synslot fillers were represented by the semlabels. To obtain their semantic similarity, we first built a corpus of pattern definitions and implicatures from the entire PDEV. Then we fed this corpus to a neural network, which created a vector representation for each word.[5] We defined the mutual similarity of each two words by the cosine similarity of their vectors. For more details see [2].

[5] *text2vec* [22] – an implementation of the *word2vec* [16] neural network for R. The original task on which the neural network was trained was guessing context around each word. Its practical use draws on the so-called Distributional Hypothesis[10], according to which words with similar context distribution are more semantically related than those with dissimilar context distribution. The network creates a vector representation of each word, with the dimensions of each word vector being the other words. The similarity of two vectors reflects the distributional (and hence semantic) similarity of two words.

3.6 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. *reads, are reading, (they) read, čtu, čtěte, chtěl by, 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* (incorrect). 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.

3.7 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 follow:

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?)

4 Textual Entailment Annotation

4.1 Annotation Procedure

Three annotators[6] obtained paired implicatures of colempats of each target verb and judged whether one entailed the other (specifying the direction), or whether the entailment is bidirectional or absent (cf. Section 2.3). The definition of entailment used here is based on the conception of *textual entailment* coined by Dagan et al. (cf. **RTE**, [6]). For the purposes of this paper, we collapsed the annotation into entailment presence-absence judgments.

4.2 Annotation Results

The three annotators processed 1,091 implicature pairs (both implicatures always belonged to the same lemma). The annotators were allowed to see the entire entry including example sentences, but they were told to focus on the implicatures. Their pairwise percentual agreement scores were 73.8, 74.6, and 83.3. Fleiss' kappa was moderate: 0.41. While RTE annotations usually reach 0.6 desired for semantic annotations, our worse result is understandable: the PDEV implicatures are much more abstract and hence more vague than regular text, since the arguments of the target verb are described by ontology labels. See an example of two pattern implicatures of the verb *seal*:

> *Human covers the surface of Artifact with Stuff.*
> *Human encloses Physical Object in an airtight Container.*

We merged the three annotations by taking the means of "yes" and "no" judgments replaced with 1 and 0, respectively. With this setup, the judgments could acquire only four values: 0, 0.33, 0.66, and 1. We treated them as values of a categorical ordinal variable.

Fig. 4 shows the annotation results for each lemma. To facilitate the reading, we displayed the judgments as the number of annotator votes in favor of entailment. The proportions are compared to the verb *see*, with its 192 colempat pairs. The annotator disagreement is represented by 1 and 2 votes. In terms of proportions within the given lemma, the most problematic verbs were the small[7] verbs *abolish, cancel, hire*, and *praise*, along with the large verbs *act*, *point*, and *talk*.

A typical colempat pair with full agreement on no implicature entailment is e.g. *act_10-12*. The example also includes the pattern definition for better understanding:

Pattern: *Phrasal verb.* Human *acts* Event *or* Human Role *or* Emotion *out.*
Implicature: Human *performs* Role*, not necessarily sincerely, or behaves as if feeling* Emotion.

Pattern: Idiom. Human *acts* POSDET age.
Implicature: *Human* behaves in a manner appropriate to their *age*.

Although both these events have something to do with behavior, we can neither normally assume that someone who acts their emotions out is necessarily behaving according to their age, nor the other way round. Thus we observe no implicature entailment relation between these two colempats.

A typical colempat pair with full agreement on implicature entailment is e.g. *act_1-9*. This example also illustrates that entailment does not require synonymy. The second implicature entails the first; that is, when an actor performs a character on theater, they are – normally – pursuing a motivated action by pretending to be a particular character for their audience.

Pattern: Human *or* Institution *or* Animal *or* Machine *acts*
Implicature: Human *or* Institution *or* Animal *or* Machine = Agent *performs a motivated* Action

Pattern: Human *acts (*Role*) (in* Performance*)*
Implicature: Human *plays* Role = Theatrical *(in* Performance*)*
However, the general nature of the implicatures makes the entailment annotation difficult. Below follows an example where one annotator voted against the entailment, the *act_1-11* pair. The *act_1* colempat is listed in the previous example. Here follows the *act_11* colempat:

Pattern: Phrasal verb. Human *acts up.*
Implicature: Human *behaves badly.* Human *is typically a naughty child.*.

The annotators clearly disagree on whether bad behavior is normally perceived as a motivated action. They were instructed to focus only on the implicature. At the same time, they were allowed to see the entire entry. Most likely with this entry, two annotators were influenced by the very verb *act up*. The verb *act up* suggests a motivated action (e.g. start screaming to attract attention, this being perceived as bad manners in the given situation). The plain implicature leaves leeway for considering non-motivated actions (can very young infants act consiously?) or non-actions perceived as bad behavior (even a child can behave badly by *not* acting e.g. to someone's help).

The reasons for annotator disagreements are very diverse, including obvious annotation errors, and their ex post analysis is often subjective. We show a case from still the same verb, *act_1-12*. See *act_1* above again, *act_12* follows:

[6] linguists familiar with PDEV as well as with RTE, professional but non-native English speakers

[7] i.e. with a small number of colempat pairs

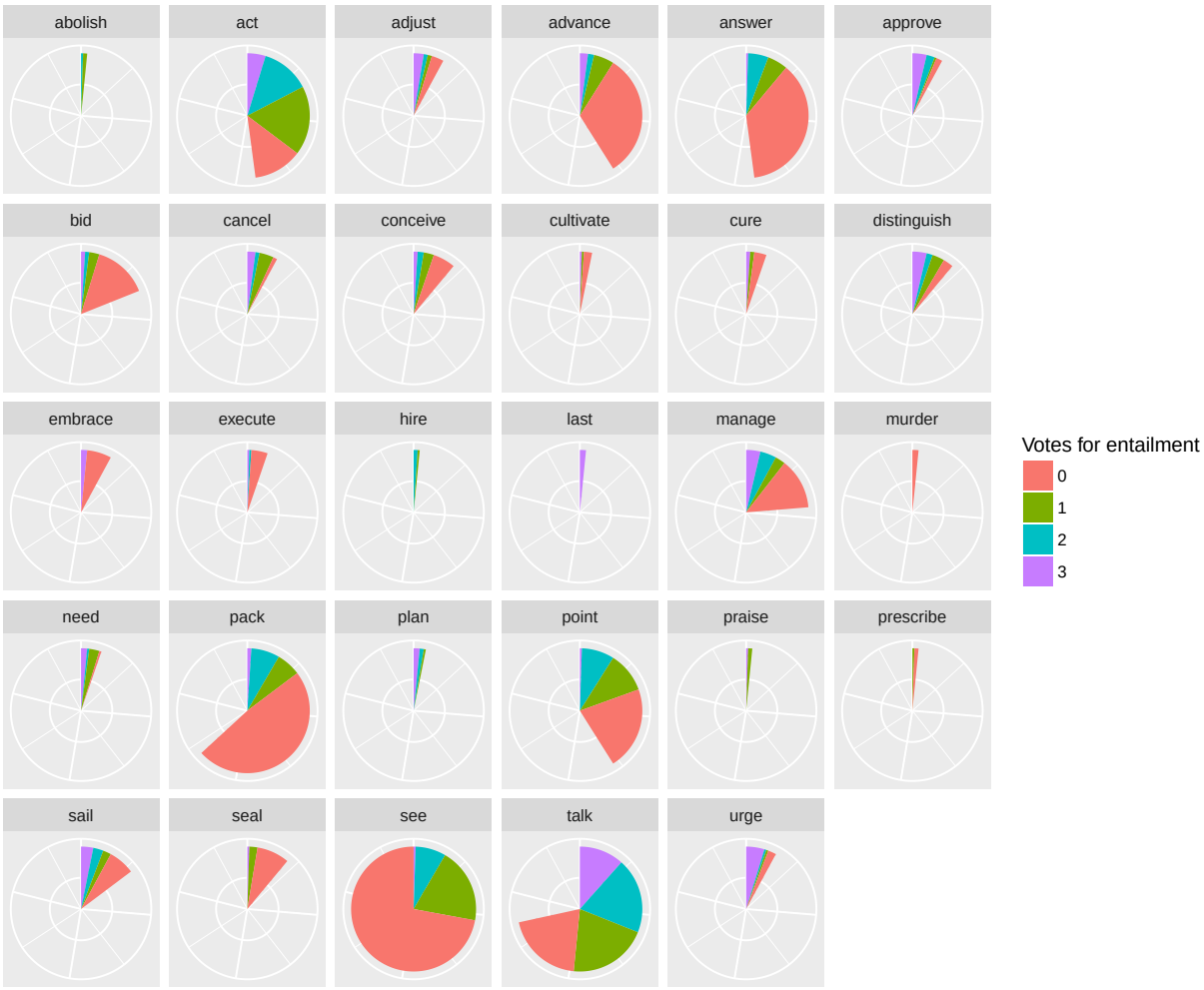

Figure 4: Proportional distribution of entailment judgments in individual lemmas, relating to the verb *see*, which has 190 possible colempat pairs.

Pattern: Phrasal verb. Machine *acts up.*
Implicature: Machine *fails to function correctly.*

Here, the pro-entailment decision by two annotators was most likely motivated by the fact that *act_1* specifies *Machine* as *Agent* and lets it perform a *motivated* action. Then, naturally, even malfunction can be a motivated action. The remaining annotator, on the other hand, did not accept malfunction as a motivated action.

Often the uncertainty lies in the interpretation of the semantic types. For instance, *hire_1-2* differ in the object of hiring. In the first colempat it is *Human* or *Institution*, whose services are obtained for payment. In the second colempat it is a *Physical Object*, which is used for an agreed period of time against payment. In real life, this corresponds to e.g. hiring a gardener to take care of a garden vs. hiring an apartment. Such two events naturally do not entail each other in any way. However, the general wording of implicatures allows one annotator to regard the use of a *Physical Object* against payment as a service provided by a *Human* or *Institution*. Consider e.g. *Mary hires John to let her live in an apartment that belongs to him.*.

5 Association between Implicature Entailment and Rivalry

5.1 Linear Model with Rivalry Abstracted from Individual KWICs

While the textual entailment is observed between two colempat implicatures independently of their instances in corpus evidence, rivalry is always associated with both the given pair of colempats and the KWIC, with respect to which their appropriateness was judged (cf. Section 3.3). We had 50 rivalry scores for each colempat pair, since the VPS-GradeUp annotators were judging the appropriateness of each pattern for each of the 50 KWICs per lemma.

For each colempat pair we selected the KWIC on which their rivalry was highest.

To examine the association between the textual entailment between implicatures and rivalry between colempats, we built this linear regression model using the `lm()` function in base R [20].

```
Call:
lm(formula = abstr_rivalry ~ factor(numMeans),
data = all_entailment)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17193 -0.02510 -0.01854  0.01936  0.38500

Coefficients:
            Estimate Std.Error t-value Pr(>|t|)
(Intercept) 0.167961 0.002754  60.984  < 2e-16 ***
numMeans0.3 0.041712 0.005380   7.753 2.05e-14 ***
numMeans0.6 0.084065 0.006107  13.765  < 2e-16 ***
numMeans1   0.155577 0.007134  21.806  < 2e-16 ***
---
Signif.
codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '1

Residual standard error: 0.06808 on 1087 degrees of
freedom
Multiple R-squared:  0.348,     Adjusted R-squared:
0.3462
F-statistic: 193.4 on 3 and 1087 DF,
p-value: < 2.2e-16
```

According to the Adjusted R-squared, it explains approximately 35% of the variance of rivalry. This means that entailment is quite a strong predictor. Apart from that, the individual coefficient values in the model nicely confirm our assumption that entailment causes rivalry increase: One vote for entailment (i.e. value 0.3) increases the rivalry coefficient by 0.04, two votes increase it by 0.08, and three votes increase it by 0.15. Their individual standard errors are one decimal point smaller than the coefficients themselves, which means that they would not overlap; that is, every single entailment vote matters. The model is highly significant, and so are all levels of the entailment values (p-value always much smaller than 0.05). This, along with the randomness of lemma selection, means that we can expect the results to be similar with other equally annotated verbs.

5.2 Linear Model with Rivalry on All KWICs

We ran the same experiment also without abstracting from the KWICs. The model is still highly significant, but extremely weak (explaining about 20% of the rivalry variation). This makes sense, since this time we also included observations with the same entailment conditions but lower rivalry. This way we introduced KWICs where the positive effect of entailment can have been overcome by the negative effect of other predictor values, which we have not included into the model.

```
Call:
lm(formula = rivalry ~ factor(entail_numMeans),
data = vyplyv)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8233 -0.8442 -0.5540  0.2811 15.1558

Coefficients:
                  Estimate Std.Err t-value Pr(>|t|)
(Intercept)       3.55396  0.01149  309.38 <2e-16 ***
fctr(numMeans)0.3 1.20615  0.02244   53.74 <2e-16 ***
fctr(numMeans)0.6 1.12526  0.02547   44.17 <2e-16 ***
fctr(numMeans)1   3.26938  0.02976  109.84 <2e-16 ***
---
Signif.
codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '1

Residual standard error: 2.008 on 54534 degrees
of freedom
Multiple R-squared:  0.1985,
Adjusted R-squared:  0.1984
F-statistic:  4501 on 3 and 54534 DF,
p-value: < 2.2e-16
```

6 Discussion

We have observed a statistically significant positive effect of textual entailment of colempat implicatures on the rivalry between colempats in PDEV. It is evidently not the only cause of increasing rivalry, as shown by the weakness of the model, but has the strongest effect. The implicature is the part of patterns that corresponds to classic word senses in traditional lexicons. This suggests that the traditional conception of word senses as semantic definitions rather than usage definitions is very useful in sense distinction, whenever annotators agree. On the other hand, like with traditional word senses, the interannotator agreement is low. Like traditional word senses shaped as lexicon glosses/definitions, the implicatures are too abstract to bode well for interannotator agreement. The issue persists even when the annotation task is set up as an RTE task rather than recognizing synonymy and mutual exclusivity (according to which traditional WSD annotation decisions are taken)[8].

Apart from the textual entailment, we have been preliminarily examining other features suspect of increasing rivalry, such as the explicit presence/absence of relevant arguments (argument opacity, Section 3.7), semantic distance between labels used in corresponding syntactic positions within a colempat pair (based on text2vec [22]), and finiteness of the target verb in the KWICs (Section 3.6). A statistically significant linear model predicting rivalry finds all these predictors significant (Fig. 5).

However, the textual entailment turns out to be most effective rivalry increaser, raising each rivalry unit by 2.55 (to the extent we can believe averaged human judgments on implication). Interestingly, verb finiteness (promising more explicit contexts) does not help distinguish between patterns but in fact *increases* rivalry (i.e. blurs distinctions between colempats). Considering the argument opacity, opaque object is the most rivalry increasing predictor from the opacity family (coeff. 1.42). We have also been considering the factuality[9] of the events described by the tar-

[8] The RTE annotation task would possibly benefit from graded annotation by many annotators like word-similarity/relatedness experiments, e.g. [11].

[9] [21]

```
Call:

lm(formula = rivalry ~ w2vec_hsdrff_Sum + z_finite + z_args.opaque
                     +  entail_mean, data = rival)
Residuals:
                    Min        1Q        Median      3Q         Max
                    -44 145    -0.7944   -0.4442     0.3024     161 824
Coefficients:
                    Estimate   Std. Err  t value     Pr(>|t|)
(Intercept)         385 483    0.04893   78 785      < 2e-16    ***
w2vec_hsdrff_Sum    -0.01200   0.00110   -10 908     < 2e-16    ***
z_finitey           0.34715    0.01713   20 264      < 2e-16    ***
z_args.opaquey      123 175    0.23520   5 237       1.64e-07   ***
z_args.opaqueobj    141 808    0.36389   3 897       9.75e-05   ***
z_args.opaquesubj   0.20601    0.02265   9 097       < 2e-16    ***
entail_mean         255 232    0.02152   118 592     < 2e-16    ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.992 on 54531 degrees of freedom
Multiple R-squared:  0.2112, Adjusted R-squared:  0.2111
F-statistic:  2433 on 6 and 54531 DF,  p-value: < 2.2e-16
```

Figure 5: A linear model predicting rivalry from semantic distance, verb finiteness, argument opacity and textual entailment

get 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 even more speculative than those of textual entailment between colempat implicatures, so we have not included it in the model.

All the aforementioned predictors are apparently not general enough to beat the effects of individual lemmas: most lemmas are significant, have high coefficients, and increase the predictive power of the model in Fig. 6; cf. R-squared in both models: despite efforts to find universal linguistic features, each verb appears to remain a little universe in its own right.

7 Conclusion

We have confirmed that textual entailment between two colempat implicatures increases rivalry between these colempats. We also see that the more the annotators agree on the presence of entailment, the stronger its effect is: it grows with each annotator vote to even double when all three annotators agree, compared to two annotators.

8 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.

References

[1] Vít Baisa, Silvie Cinková, Ema Krejčová, and Anna Vernerová. VPS-GradeUp: Graded Decisions on Usage Patterns. In *LREC 2016 Proceedings*, Portorož, Slovenia, May 2016.

[2] Silvie Cinková and Zdeněk Hlávka. Modeling Semantic Distance in the Pattern Dictionary of English Verbs. *Jazykovedný časopis*, to appear.

[3] Silvie Cinková, Martin Holub, Adam Rambousek, and Lenka Smejkalová. A database of semantic clusters of verb usages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3176–3183, Istanbul, Turkey, 2012. European Language Resources Association.

[4] Silvie Cinkova, Ema Krejčová, Anna Vernerová, and Vít Baisa. Graded and Word-Sense-Disambiguation Decisions in Corpus Pattern Analysis: a Pilot Study. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, 2016. European Language Resources Association (ELRA).

[5] Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii, 2009.

[6] Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2013.

[7] Katrin Erk, Diana McCarthy, and Nicholas Gaylord. Investigations on Word Senses and Word Usages. In *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*, pages 10–18, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[8] C. Fellbaum, J. Grabowski, and S. Landes. Performance and confidence in a semantic annotation task. In *WordNet: An Electronic Lexical Database*, pages 217–238. Cambridge (Mass.): The MIT Press., Cambridge (Mass.), 1998. 00054.

[9] Patrick Hanks. *Pattern Dictionary of English Verbs*. http://pdev.org.uk/, UK, 2000.

[10] Zellig Harris. Distributional structure. *Word*, 23(10):146–162, 1954. 01136.

[11] Samer Hassan and Rada Mihalcea. Cross-lingual semantic relatedness using encyclopedic knowledge. In *In EMNLP 2009. Association for Computational Linguistics*, 2009. 00044.

[12] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. 00346.

[13] Adam Kilgarriff. "I Don't Believe in Word Senses". *Computers and the Humanities*, 31(2):91–113, 1997.

[14] Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40, 2008. 00139.

[15] Ramesh Krishnamurthy and Diane Nicholls. Peeling an Onion: The Lexicographer's Experience of Manual Sense Tagging. *Computers and the Humanities*, 34:85–97, 2000. 00000.

[16] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL*, pages 746–751. The Association for Computational Linguistics, 2013.

[17] Roberto Navigli. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 105–112, Sydney, Australia, 2006.

[18] Roberto Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69, February 2009. 00697.

[19] James Pustejovsky, Catherine Havasi, Jessica Littman, Anna Rumshisky, and Marc Verhagen. Towards a generative lexical resource: The Brandeis Semantic Ontology. In *Proceedings of the Fifth Language Resource and Evaluation Conference*, 2006.

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2014.

[21] Roser Saurí and James Pustejovsky. Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text. *Comput. Linguist.*, 38(2):261–299, June 2012. 00068.

[22] Dmitriy Selivanov. *text2vec: Modern Text Mining Framework for R*. R Foundation for Statistical Computing, 2016.

```
Call:

lm(formula = rivalry ~ w2vec_hsdrff_Sum + z_finite + z_args.opaque
+ entail_mean + lemmas, data = rival)
Residuals:
Min                 1Q            Median       3Q        Max
-59 380             -0.7319       -0.1572      0.1800    159 160

Coefficients:
(Intercept)         73522017      0.1532486    47 976    < 2,00E-16    ***
w2vec_hsdrff_Sum    -0.0003296    0.0011026    -0.299    0.7650
z_finitey           0.1455412     0.0161223    9 027     < 2,00E-16    ***
z_args.opaquey      0.5009538     0.2156543    2 323     0.0202        *
z_args.opaqueobj    0.2547546     0.3372427    0.755     0.4500
z_args.opaquesubj   0.0532390     0.0217931    2 443     0.0146        *
entail_mean         1.8818343     0.0217515    86 515    < 2,00E-16    ***
lemmasact           -3.7824581    0.1512543    -25 007   < 2,00E-16    ***
lemmasadjust        -2.7990281    0.1619012    -17 288   < 2,00E-16    ***
lemmasadvance       -4.2765385    0.1515831    -28 213   < 2,00E-16    ***
lemmasanswer        -4.2405515    0.1508514    -28 111   < 2,00E-16    ***
lemmasapprove       -3.1989511    0.1621494    -19 728   < 2,00E-16    ***
lemmasbid           -3.9934306    0.1548404    -25 791   < 2,00E-16    ***
lemmascancel        -2.8219473    0.1621358    -17 405   < 2,00E-16    ***
lemmasconceive      -2.2675897    0.1583548    -14 320   < 2,00E-16    ***
lemmascultivate     -2.7869641    0.1816034    -15 346   < 2,00E-16    ***
lemmascure          -3.8352304    0.1688616    -22 712   < 2,00E-16    ***
lemmasdistinguish   -2.9855282    0.1580461    -18 890   < 2,00E-16    ***
lemmasembrace       -3.3944366    0.1624320    -20 898   < 2,00E-16    ***
lemmasexecute       -2.2898572    0.1686455    -13 578   < 2,00E-16    ***
lemmashire          -3.4752011    0.2089821    -16 629   < 2,00E-16    ***
lemmaslast          -1.2512805    0.2101987    -5 953    2.65e-09      ***
lemmasmanage        -2.9204488    0.1531206    -19 073   < 2,00E-16    ***
lemmasmurder        -3.5778433    0.2101611    -17 024   < 2,00E-16    ***
lemmasneed          0.2515703     0.1692646    1 486     0.1372
lemmaspack          -4.3029164    0.1501447    -28 658   < 2,00E-16    ***
lemmasplan          -1.7058389    0.1817877    -9 384    < 2,00E-16    ***
lemmaspoint         -3.2865632    0.1512721    -21 726   < 2,00E-16    ***
lemmaspraise        -0.2847921    0.2091486    -1 362    0.1733
lemmasprescribe     -0.2380621    0.2091980    -1 138    0.2551
lemmassail          -1.8942963    0.1567161    -12 087   < 2,00E-16    ***
lemmasseal          -3.8569221    0.1581722    -24 384   < 2,00E-16    ***
lemmassee           -4.3824168    0.1498710    -29 241   < 2,00E-16    ***
lemmastalk          -3.7339660    0.1502380    -24 854   < 2,00E-16    ***
lemmasurge          -0.8541827    0.1623112    -5 263    1.43e-07      ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.809 on 54503 degrees of freedom
Multiple R-squared:  0.3497,        Adjusted R-squared:  0.3493
F-statistic: 862.2 on 34 and 54503 DF,  p-value: < 2.2e-16
```

Figure 6: Linear model enriched with lemmas as predictors