What Do Graded Decisions Tell Us about Verb Uses

Silvie Cinková[1], Ema Krejčová[1], Anna Vernerová[1], Vít Baisa[2]

Charles University in Prague[1], Masaryk University Brno[2]

e-mail: {cinkova, krejcova,vernerova}@ufal.mff.cuni.cz, xbaisa@fi.muni.cz

Abstract

We work with 1450 concordances of 29 English verbs (50 concordances per lemma) and their corresponding entries in the Pattern Dictionary of English Verbs (PDEV). Three human annotators working independently but in parallel judged how well each lexical unit of the corresponding PDEV entry illustrates the given concordance. Thereafter they selected one best-fitting lexical unit for each concordance – while the former setup allowed for ties (equally good matches), the latter did not. We measure the interannotator agreement/correlation in both setups and show that our results are not worse (in fact, slightly better) than in an already published graded-decision annotation performed on a traditional dictionary. We also manually examine the cases where several PDEV lexical units were classified as good matches and how this fact affected the interannotator agreement in the best-fit setup. The main causes of overlap between lexical units include semantic coercion and regular polysemy, as well as occasionally insufficient abstraction from regular syntactic alternations, and eventually also arguments defined as optional and scattered across different lexical units despite not being mutually exclusive.

Keywords: Word Sense Disambiguation; usage patterns; computational lexicography; graded decisions; Likert scales; Corpus Pattern Analysis; Pattern Dictionary of English Verbs; regular polysemy; coercion

1 Introduction

1.1 The Importance of Modelling Lexical Meaning

Modelling of lexical meaning is interesting with respect to a range of language-related disciplines, e.g.: 1) human lexicography; 2) design and use of lexical databases in NLP; 3) ontology design and use. Semantic modelling of a lexical item or a concept as an inventory of senses (hotly debated in 2, but inevitable in 3 and, particularly, in 1) tends to involve different relations between the senses, spanning from mutual exclusivity to broad overlap. In certain tasks (e.g. Word Sense Disambiguation), sense overlaps pose a problem, while in cognitively oriented computational tasks, they represent a valuable source of world knowledge (e.g. *institutions often act as humans*). An analysis of the causes of sense overlaps can promote a more conscious tailoring of these reference sources to their different purposes as well as facilitate automatic sense clustering/classification procedures.

We report on an experiment exploring how human dictionary users match authentic verb uses to an inventory of corpus-based usage patterns, given the option of graded rating on the one hand, and being coerced into selecting one best option on the other hand. We compare the inter-annotator correlation in both setups and manually analyze the cases in which the option eventually chosen as the best fit had outcompeted others with an equally good rating.

Our work is based upon the Pattern Dictionary of English Verbs (PDEV).[1] PDEV was our fist choice, since, to the best of our knowledge, its structure provides the most elaborate lexicographical account of the usage of English verbs, and it tracks the individual lexicographical decisions most explicitly of all lexical resources we are familiar with.

[1] http://pdev.org.uk

PDEV is methodologically based on the Corpus Pattern Analysis (e.g. Hanks & Pustejovsky 2005) as well as the Theory of Norms and Exploitation (Hanks 2013) and comes with its specific terminology: the microstructural units of PDEV, called *patterns* or *pattern definitions*, are explicitly declared not to correspond to what is traditionally called *dictionary senses*.

1.2 Terminological Clarifications

Here we have to tackle a terminological issue before we go on any further. After thorough considerations, we have decided to refer to subunits of a dictionary entry as *lexical units*. We understand and respect the theoretical motivation for maintaining a difference between patterns and senses stressed by PDEV, but we consider this difference irrelevant for the purpose of this paper and need a more general term that would abstract from it. We are not investigating the differences between lexicographical approaches; on the contrary: we are looking into human syntactico-semantic clustering decisions, given a closed inventory of cluster definitions and examples, no matter on which theoretical background that inventory (that is, a dictionary entry) relies. To remain as theory-neutral as possible, we will refrain from both *pattern* and *sense*, and we will resort to *lexical unit*, which we believe to be free of the polarity that *sense* acquires in this particular research context. Another reason for avoiding the PDEV-specific terminology is that both *patterns* and *pattern definitions* seem to be used for two things simultaneously – the entire unit on the one hand and the first part of the unit on the other hand – whereas we need to refer to each separately. Even a quite recent methodological paper (El Maarouf 2013) preserves this ambiguity. Eventually, we want to stress our belief that our findings on PDEV deserve to be considered even beyond the realm of PDEV and CPA, as CPA has been along and influential for several decades by now, and its principles regarding verb entries are making their way into the verb entries of many “traditional” dictionaries.

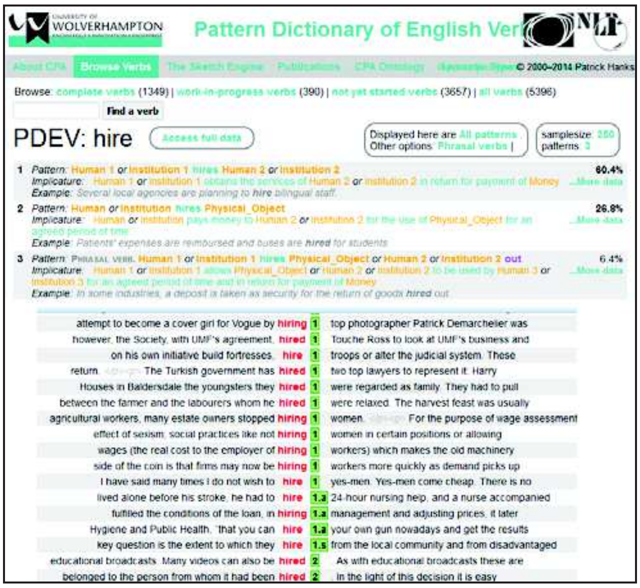
1.3 Research Outline

The paper is structured as follows: first we briefly describe PDEV, then our experimental data and the annotation procedure. We guide the reader through a simple quantitative analysis of the results, and, to conclude, we discuss the relative importance of the two main sources of overlap in the lexical units (as indicated by annotator disagreement): entry design and contextual factors. We have found several recurrent types of interference between lexical units as well as several context features that make the 1:1 match between a concordance and a lexical unit a more or less random choice within a well-motivated subgroup of the available lexical units.

2 Source Lexicon and Annotated Data

The lexicon under scrutiny is the Pattern Dictionary of English Verbs, which is meant both for human use and for computational-linguistic applications. From the human-use perspective, it is a monolingual learner’s dictionary: it captures normal uses of verbs presenting each as a combination of a *pattern* and an *implicature*. In this paper, we refer to them as *lexical units.*[2] The *pattern* (the first line in each lexical unit) is a finite clause template with the target verb as the main predicate and its relevant arguments and adjuncts directly listed or represented by ontological labels, such as *Human, Eventuality*, etc., or alternatively by structural labels (*QUOTE, THAT-clause*). The implicature is a definition or paraphrase, also formed as a finite clause with arguments represented by their ontological and structural labels. Each lexical unit is illustrated with numerous examples from BNC50, which is a subcorpus of BNC (British National Corpus 2007) compiled for the purpose of PDEV. BNC50 comprises contemporary written documents, approximately one half of BNC. For illustrations of PDEV entries and annotation see Figures Figure 1 and Figure 5.

[2] See Section 1.2.



UNIVERSITY OF WOLVERHAMPTON

Pattern Dictionary of English Ver

About CPA | Browse Verbs | The Sketch Engine | Publications | CPA Ontology | Apresjan Types © 2000–2014 Patrick Hanks

Browse: complete verbs (1349) | work-in-progress verbs (390) | not yet started verbs (3657) | all verbs (5396)

Find a verb

PDEV: hire

Access full data

Displayed here are All patterns
Other options: Phrasal verbs |

samplesize: 250
patterns: 3

1 *Pattern:* Human 1 or Institution 1 hires Human 2 or Institution 2 — 60.4%
Implicature: Human 1 or Institution 1 obtains the services of Human 2 or Institution 2 in return for payment of Money — ...More data
Example: *Several local agencies are planning to **hire** bilingual staff.*

2 *Pattern:* Human or Institution hires Physical_Object — 26.8%
Implicature: Human or Institution pays money to Human 2 or Institution 2 for the use of Physical_Object for an agreed period of time — ...More data
Example: *Patients' expenses are reimbursed and buses are **hired** for students*

3 *Pattern:* PHRASAL VERB. Human 1 or Institution 1 hires Physical_Object or Human 2 or Institution 2 out — 6.4%
Implicature: Human 1 or Institution 1 allows Physical_Object or Human 2 or Institution 2 to be used by Human 3 or Institution 3 for an agreed period of time and in return for payment of Money — ...More data
Example: *In some industries, a deposit is taken as security for the return of goods **hired** out*

attempt to become a cover girl for Vogue by	hiring	1	top photographer Patrick Demarchelier was
however, the Society, with UMF's agreement,	hired	1	Touche Ross to look at UMF's business and
on his own initiative build fortresses,	hire	1	troops or alter the judicial system. These
return. The Turkish government has	hired	1	two top lawyers to represent it. Harry
Houses in Baldersdale the youngsters they	hired	1	were regarded as family. They had to pull
between the farmer and the labourers whom he	hired	1	were relaxed. The harvest feast was usually
agricultural workers, many estate owners stopped	hiring	1	women. For the purpose of wage assessment
effect of sexism, social practices like not	hiring	1	women in certain positions or allowing
wages (the real cost to the employer of	hiring	1	workers) which makes the old machinery
side of the coin is that firms may now be	hiring	1	workers more quickly as demand picks up
I have said many times I do not wish to	hire	1	yes-men. Yes-men come cheap. There is no
lived alone before his stroke, he had to	hire	1.a	24-hour nursing help, and a nurse accompanied
fulfilled the conditions of the loan, in	hiring	1.a	management and adjusting prices, it later
Hygiene and Public Health, "that you can	hire	1.a	your own gun nowadays and get the results
key question is the extent to which they	hire	1.s	from the local community and from disadvantaged
educational broadcasts. Many videos can also be	hired	2	As with educational broadcasts these are
belonged to the person from whom it had been	hired	2	In the light of this decision it is easy

Figure 1: Hire entry and annotated concordances in PDEV.

3 Related Work

From the computational point of view, PDEV is a *semantic concordance*, i.e. a reference resource (lexicon) linked to an annotated corpus.[3] By its design it is comparable to WordNet/SemCor (Landes et al., n.d.) , FrameNet (Ruppenhofer et al. 2010), or the Proposition Bank (Palmer, Gildea, & Kingsbury 2005). Semantic concordances are mainly used as lexical resources for statistically driven computational linguistics and Natural Language Processing (NLP). The human judgments they contain represent the "ground truth" or "gold standard" that computers learn to mimic. The capability of computer systems to mimic human language-analytical judgments is trained and evaluated in a number of standard NLP tasks. One of them is the Word Sense Disambiguation (WSD). This task requires a text and a lexicon. For each word in context, the most appropriate sense from the lexicon is automatically selected. The decisions of the computer are learned from training data prepared by human experts. WSD is a tedious task and a bottleneck of many applications. After numerous studies had been conducted on this topic, it turned out that Word Sense Disambiguation is an unnatural task even for humans (Krishnamurthy & Nicholls 2000) and that the judgments of individual annotators often differ, either due to semantic overlap in the sense definitions, or due to context vagueness, which can be textual as well as grammatical. In these cases, the annotators easily disagree when forced to select exactly one best-fitting sense and discard the others, and the resulting gold-standard data become less consistent and hence less efficient for machine-learning. There have already been quite a few experiments with other ways of human modelling of the lexical meaning:

- by *clustering of concordances* according to their mutual similarity, as intuitively perceived by human annotators (Rumshisky, Verhagen, & Moszkowicz 2009)
- by *graded decisions*, i.e. scoring each lexicon sense with respect to how well it illustrates a given concordance (Erk, McCarthy & Gaylord 2009) and (Erk, McCarthy & Gaylord 2013).

Our previous research (Cinková et al. 2012) focused on the optimization of the design of a CPA-based dictionary entry to increase the interannotator agreement in the classical WSD task. In the present paper, we draw on (Rumshisky, Verhagen, & Moszkowicz 2009) and (Erk, McCarthy & Gaylord 2009 and 2013) when measuring the interannotator agreement and analyzing

[3] Cf. (Miller et al., 1993)

disagreements in graded judgments on the individual match between a target verb in the context of a random BNC50-concordance and each lexical unit of the corresponding PDEV entry. Unlike in (Cinková et al. 2012), the entries were not subject to revisions during the task and no annotation was repeated.

4 The Task

4.1 Lemma Selection and Annotator Instructions

Three independently working annotators were processing 29 randomly selected verb entries from PDEV.[4] The verb lemmas were: *abolish, act, adjust, advance, answer, approve, bid, cancel, conceive, cultivate, cure, distinguish, embrace, execute, hire, last, manage, murder, need, pack, plan, point, praise, prescribe, sail, seal, see, talk*, and *urge*. For each of the 29 verbs, the annotators got the complete corresponding PDEV entry including the original PDEV annotation made by the PDEV team, along with a 50-page online form. Each form page contained one random BNC50 concordance of the given target verb and a list of the lexical units in the corresponding PDEV entry. For each lexical entry they were supposed to mark a point on a 7-point Likert scale, answering the question "How well do these lexical units illustrate the use of the target verb in the sentence above?". The responses could range from "Exact match" to "Irrelevant". To sum this information in amount of judgments made: the final number, which exceeds 11,000 judgments, comes from 29 verbs *times* 50 sentences per verb *times* the sum of the numbers of lexical units of all verbs.
Table 1 presents the numbers of lexical units in the entries.

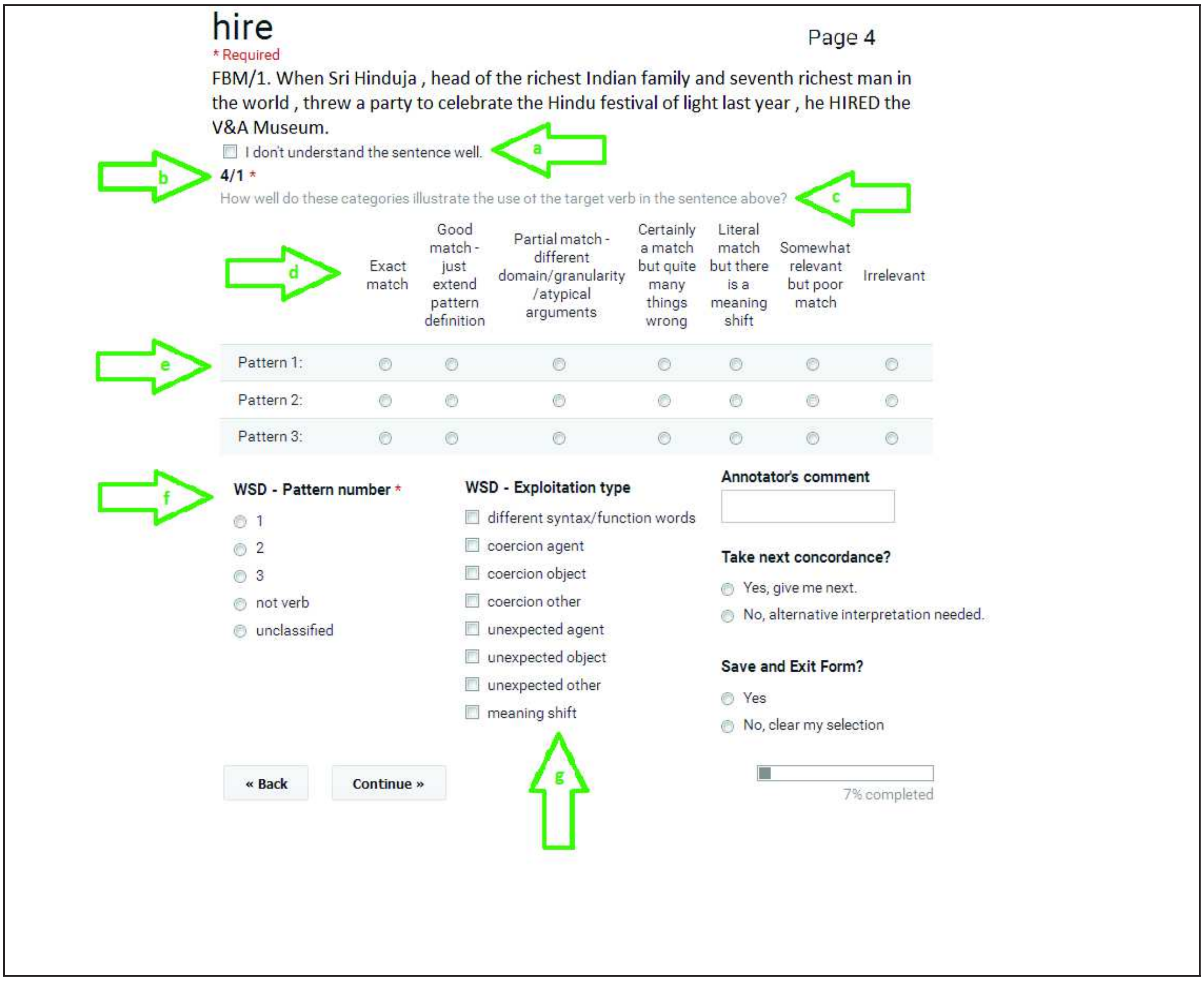
number of lexical units	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**13**	**14**	**16**	**18**	**21**
frequency	6	2	3	5	3	1	1	1	2	2	1	1	1

Table 1: Frequency distribution of the number of lexical units in the random sample of PDEV entries.

4.2 The Survey Form

The survey form had, loosely speaking, one page per concordance; that is, 50 pages; however, strictly speaking, it had 100 pages, of which almost one half remained blank, as each concordance was allowed one additional interpretation in case the concordance was ambiguous. Figure 2 presents a standard form page. It is introduced by the verb lemma (in this case *hire*) and the given concordance. The annotator was supposed to read the entire concordance carefully. If they were not sure about its correct interpretation, they were to indicate a comprehension problem in a tick box (a). Each concordance is accompanied by its identifier (unique within one verb lemma), the annotation question "How well do these categories illustrate the use of the target verb in the sentence above?" (c), and the anchor descriptions of the Likert items (d) mapping to numbers 7 - 1, with one Likert scale per lexical unit (e). The next part contains the best-fit-lexical-unit decision (f). Conforming to the Theory of Norms and Exploitations (Hanks 2013), the best-fit-lexical-unit decision is complemented by exploitation markup (g).
In this way, we get a solid comparison of the two annotation setups: the classical best-fit vs. the graded decision setup.

[4] A few selection restrictions were applied, though: the minimum number of lexical units was 3. They were not allowed to be contained in the VPS-30-En data set (cf. Cinková et al. 2012), and there had to be at least 100 unannotated concordances left in BNC50, of which we then randomly selected 50 (in order not to have the judgments in a particular sentence biased by annotations conducted by the PDEV team). As a final step, we excluded one verb lemma with more than 60 lexical units (*blow*) as a clear outlier. A careful random selection was a necessary prerequisite for any subsequent statistical analysis. For more detail see (Baisa et al. 2016).



hire

Page 4

* Required

FBM/1. When Sri Hinduja , head of the richest Indian family and seventh richest man in the world , threw a party to celebrate the Hindu festival of light last year , he HIRED the V&A Museum.

□ I don't understand the sentence well.

4/1 *

How well do these categories illustrate the use of the target verb in the sentence above?

	Exact match	Good match - just extend pattern definition	Partial match - different domain/granularity /atypical arguments	Certainly a match but quite many things wrong	Literal match but there is a meaning shift	Somewhat relevant but poor match	Irrelevant
Pattern 1:	○	○	○	○	○	○	○
Pattern 2:	○	○	○	○	○	○	○
Pattern 3:	○	○	○	○	○	○	○

WSD - Pattern number *

○ 1
○ 2
○ 3
○ not verb
○ unclassified

WSD - Exploitation type

□ different syntax/function words
□ coercion agent
□ coercion object
□ coercion other
□ unexpected agent
□ unexpected object
□ unexpected other
□ meaning shift

Annotator's comment

Take next concordance?

○ Yes, give me next.
○ No, alternative interpretation needed.

Save and Exit Form?

○ Yes
○ No, clear my selection

Figure 2: The annotation form using Google Forms.

4.3 Likert Scales

To capture the graded annotator decisions, we used 7-item Likert scales[5] with verbal anchors. The Likert scales are a standard survey technique that enables the respondents to scale their responses according to anchor points (aka Likert items). Our anchors lie on the scale “Good match – Poor match”, where the goodness of match is associated with semantic as well as morphosyntactic criteria. Our previous research as well as our previously elaborated annotation instructions (see also (Cinková & Hanks 2010)) conclude that semantic relevance is the first prerequisite for any further relevance considerations; hence the matches “worsen” gradually from morphosyntactic mismatches over the conceptual features of the arguments to semantic mismatches between the given concordance and the implicature of the lexical unit in question. An advantage of a fine-grained Likert scale is that the measured values can be converted to numbers, so that, for instance, we can calculate a middle value for a match from all annotators. In our case, we compute the median instead of the intuitively preferred mean, since we cannot be sure whether the distances between the adjacent items are identical throughout the entire scale. Another advantage of this survey design is that it allows for ties; i.e., the annotator can classify several lexical items as equally good matches for the given concordance.

[5] https://en.wikipedia.org/wiki/Likert_scale

Interannotator Agreement and Interannotator Correlation

4.4 Comparison to Related Experiments

Before manually analyzing the disagreements, we measured the inter-annotator agreement in the best-fit task and the interannotator correlation in the graded-decision task. Both measures help us estimate whether the task has a generally acceptable solution for most cases or whether the annotators were rather throwing in random decisions. In the latter case the data would be of little use for machine learning.

Since we knew from our earlier research that the interannotator agreement is particularly threatened when a verb is used in the past or present participle form in a syntactic position that makes it act as a noun or adjective (e.g. *a stolen car*), we separated out such cases and provided the agreement and correlation measurements for two datasets called Complete and VerbsOnly (Figure 3). The VerbsOnly dataset is a subset of Complete, excluding KWICs classified as "not verb" by at least one annotator, since we also wanted a view abstracting from the well-known part-of-speech ambiguity of participle verb forms and from tagging errors in the BNC. To compare our inter-annotator agreement with (Erk et al. 2009), we used Spearman's ρ. For the graded decisions on Complete, the pairwise correlations were $\rho = 0.658$, $\rho = 0.656$, and $\rho = 0.675$. For the best-fit-lexical-unit decisions in VerbsOnly, the pairwise correlations were $\rho = 0.785$, $\rho = 0.743$, and $\rho = 0.792$. The Fleiss' kappa for the best-fit-lexical-unit task is 0.76 (the VerbsOnly set), ignoring the exploitation markup. All correlations are highly significant with $p < 2.2e\text{-}16$. The observed correlations are even higher than those reported by (Erk et al. 2009). In practice, Fleiss' kappa above 0.6 on a semantic task is considered a reasonable agreement. [6] Hence the results for VerbsOnly were very satisfactory.

5 Manual Annotation Analysis

To analyze the inter-annotator agreement of the graded-decision setup for all annotators simultaneously, we converted the anchor points to numbers 1 (*irrelevant*) – 7 (*perfect match*) and treated them as values of an ordinal variable, observing the *median* and *range* (i.e., the difference between the maximum value and the minimum value) of each decision triple. The first sight at the data revealed that, except for the verb *murder*, the annotators had used the entire Likert scale. The extreme median values 1-2 (poor matches) and 7 (perfect match) occurred generally more frequently and with lower ranges than the moderate values; that is, the annotators had a good agreement on good candidates for the best-fit lexical unit as well as on extremely poorly matching lexical units. No substantial personal bias was observed.

To perform a manual analysis of lexical unit overlap, we singled out *all graded decisions in which the judged lexical unit had also appeared as the best-fit* (Figure 4). A graded decision is identified by the KWIC ID ("SentID"), the lemma, and the lexical unit number ("LikID"). We divided the graded decision triples into three sets according to the number of annotators who had selected that particular lexical unit as the best-fit lexical unit, referring to these groups as WSD3, WSD2, and WSD1. We have quantitatively compared these three sets[7] with respect to the median values. All three sets turned out to significantly differ from one another, with the difference between WSD2 and WSD3 much less significant than in WSD1 vs. WSD3 and WSD1 vs. WSD2. At the same time, the typical range is smallest in WSD3 and highest in WSD1. This is perfectly in line with the intuition that the more of the annotators rate a lexical unit high in their graded decisions, the more likely this lexical unit is to be unanimously voted for as the best fit.

[6] The Fleiss' kappa scores lie between 0 and 1.

[7] using the Fisher-Yates exact test

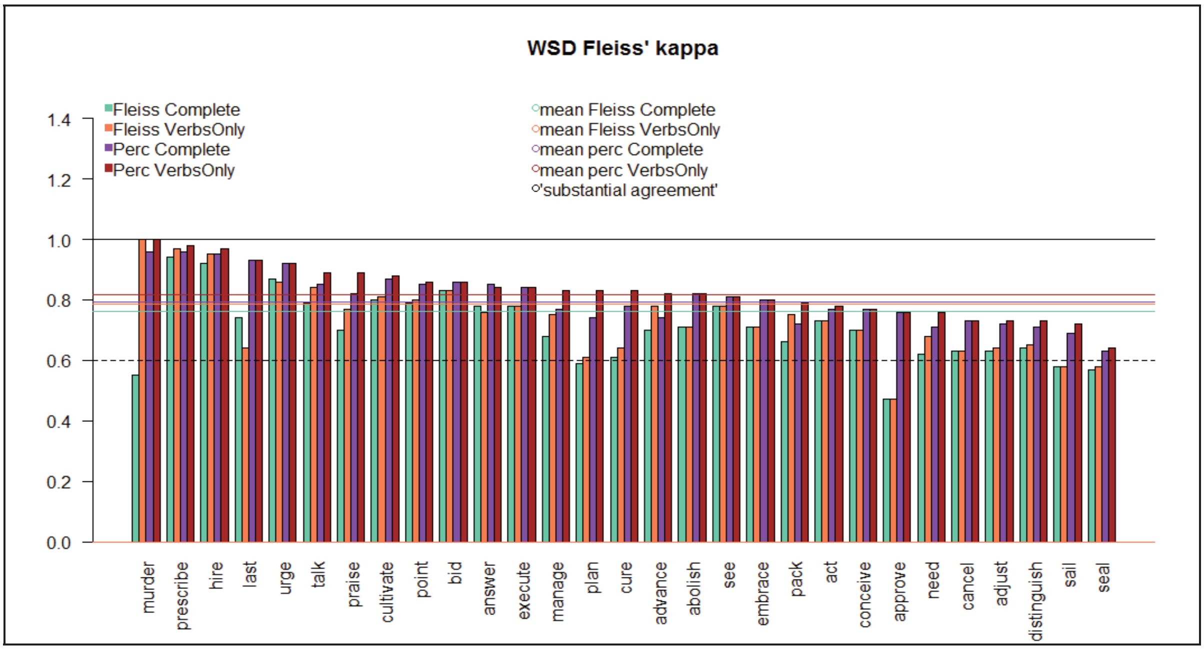
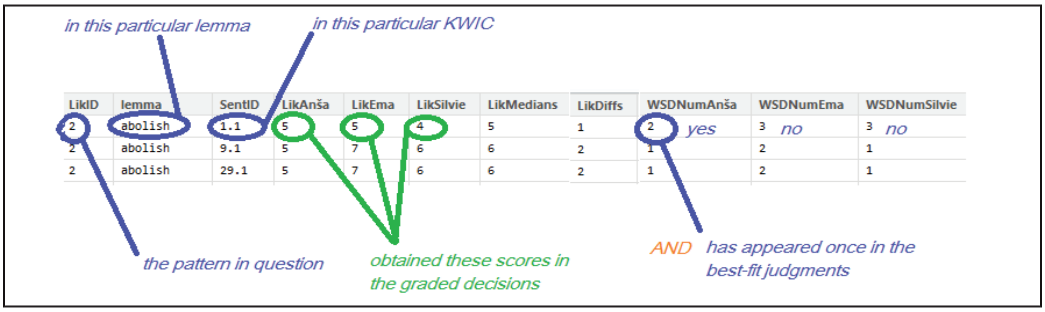


Figure 3: Lemma-wise values of Fleiss kappa and percentual agreements on Complete vs. VerbsOnly, compared to the means for the entire data set and the 0.6 score regarded as the level of "substantial agreement" for Fleiss kappa. The lemmas are presented in descending order according to the percentual agreement on VerbsOnly.


in this particular lemma
in this particular KWIC
yes
no
no
the pattern in question
obtained these scores in the graded decisions
AND has appeared once in the best-fit judgments

LikID	lemma	SentID	LikAnša	LikEma	LikSilvie	LikMedians	LikDiffs	WSDNumAnša	WSDNumEma	WSDNumSilvie
2	abolish	1.1	5	5	4	5	1	2	3	3
2	abolish	9.1	5	7	6	6	2	1	2	1
2	abolish	29.1	5	7	6	6	2	1	2	1

Figure 4: Data for the lexical unit-overlap analysis: Lexical unit ID and how many annotators selected this particular lexical unit in the best-fit lexical unit task (each observation is related to a given KWIC identified by its SentID and lemma). This example captures Lexical unit 2 of the lemma *abolish* in Sentence 1.1. The annotators were generally not considering it an optimum match, but one annotator still judged it as the best-fit lexical unit.

As the next step, we checked for each retrieved KWIC whether there were other competing lexical units, i.e. lexical units with the same median or higher, considering only KWIC with median values 5 and above. We set this threshold because we were only interested in those lexical units that had the potential to become the best fit in the WSD setup, which is reflected by the median. We found 8 such KWICs above the median 5 in WSD3 (of 9), 50 in WSD2 (of 52), and 5 in WSD1 (of 114). This means: The WSD3 group, i.e. cases where all three annotators agreed on the best-fit lexical unit, contained 9 concordances where somebody also gave an equally high rating to other lexical unit(s). Of these, 8 achieved the median 5 and above. Similarly, the WSD2 group contained 52 concordances, in which two of the three annotators agreed on the best-fit lexical unit, but somebody also rated some other lexical unit(s) equally high. Of these, 50 concordances achieved the median 5 and above in the graded decisions. Group WSD1, in which only one annotator selected the given lexical unit as the best-fit unit and any of the annotators rated a different pattern equally high, contained 114 concordances, but only 5 of them had the median 5 and above in the graded decisions. This group contains mostly extreme individual judgments, some of which can even be annotation errors.

We performed a manual analysis of the competing pairs (triples, having occurred rarely, were disassembled into pair combinations). The verbs with most competing lexical units across all three WSD groups turned out to be *approve* (8), *advance* (5)*, act, cancel, cure, embrace, execute*, and *manage* (each 4). The following section discusses the possible reasons for the lexical unit overlap.

6 Discussion

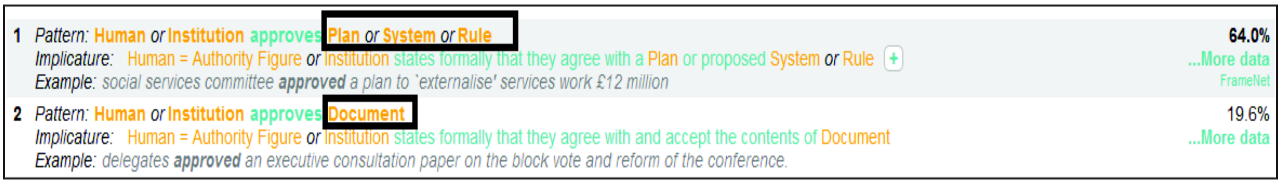
The most frequent competing pairs were *cancel 2 /cancel 6* and *approve 2/approve 1*, with the following feature in common: the KWICs contain objects that happen to match two semantic types at the same time (due to regular polysemy and/or semantic coercion).[8] These semantic types are the only aspect in which the two lexical units differ (cf. Figure 5). Indeed, one lexical unit entails another: there is an event (coerced into a *plan, system,* or *rule*) to be approved, and the approval is typically performed and/or recorded by a document. The word sketch[9] of the objects of *approve* reveals that the approved event and the document always occur intertwined (unlike e.g. in *allow*, which does not imply the use of any document at all). Although the distinction between a document and an event in general is obvious, a large set of noun-verb collocations *abstract from this distinction*, thus making it a weak foundation for splitting.

Regular polysemy and semantic coercion are a common phenomenon, which PDEV largely reflects, in that it e.g. systematically alternates the semantic types Human and Institution within one lexical unit rather than splitting them into two lexical units. Even so, polysemy occurs in 56% of investigated KWICs, and most cases qualify as regular polysemy not captured in the same lexical unit, making undetected regular polysemy the dominating cause of lexical unit overlap. The regular polysemy problem can also be considered a special case of one-directional entailment between a more specific and a more general lexical unit, which regularly obtain identical or very similar scores in the graded decisions. When the granularity difference is evident, entailment does not impair the *manual* best-fit results, as the annotators were instructed to make their decisions as fine-grained as possible. Nevertheless, granularity-based human decisions can be hard to follow with automatic procedures.

Another recurring issue that specifically harms the best-fit results (because they do not allow ties), is (the lack of) abstraction from regular syntactic alternations in patterns. Different realizations of a regular syntactic alternation are often presented by separate lexical units. For instance, the verb *distinguish* has three separate lexical units for different realizations of reciprocity: *distinguish between Entity 1 and Entity 2* vs. *distinguish Entity 1 from Entity 2* vs. *distinguish Entity = Plural*. To name another example: lexical units of verbs of speaking are sometimes split due to the structural (and hardly semantically relevant) difference between a quote and a *that*-clause, which are hard to tell apart whenever *that* is omitted or quotes are missing *(urge 3,4).* The same applies to the difference between adverbials realized as adverbs and those realized as prepositional groups (*act 1,2*). Prepositions and multiword preposition-like expressions pose a challenge, too, when each is recorded in a separate lexical unit, since multiword expressions display substantial variability and are impossible to list exhaustively. So *act in the interest of*, which is not recorded (although evidently a normal use), can be randomly assigned to either *act on behalf of (act 5)* or *act for (act 4),* or the annotator has to fall back to a more general lexical unit *act 1*. In general, as soon as a KWIC contains a regular syntactic realization not explicitly listed in the inventory of lexical units or (however slightly) deviant from the listed ones, all the syntactic realizations listed as lexical units become equally suitable best-fit candidates, as our data also reflect (most cases in WSD2).

[8] Cf. e.g. the annotation work of Martínez Alonso, Sandford Pedersen, and Bel 2013 and Pustejovsky et al. 2009.

[9] (Kilgarriff et al. 2004)



1 *Pattern:* **Human** *or* **Institution approves Plan** *or* **System** *or* **Rule** **64.0%**
Implicature: Human = Authority Figure *or* Institution states formally that they agree with a Plan or proposed System *or* Rule + ...More data
Example: *social services committee* ***approved*** *a plan to `externalise' services work £12 million* FrameNet

2 *Pattern:* **Human** *or* **Institution approves Document** 19.6%
Implicature: Human = Authority Figure *or* Institution states formally that they agree with and accept the contents of Document ...More data
Example: *delegates* ***approved*** *an executive consultation paper on the block vote and reform of the conference.*

Figure 5: Lexical units 1 and 2 of the verb *approve* – an example of systematic coercion.

Yet another recurrent cause of lexical unit overlap pose lexical units whose patterns are defined by the optional presence of different adverbials or the optional presence vs. absence of an adverbial (e.g. *sail 1,3*), with implicatures differing just with respect to the adverbial differences. Whenever no adverbial or both adverbials are present, the best-fit decision inevitably becomes random.

7 Conclusion

We have performed a quantitative analysis of 11,350 graded human decisions regarding the 50-KWIC batches of 29 verbs captured by completed PDEV entries, followed by a qualitative examination of lexical unit overlap causes on all cases where two or more lexical units had identical scores high enough to be eligible for the best-fit lexical unit in the classical Word Sense Disambiguation Setup. The data are publicly available at http://hdl.handle.net/11234/1-1585.

Our results touch two topic areas:

- annotation and its usability in statistical machine learning,
- causes of semantic overlap in PDEV lexical units.

Concerning the annotation results, we applied the graded-decision approach to a lexicon based on usage patterns. Our interannotator agreement/correlation results show that *this data is no less consistent than graded-decision judgments produced with a traditional dictionary*.

As for the latter point: we are well aware of the fact that the aforementioned observations have different implications depending on the purpose to which PDEV is used. In the Word Sense Disambiguation context, the ideal dictionary would be one that gets the user/computer to select exactly one lexical unit, and in this respect, our observations could represent revision suggestions. Human users of PDEV could profit from a sensibly reduced number of lexical units, e.g. by having several syntactically alternating patterns share one implicature.

Although the theoretical background of PDEV rejects word senses and replaces them with usage patterns, PDEV entries, as a more radical version of the influential Collins Cobuild conception (Birmingham 1996), have been widely adopted in the modern lexicographical standards (cf. e.g. (Rundell 2002) and (Oxford Learner's Dictionaries 2015)). If there used to be a gap between the collocation-oriented Sinclairian approach and the “traditional” sense defining lexicographical approach in the 1990s when the first theoretical foundations of PDEV were laid, it seems to be vanishing. Therefore we hope that our findings based on PDEV will also be relevant for other modern learners' dictionaries as well as for anyone willing to launch a semantic annotation on verbs.

8 References

Baisa, V., Cinková, S., Krejčová, E., & Vernerová, A. (2016). VPS-GradeUp: Graded Decisions on Usage Patterns. In *LREC 2016 Proceedings*. Portorož, Slovenia.

Collins Cobuild Grammar Patterns: Helping Learners with Real English. HarperCollins Publishers Limited. University of Birmingham (2006).

Cinková, S., & Hanks, P. (2010). Validation of Corpus Pattern Analysis - Assigning pattern numbers to random verb samples. Retrieved from http://nlp.fi.muni.cz/projekty/cpa/CPA_valiman.pdf, http://ufal.mff.cuni.cz/spr/data/publications/annotation_manual.pdf

Cinková, S., Holub, M., Rambousek, A., & Smejkalová, L. (2012). A database of semantic clusters of verb usages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)* (pp. 3176–3183). Istanbul, Turkey: European Language Resources Association.

El Maarouf, I. (2013). Methodological Aspects of Corpus Pattern Analysis. *ICAME Journal*, *37*, 119–148.

Erk, K., McCarthy, D., & Gaylord, N. (2009). 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* (pp. 10–18). Suntec, Singapore: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P/P09/P09-1002

Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring Word Meaning in Context. *Computational Linguistics*, *39*(3), 511–554.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press.

Krishnamurthy, R., & Nicholls, D. (2000). Peeling an Onion: The Lexicographer’s Experience of Manual Sense Tagging. *Computers and the Humanities*, *34*, 85–97.

Landes, Leacock, C., Tengi, R. I., & et al. (n.d.). SemCor. Princeton University. Retrieved from http://multisemcor.fbk.eu/semcor.php, http://www.cse.unt.edu/~rada/downloads.html#semcor

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993). A semantic concordance. In *Proceedings of ARPA Workshop on Human Language Technology*.

Oxford Learner's Dictionaries. (n.d.). Retrieved from http://www.oxfordlearnersdictionaries.com/

Palmer, M., Dan Gildea, & Paul Kingsbury. (2005). The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics Journal*, *31*(1).

Rumshisky, A., Verhagen, M., & Moszkowicz, J. L. (2009). The Holy Grail of Sense Definition: Creating a Sense-Disambiguated Corpus from Scratch. In *Fifth International Workshop on Generative Approaches to the Lexicon (GL 2009)*. Pisa, Italy.

Rundell, M. (2002). *Macmillan English Dictionary for advanced learners*. Macmillan Education.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., & Scheffczyk, J. (2010). *FrameNet II: Extended Theory and Practice*. ICSI, University of Berkeley. Retrieved from https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf

Acknowledgements

This project was supported by the Czech Science Foundation grant GA-15-20031S and the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic.

This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth, and Sports of the Czech Republic (project LM2015071). For most implementation we used R (R Core Team, 2015).