A database of semantic clusters of verb usages

Silvie Cinková*, Martin Holub*, Adam Rambousek†, Lenka Smejkalová*
*Charles University in Prague, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics
Malostranské náměstí 25, Praha 1, Czech Republic
†Masaryk University, Centre for Natural Language Processing, Faculty of Informatics
Botanická 554/68a, 602 00 Brno, Czech Republic

E-mail: {cinkova,holub,smejkalova}@ufal.mff.cuni.cz, xrambous@fi.muni.cz

Abstract

We are presenting **VPS-30-En,** a small lexical resource that contains the following 30 English verbs: *access, ally, arrive, breathe, claim, cool, crush, cry, deny, enlarge, enlist, forge, furnish, hail, halt, part, plough, plug, pour, say, smash, smell, steer, submit, swell, tell, throw, trouble, wake* and *yield*. We have created and have been using VPS-30-En to explore the interannotator agreement potential of the Corpus Pattern Analysis. VPS-30-En is a small snapshot of the Pattern Dictionary of English Verbs (Hanks and Pustejovsky, 2005), which we revised (both the entries and the annotated concordances) and enhanced with additional annotations. It is freely available at **http://ufal.mff.cuni.cz/spr.** In this paper, we compare the annotation scheme of VPS-30-En with the original PDEV. We also describe the adjustments we have made and their motivation, as well as the most pervasive causes of interannotator disagreements.

Keywords: Corpus Pattern Analysis, clustering, lexical semantics

1. Introduction

We are presenting **VPS-30-En** (Verb Pattern Sample, 30 English verbs) – a pilot lexical resource of 30 English lexical verb entries enriched with semantically annotated corpus samples (henceforth **VPS**). VPS is publicly available at **http://ufal.mff.cuni.cz/spr.** It describes regular contextual patterns of use of the selected verbs in the BNC (The British National Corpus, 2007). VPS has arisen as a result of a previous cooperation with Patrick Hanks, drawing on his Pattern Dictionary of English verbs – PDEV (Hanks and Pustejovsky, 2005). VPS contains the following verbs: *access, ally, arrive, breathe, claim, cool, crush, cry, deny, enlarge, enlist, forge, furnish, hail, halt, part, plough, plug, pour, say, smash, smell, steer, submit, swell, tell, throw, trouble, wake* and *yield*.

PDEV is publicly available at http://deb.fi.muni.cz/pdev/ (Horák et al., 2008). Hanks' approach to lexical description is innovative and intuitively very appealing in terms of NLP. Like e.g. (León et al., 2009, Rumshisky and Pustejovsky, 2006), we seek to explore its potential for automatic lexical disambiguation. Unfortunately, PDEV has still very limited corpus coverage (approx. 10 % of verb occurrences) and, to the best of our knowledge, no large-scale expansion is underway. Taking this into account, we seek to address the following issues in the long term:

1) Can an on-the-fly automatic pattern creation be learned, which would be tailored to each individual data set or application?
2) If 1 turns out to be feasible, will this improve the performance of any practical NLP task?

These questions are admittedly far too complex. In the following sections, we are making first steps towards tackling what we believe are the prerequisites:

1) Can a reasonable interannotator agreement (IAA) be achieved in assigning the pattern numbers?
2) Can we identify any types of disagreements that are not caused by bad pattern design?
3) Can automatic pattern assignment be trained with reasonable performance?

To start with, we have put down annotation guidelines, cleaned up the original PDEV data to facilitate the interannotator agreement, run several annotation rounds, as well as recorded and analyzed the behavior of the annotators. This paper is also a report on the experiments and the resulting observations.

2. The original resource - PDEV

2.1 Availability and coverage

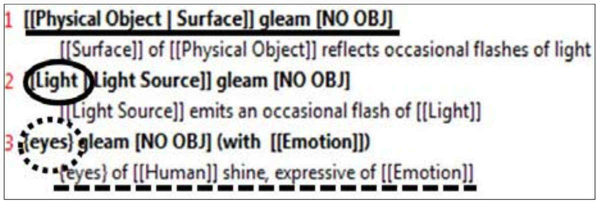
PDEV is publicly available at http://deb.fi.muni.cz/pdev/ (Horák et al., 2008). Almost all entries were created and edited by P. Hanks. The entries were created on the basis of BNC50, which is a subset of BNC comprising around 50 million tokens, cleared of spoken documents and some older fiction (Hanks, personal communication, 2008–2010). All statistics in this paper also refer to BNC50.

BNC50 contains 99 high-frequency verb types (>9,999 occurrences), 573 verb types with 10,000 > f > 999, 576 verb types with 1,000 > f > 349 and 4533 entries with even lower frequency.

PDEV contains 694 lexical verbs with the status "complete", which altogether comprise 2663 patterns (March 13, 2012). There are 9 completed verbs with frequency over 10,000: *say, need, call, tell, lead, claim, accept, argue* and *explain*. The frequency ranking in BNC50 closely corresponds to the one observed in the entire BNC. Our snapshot of PDEV from the early 2011 misses 24 new verbs present in the current web version. Nevertheless, of these new verbs only *argue* has more than 10,000 occurrences (11,362). The verb *cry* has 1,200 and the others do not even reach 1,000 occurrences. Most completed PDEV verbs (428) do not reach 100 and 131 other completed verbs do not reach 350 occurrences. 63 verbs have frequency 350 <= f <= 999 and other 63 have frequency 1000 <= f <= 9999.

2.2 PDEV entry structure

The PDEV scheme has undergone slight changes since 2011. We describe the current scheme. Each lexical entry consists of **categories** (numbered in Fig. 1). Each category consists of a **pattern** (marked with a full line) and an **implicature** (dotted line). The pattern represents the morphological, syntactic and lexical characteristics of the verb used in a certain context. Characteristic contexts activate different aspects of the **meaning potential** of the verb (for details see Hanks and Pustejovsky, 2005). The meaning is represented by the implicature. The pattern takes the form of a predication. The pattern-defining verb complements, which we call **slots**[1] are represented by **semantic types** (full oval) or **lexical sets** (dotted oval). A lexical set is a list of characteristic collocates in one slot (e.g. the object of *cool:* ***atmosphere, tempers***). Semantic types are items in Hanks' ontology (part of PDEV). The ontology development is still in progress, but the number of semantic types remains stable around 200 labels. The implicature is a paraphrase or reformulation of the proposition rendered by the pattern. Whenever possible, it mirrors the slots contained in the pattern.



1 [[Physical Object | Surface]] gleam [NO OBJ]
[[Surface]] of [[Physical Object]] reflects occasional flashes of light
2 [[Light | Light Source]] gleam [NO OBJ]
[[Light Source]] emits an occasional flash of [[Light]]
3 {eyes} gleam [NO OBJ] (with [[Emotion]])
{eyes} of [[Human]] shine, expressive of [[Emotion]]

Fig. 1: A PDEV entry with 3 categories

2.3 Slot features

Each category contains a global description part that applies to the entire category. It contains information about domain and register, as well as it indicates that the pattern is an idiom or a phrasal verb. In addition, essential features of each particular slot are recorded in a slot description form (Fig. 4). The scheme distinguishes the following slot types: **subject, object, clausal object, indirect object, complement, clausal** and **adverbial**. The slots "object" and "adverbial" have a tick box for indicating that the pattern is determined by the absence of these slots, and another one for marking their optionality. The slots "subject", "object", "indirect object" and "adverbial" are meant for nouns. The form contains fields for the semantic type and role, lexical set, quantifier/determiner and attribute. In addition, "adverbial" contains fields for preposition or particle and adverbs. "Clausal object" and "clausal" have tick boxes for the following clause types: to+infinitive, -ing, that-clause, wh-clause and quote. The additional item "semantics" contains the same fields as the slots rendered by nouns. The "complement" slot can be either classified as subject complement or as object complement.

2.4 Annotated concordance samples

Each verb entry is associated with a lexicographer -annotated concordance sample, which we call **reference sample**. In verbs less frequent than 250 occurrences it usually takes all occurrences. In more frequent verbs it is typically 250 occurrences. The sample is larger in very complex verbs, e.g. *throw*. Besides, there are semi-automatically annotated concordances. They arose by manual but global assignment of a common pattern number to concordances sorted by typical collocates in certain syntactic positions by the Sketch Engine (Kilgarriff et al., 2004). These contain some noise.

2.5 Concordance classification

There are four types of labels in the samples: plain number, "number .e", "u" and "x". Meta uses of verbs, verbs as parts of named entities and non-verbs mistakenly tagged as verbs are marked with "x". Sensible verb uses that match no existing patterns are marked with "u". Concordances that match a pattern without reservation (prototypical uses) get the number of that pattern. The mark ".e" (exploitation) in combination with a pattern number is used in concordances that intuitively belong to an existing pattern but deviate in syntax, or the verb use is figurative or ironic, or the arguments do not match the semantic types.

3 VPS deviations from the PDEV scheme

3.1 Motivation

While the goal of the original PDEV is to become a large-scale lexical resource, VPS has been built solely as a sample of the cleanest CPA-based data available for experiments on whether or not human annotators can agree on the CPA-based semantic clustering. We intend to use this sample as a gold-standard data set to support automatic semantic clustering of verb concordances. This implies a different focus. Much more than PDEV, we make sure to keep the data in line with the entries after each entry revision, and the entry revisions we make are supported by annotation experiments.

The initial experiments performed during 2009 and 2010 revealed the need of a PDEV snapshot serving as a sandbox that we could adjust to our needs without destroying the original data. Before parting from the original PDEV, we created a tentative annotation scheme in cooperation with P. Hanks and put down explicit annotation guidelines for pattern assignment. We have been sticking to the original approach as much as we were able to. The following sections mention deviations that evolved nevertheless.

3.2 Minor alterations

Since we have been heading for automatic pattern assignment in an automatically parsed corpus, we have been seeking to model the syntactic dependencies in each pattern to facilitate the mapping of patterns on parsed sentences. We have therefore expanded the annotation scheme to model the syntactic features of the described verb in more detail. For instance, we have introduced fields that say that the described verb is in coordination with another verb or that it is governed by another verb, that it is typically introduced by a subordinating

[1] Since we would like to preserve the term *complement* for one particular type of verb argument, we will henceforth refer to the pattern-defining verb complements as (valency) **slots**.

conjunction (e.g. in the phrase *if you please*), that negation is typical for that given pattern or that the described verb is in this case a light verb.

We distinguish several types of noun modifiers in the inner structure of the nodes (e.g. possessive pronoun or genitive, adjective or prepositional phrase, quantifier or determiner and pre-determiner would each be considered a different modifier type). We indicate slots that are reciprocal and can thus undergo various syntactic alternations associated with reciprocity. We do not make any difference between subject and object complement, but we observe whether it is a noun or an adjective and the presence of prepositions (typically *as*). The original PDEV would probably regard a complement with *as* as an adverbial. In adverbials, we distinguish among particles, adverbs, prepositional phrases and clauses. In indirect object (which is mostly the "dative" object or beneficiary) we have an extra field for preposition. While the original PDEV scheme only regards objects with the *to*-alternation as indirect objects, we also include *for*. There is an additional field for other prepositions.

We have certainly not tackled the syntactic issues quite consistently. For instance, prepositional phrases as obligatory arguments remain a classification problem, like in the original PDEV. A prepositional object as the only object like *rely on* or *indulge in* will be classified as adverbial just because of its position at the end of the sentence and the practical assumption that in abstract events like *take something into account* it is hard to decide whether *into account* is an indirect object or an adverbial and most automatic parsers would classify it as an adverbial anyway. On the other hand, we stuck to verb complement rendered by a noun as a slot class, although parsers are likely to confuse them with indirect objects, simply because they are intuitively much easier to tell apart for humans than telling the prepositional objects from adverbials. Our scheme is admittedly somewhat clumsy in that it displays the noun-adjective option as two separate verb complements present at the same time.

We do not consider the alterations of the slot description an essential deviation from the original approach. The detailed specification is available on our web page.

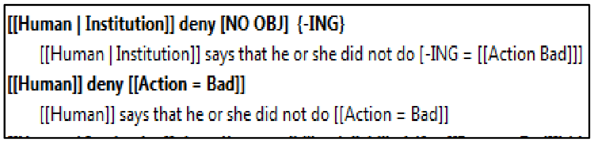
3.3 Major alterations

There are, nevertheless, four deviations from the original approach that we consider quite essential and worth a more detailed description:

1. merging of clausal slots with the noun slots
2. emphasis on the semantic distinctiveness of implicatures
3. separate patterns for participial verb forms
4. more nuanced exploitation types

3.3.1 Merging of clausal slots

In the pre-VPS PDEV scheme, all clausal verb arguments were regarded as one syntactic element in its own right. (The "clausal object" label is more recent.) For instance, a that-clause or quote attached to *say* did not count as an object. The pattern of *deny* explicitly indicates the absence of direct object and the presence of a gerundial clause. Hence it takes two patterns to describe that a verb often (but not always) has an argument rendered by a clause. Their implicatures only differ in the description of that particular argument (Fig. 2). The VPS scheme, on the other hand, allows saying that a verb has a direct object that has a given semantic type when rendered by a noun and a given form, when rendered by a clause (Fig. 3 and Fig. 4 – see tick box alternatives in the form). The reasons for classifying arguments according to their function were mainly that we are used to think in terms of dependency grammar, which has no problem with clauses as arguments, and, secondly, that we wanted the distinctions between categories to be purely semantic. We accept different categories e.g. for different diatheses, since they at least shift the perspective of the event (e.g. *sun radiates heat* vs. *heat radiates from the sun*), but we do not see any semantic difference between an event noun, a nominalization and a verbal clause, in particular when they trigger the same implicature. Therefore our scheme does not represent them in separate categories.



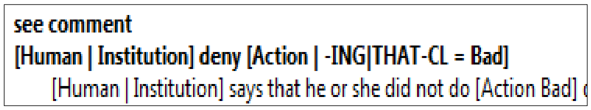
[[Human | Institution]] deny [NO OBJ] {-ING}

[[Human | Institution]] says that he or she did not do [-ING = [[Action Bad]]]

[[Human]] deny [[Action = Bad]]

[[Human]] says that he or she did not do [[Action = Bad]]

Fig. 2: Separate indication of arguments shaped as a clause in the original PDEV

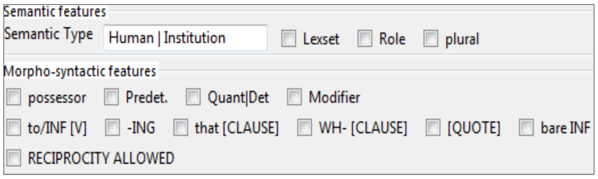


see comment

[Human | Institution] deny [Action | -ING|THAT-CL = Bad]

[Human | Institution] says that he or she did not do [Action Bad]

Fig. 3: Merging of noun and clausal arguments in VPS



Semantic features

Semantic Type: Human | Institution □ Lexset □ Role □ plural

Morpho-syntactic features

□ possessor □ Predet. □ Quant|Det □ Modifier

□ to/INF [V] □ -ING □ that [CLAUSE] □ WH- [CLAUSE] □ [QUOTE] □ bare INF

□ RECIPROCITY ALLOWED

Fig. 4: a slot form in VPS-30-En

3.3.2 Semantic distinctiveness of implicatures

The current parsers would enable us to automatically cluster concordances of a verb according to their syntactic similarity. What we, nevertheless, still lack resources for is learning how to cluster concordances according to their **semantic** similarity. A lexical resource could be helpful that would consequently cluster concordances based on the semantic similarity (i.e. a common implicature) and describe the morphosyntactic and lexical features that – in the human intuition – make a bunch of concordances trigger the same implicature.

The original PDEV has been evolving through a long period of time, and the focus seems to have been shifting between the emphasis on syntax (entries where several patterns with minor syntactic differences have identical or very similar implicatures) and on semantics (implicatures are clearly different even when the patterns look similar). Having noticed this, we seek to prioritize the latter approach. This has implications. On one hand, we systematically merge patterns that have identical implicatures into one common pattern, but, on the other hand, the 30-verb sample suggests that we generally tend to split patterns. The average number of patterns is 13.6 in VPS compared to 9.23 on the same set of verbs in PDEV (March 2012). The number of patterns is only higher in PDEV in case of *submit* and *tell*. In case of *throw,* where the difference is truly striking, the explanation is that the annotated concordance sample (on which the entry was based) was larger in VPS than in PDEV and simply more patterns emerged in the concordances (72 vs. 46). Compared to PDEV, we were not equally strict about the criterion of frequency. A recognized idiom got its own pattern in VPS even if observed only once. Also, we consulted the COCA corpus (Davies, 2008) to check frequency in a really large corpus, whenever a pattern seemed to be frequent, even though it happened to occur only sporadically in the samples. Whenever COCA revealed that the pattern occurred in similar and unmarked contexts and was not rare, it was included into the entry. Eventually, we were seeking to formulate the implicatures as common-language paraphrases rather than in very abstract terms, though we did not always succeed. Here, collocability of the paraphrasing verbs has come into play and can have required more splits to make the participants of the events so semantically homogeneous as to find an acceptable common collocate for them. The homogeneity of event participants is an ex-post observation rather than a goal set from the beginning.

3.3.3 Participial patterns

Participles represent a transition between verbs and adjectives or between verbs and nouns. When regarded as verbs, they are inherently unspecific as to the number of event participants and causality. The association with the transitive pattern implies a passive verb form, whereas the association with an intransitive pattern implies pseudo passive. Since the difference between a pseudo passive verb form and an adjective (i.e. "x") is often blurred, we decided to draw a line between a passive form of a verb and any other participial form all the way down to a clearly manifested adjective or noun (e.g. participles as noun modifiers). All verbs that often occur in participle forms got participial patterns, which narrowed the space for disagreement from three options (passive, pseudo passive, non-verb) to two (passive, participle). In one exceptional case, a non-participle noun (*steer*) was encoded in a separate pattern, since it was extremely frequent and kept confusing the annotators (*steer prices*). The participial patterns appear only to occur separately in PDEV when the matching concordances do not occur in other forms (e.g. *seek 8* in Hanks and Pustejovsky, 2005, but not in PDEV). Neither did PDEV use any explicit criterion for classifying participial verb forms, as the reference samples showed.

3.3.4 Exploitation types

Marking exploitations has two purposes in PDEV as well as in VPS. On one hand it identifies atypical uses; on the other hand, too many exploitation cases or "u"-cases suggest the need of a revision of the entry. Too many uses marked as "u" typically mean that a pattern is missing or that several patterns would match to equal extent. A use can be classified as an exploitation of a pattern (i.e. ".e" rather than "u") for several different reasons. This markup is used both by the lexicographer and by the annotators, but there is a difference: while creating the patterns and annotating the reference sample, the lexicographer seeks to end up with as few exploitations of the same kind in the same category as possible, which takes a number of iterations between redefining the patterns and re-annotating the data. The annotators do not get them before the lexicographer believes to have achieved the best result possible. The annotators, who get additional 50 random concordances for annotation according to the patterns and the reference sample, act as proofreaders. They are trained to read the patterns closely and indicate any mismatches between the patterns and the data.

In cooperation with P. Hanks, we identified the following exploitation types, which are indicated in the annotation:

- figurative use, idiom, creative metaphor (.f);
- different auxiliary words than defined by the pattern (*enlarge on* vs. *upon*), a diathesis *(spray the wall with paint/spray paint on the wall, the bibliography is* <u>*being accessed*</u> [causative] *to the Cambridge computer by staff at the Cambridge Group for the History of Population)* or ellipsis indicating a generic participant *(we punish too much and imprison too much)* (.s);
- an event participant (content word) atypical of the sematic type defined by the pattern *(ride a caterpillar)*, without any metaphorical shift (.a);
- metonymy or coercion (*drink [Beverage]* vs. *He drank a cup.*) (.c).

The ".a" and ".s" exploitations indicate most often that the pattern is defined too narrowly rather than that there are many atypical uses. For instance, the semantic type of a position is often defined as "Human", but the data reveal that institutions are equally represented. The association between humans and institutions is so natural, that the lexicographer easily forgets to list "Institution" as a separate semantic type, but the annotators usually catch it. In contrast, many concordances marked with ".f" do not necessarily imply that a new pattern should be created, since the metaphorical shifts can be quite heterogeneous and impossible to encompass by one single implicature.

The type ".c" is regarded as a subset of ".a", since the annotators turned out to have a problem telling it apart from ".a".

We have also introduced a hierarchy of matches. In a nutshell: the annotator compares the given concordance to all implicatures. When only one implicature matches, the annotator observes the pattern of the candidate category. When also the pattern matches in the semantic types or lexical sets and the auxiliary words match as well, the sheer pattern number is assigned to that concordance. When there are minor deviations but the syntax still remains similar (or it is a regular diathesis), the mismatch is classified with ".s". When the syntax is very dissimilar, the concordance is to be marked as "u".

When a usage is clearly a creative exploitation of a given category, that category is assigned with ".f". ".f" certainly implies that even the event participants can be atypical.

When it is not a creative metaphorical use, the category with the better matching implicature is to be preferred and the deviations in the semantic types/lexical sets or in the syntax should be indicated. When both lexical (".a", ".c") and syntactic (".s") deviations occur, ".a" is to be preferred.

Special rules had to be introduced for evident phrasemes and idioms. The categories that describe idioms are marked as idioms, also in the original PDEV. We distinguish between "phraseological adverbials" and full-fledged phrasemes. When the construction matches the implicature of a regular (non-idiom) pattern as well as its syntax (having the same configuration of subject/object(s)) but it contains an additional phraselogical element, typically an adverbial, it is not regarded as exploitation. For instance: *frighten sb to death, out of one's skin* is still regarded as a normal use of *frighten*. Under a full-fledged phraseme we understand a stable combination of content words that constitute a new meaning altogether that cannot be derived from the single elements, e.g. *Human 1 hits Human 2 below the belt* (behaves in an unfair way). When no idiom category matches in the entry, the annotator is encouraged to pick a category that matches the **literal meaning** of the idiom as well as the syntax and semantic types/lexical sets and mark it with ".f". When an implicature of another idiom category matches well, that category is assigned with ".s". With this set of exploitation preferences, the participle patterns and mainly with the emphasis on the implicature part of the category, we seek to decrease unnecessary interannotator disagreements without eliminating the individual judgment.

4 VPS annotation

4.1 Verb selection

We have selected the following verbs: *access, ally, arrive, breathe, claim, cool, crush, cry, deny, enlarge, enlist, forge, furnish, hail, halt, part, plough, plug, pour, say, smash, smell, steer, submit, swell, tell, throw, trouble, wake* and *yield*.

All have higher frequency than 300. *Say*, *tell*, and *claim* have over 10,000 occurrences in BNC50. *Arrive, claim, deny, throw, submit, yield* and *cry* have between 1,000 and 10,000 occurrences. *Throw* has only 3,710, but is known to be a complex verb, both semantically and syntactically, with limited potential to act as a light verb (*throw a punch, throw a wink*).

All verbs had the "complete" status in PDEV. The minimum frequency threshold was set to 300, later changed to 350. Considering the number of patterns, we wanted to explore a few with low pattern numbers (minimum 3), but we assumed that those verbs of all frequency ranges would be interesting that have a higher number of patterns. At the same time, we preferred verbs that we had not met in a preliminary 2009 annotation experiment (which excluded some frequent complete verbs as *need, call, lead, accept, argue* and *explain*, but on the other hand we kept *claim*). Since none of us has a lexicographical experience comparable to P. Hanks, we decided, as a precaution, to start with verbs that are somewhat complex (most verbs have 10-20 patterns) but just so frequent that we were able to go through all their occurrences manually (300 - 1,000 occurrences). Later, having gained some practice both as lexicographers and as annotators, we proceeded to the more frequent ones.

4.2 Entry compilation

The lexicographer revises the reference sample annotated by P. Hanks according to the original PDEV entries. All exploitations marked as ".e" are classified according to the new scheme. When appropriate, also the entry is revised. Typically, semantic types and items in lexical sets are added. Sometimes, categories are merged and split or new patterns are added until the lexicographer believes to have reached the optimum match between the entry and the reference sample. COCA and printed dictionaries are consulted during the work.

4.3 Random sample annotation

The annotators get a random 50-concordance sample along with the lexicographer-annotated reference sample and the entry. They match each random concordance to the categories according to the similarity of implicatures, the similarity of the patterns and, not least, according to the overall similarity of the concordance to the concordance clusters associated with the respective categories.

4.4 Disagreement analysis and adjudication

After each annotation round, IAA is measured and disagreements are manually analyzed. The disagreement analysis is supported by confusion matrices computed for each annotator pair. Provided the annotation of the random sample reached a satisfactory IAA, the disagreements are manually adjudicated by the lexicographer in a spreadsheet table: the lexicographer highlights evident annotation errors, lists all acceptable values and "one best choice" in a separate column to each concordance. The "one best" annotation is typed back into the user interface as part of the **gold standard data set**.

4.5 Sample revisions

The IAA from the multiple annotations along with the confusion matrices gives a hint on whether the entry is designed adequately for the given data. When the entry does not appear to need any further revisions, the reference sample is not revised. The "one best" annotation is added into the interface. Together they form the gold standard data set. The most common case is, however, that the first annotation round triggers revisions of the entry. When the entry is revised, the reference sample is revised too to remain in line with the entry. No adjudication is performed on the random sample. Instead, the lexicographer annotates it in the same way as the reference sample, considering the opinions of the annotators. Now, the 50-concordance random sample had got the same status as the reference sample. Actually, it should become part of the reference sample, but the design of the interface does not allow it. As a make-do solution, we call the original reference sample "original sample" and the results of all non-final annotation rounds "trial samples". The final sample (which triggered no entry revision) is called "adjudicated sample".

The original sample along with the trial samples and the adjudicated sample constitutes the final reference sample. The entire reference sample is revised by the lexicographer after every alteration of the entry.

4.6 Gold standard data set

The gold standard data set consists of the reference sample and of the adjudication table for the adjudicated sample. It contains minimally 300 concordances. As a rule, it consists of 350 concordances (a 250-concordance original sample, one trial and one adjudicated sample).

4.7 Infrastructure

Entries have been compiled and the samples annotated in a web-based user interface developed and maintained primarily for PDEV by the NLP Lab at the Masaryk University in Brno (Horák et al., 2008).

5 Interannotator agreement analysis

5.1 IAA results

We divided the selection into two parts: 17 "warm-up" verbs and 13 "for real". The annotators worked independently during all rounds. In the warm-up set, we discussed the disagreements with the annotators after the first round. The first annotation round with the warm-up set was taken up with 3 annotators (one of them the lexicographer), the original scheme of PDEV and the annotation manual. Fleiss' kappa (Artstein and Poesio, 2008) reached 0.6 only 6 times. After entry revision (usually one round), the IAA improved in 15 cases (13 over 0.6) but dropped in *forge* and *wake* (Fig. 5), dragging *forge* down under 0.6 from 0.685. The number of patterns increased in all entries compared to PDEV.

In the for-real set of verbs, the entries and the references sample were revised before they were given to the annotators for the first time. IAA reached well over 0.6 except in *throw* and *halt* (Fig. 6).

5.2 Lessons learned from the warm-up set

Less frequent verbs constituted the warm-up set in the hope that they would be easier to handle – for the lexicographer as well as for the annotators. The initial IAA was nevertheless disappointingly low. The disagreement analysis revealed numerous annotators' errors, such as regarding a transitive sentence in passive as an intransitive sentence and confusing adjective or pseudo passive with a regular passive. Also, the annotators were rather insensitive to shortcomings in the pattern definitions. For instance, a pattern prescribed an adverbial with the preposition *from*, but the concordance said *out of*. Otherwise it matched the category well. A skilled annotator would have marked her concordance as a syntactic exploitation, giving a hint to the lexicographer that, if this case is frequent, it should be included in the pattern. Sometimes it even happened that an annotator kept marking concordances with a given pattern number, meaning another one. These were insufficiencies that we were fighting at the beginning. They have been gradually decreasing with the growing experience.

There was, however, a more substantial cause of disagreement: when two (or more) categories came into consideration, there were individual preferences of pattern versus implicature. Both annotators would agree that the concordance was an exploitation, but one would say it is a syntactic/semantic type exploitation of Category A, whereas the other would call it a semantic shift (".f") in Category B. The original PDEV reference samples appeared to be inconsistent in this preference, too, and, to the best of our knowledge, there was no explicit rule for this case. As this problem would have compromised the IAA systematically, we decided to put more weight on implicature. We issued a guideline saying, when in doubt, prefer "A.s" or "A.a" to "B.f", and if it does not seem correct, call it "u".

This decision had implications for the entry design as well. The implicatures should be semantically distinct, small as the semantic difference could be, and they did neither have to be mutually exclusive. The next annotation round (with a different concordance sample) was more successful. The average improvement was 0.1, the highest was 0.286. We consider three factors to have affected the score: fewer evident misjudgments, emphasis on implicature and entry revisions (including a new scheme).

5.3 The for-real set

The for-real set contained 13 verbs. The entries and the reference samples were revised according to the experience with the warm-up set before the first annotation round. *Cool*, *deny* and *yield* had two annotation rounds – *cool:* up from 0.725 to 0.843, *deny*: up from 0.58 to 0.651, *yield:* up from 0.573 to 0.716. The original IAA in *cool* was satisfactory, but the disagreement analysis suggested potential for improvement by a minor entry revision, which the second annotation confirmed. After revisions based on the disagreement analysis, *cool* and *yield* overcame the 0.6 threshold.

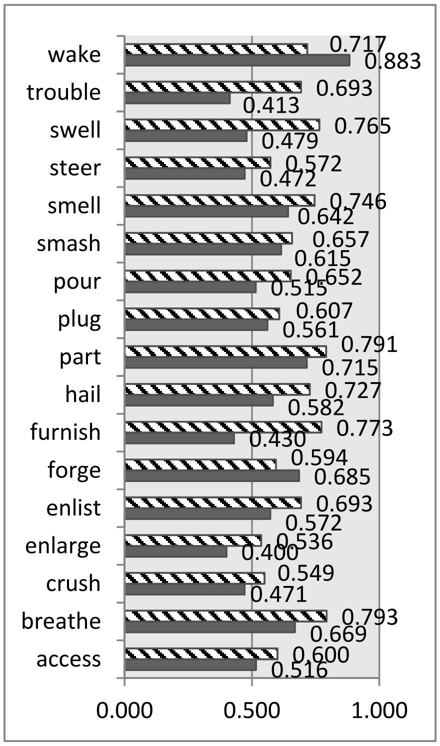


Fig. 5: The warm-up verb set. IAA in the 1st round (black) vs. IAA in the 2nd round (striped)

5.4 IAA versus intuition

To measure IAA, we use both percentage and Fleiss' kappa. Compared to the percentage count, Fleiss' kappa gives lower values. Since both measures produce values between 0 and 1, we can immediately observe that their differences are not proportional (Fig. 7).

For instance, *halt*, which has only three patterns and one is picked more frequently than the others, counts as an easy verb, since it has low **perplexity**. Fleiss' kappa penalizes disagreements in lowly perplex verbs harder than in highly perplex verbs. Hence an 80% agreement gave only 0.54 Fleiss' kappa in *halt*, while an 81.3% agreement in *part* gave a kappa of 0.791. On the other hand, both measures almost corresponded in *throw*. *Throw* is a verb with 72 patterns, of which only about 30 were quite evenly distributed in the annotated data (such that the perplexity might have grown much higher in a larger data and might have made the kappa measure more tolerant).

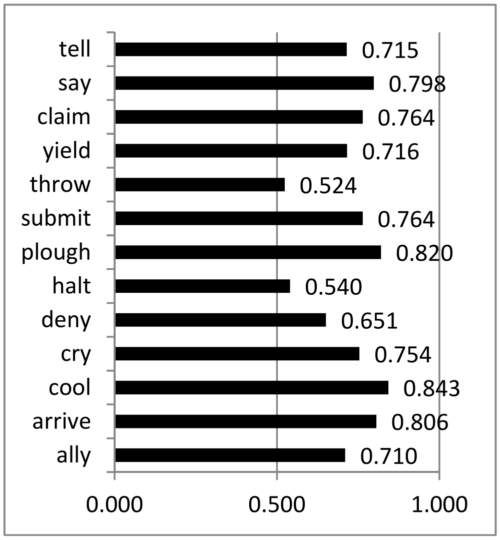


Fig. 6: IAA in the for-real verb set

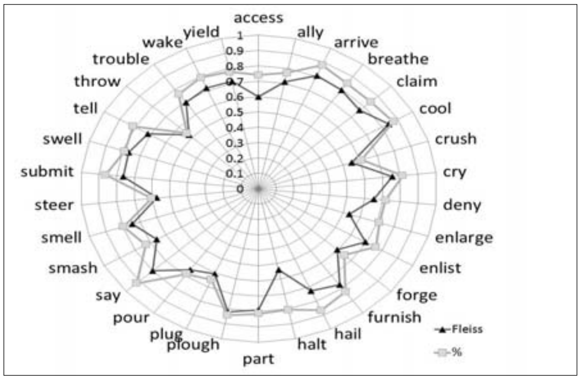


Fig. 7: Fleiss' kappa versus percentage

The annotation experience makes us believe that the intuitively sufficient IAA (as a measure of the quality of the entry) is specific to each individual verb. Moreover, if we neglect the fact that the annotators make errors, we still observe at least two factors that have a substantial effect on the agreement and we are not able to reflect them with the IAA measures – the interplay of the semantic distance between the categories and the structural or lexical opacity of the concordances.

Observing the implicatures, they are semantically distinct, yet seldom mutually exclusive. The patterns always differ at least in the semantic types or lexical set defined for a given collocate, but mostly also in the number of defined collocates and the syntactic structure in general. On the other hand, concordances can be **ambiguous** or **vague** for many reasons. For instance, the verb *hail* contains three semantically mutually exclusive categories, whose patterns have identical syntactic structure, but a different lexical population:

hail 4: [Human 1|Institution] hail [Human2 | {king|president|…} = authority] = celebrate arrival with greeting and welcoming in a ceremonial way
hail 6: [Human 1|Deity] hail [Human 2] = call sb from a distance
hail 7: [Human] hail {taxi|cab} = signal ordering a ride

These three implicatures are semantically distant. That they have the common feature of (possible) shouting is irrelevant. A speaker cannot reasonably want to say all these three things at the same time, hence the text is ambiguous. Here, an ambiguous context easily arises when the patient is populated with an unknown person name *(4, 6, 7)* or a ship *(6, 7)*. The name can belong to a taxi driver and the crew of the ship can be called by the crew of another ship. In this case, metonymy (person-car, person-ship) has caused the ambiguity. Ambiguity is, however, rare, as are rare semantically mutually exclusive categories in our selection of verbs. Mostly we face vagueness caused by syntactically opaque uses of the verb, where both categories, being not mutually exclusive, reflect the given event from a different perspective or emphasize a different aspect of it. A typical example is participial forms (see also 3.2.3). There is a passive - pseudo passive- adjective continuum: the agent is underspecified – the agent is irrelevant or by its nature not volitional – there cannot be any agent at all:

1. *[…] the house was <enlarged> in 1880 […]*
2. *[…] my lymph glands were <enlarged> as a result of HIV[…]*
3. *[…]his skills, cunning and physical powers were greatly <enlarged>, and all his tasks […] were accomplished […]*
4. *[…] the external naris and the narial fossa are not greatly <enlarged> as in sauropodomorphs […]*

The implicatures of both the participial pattern and the transitive pattern of *enlarge* imply that something becomes larger, the only difference being the added causativity. While the common knowledge says that houses are always built by people (1), it is up to individual interpretation whether HIV enlarges lymph glands or whether they enlarge themselves (2). When the "agent" in question is not even specified, the individual judgments can vary endlessly (3: magic, training, a couch?, 4: evolution?). This difference of perspectives hardly plays any role in understanding concordances like 1-4, but it drags the IAA down, whenever participial forms are frequent. This is the case of the following verbs: *ally, breathe, cool, crush, enlarge, enlist, furnish, part, plough, plug, steer, swell, trouble and wake.*

Another context-based disagreement source is semantic modulation (Cruse, 1996). The noun *traffic* denotes a lot of motor vehicles moving around (semantic type [Vehicle]) as well as the movement itself ([Action|

Event]). One category of the verb *halt* describes causative halting of events (such as *financial regression*), whereas another one describes halting vehicles. In case of *traffic*, the implicatures mean the same thing, although drawing a line between halting a convoy and halting *production* or *male vice* is intuitively easy and most concordances only belong to one category.

Nouns with strong meaning modulation potential often make the annotators mark the concordance as an exploitation of the ".a" type.

If we consider these language-inherent sources of disagreement by neglecting them as disagreements, the kappa score will be affected (Fig. 8). This time we use the average of Cohen's kappa over all annotator pairs ("pairwise Cohen's kappa"), since disregarding of selected disagreements cannot be computed in Fleiss' kappa. Nevertheless, the pairwise Cohen's kappa gives almost identical results for the basic IAA as the Fleiss' kappa (marked as "base" in Fig. 7). The "ignore expl" values indicate the IAA when exploitation marks within the same category number are ignored. The "merge part" values indicate the IAA when disagreements between participle patterns vs. other patterns are ignored. The "both" values indicate the IAA when both exploitations and participles are neglected. Neglecting exploitations and participle disagreements, 20 verbs reach over 0.8, 6 verbs reach over 0.7 and the others are all above 0.6. This is of course a very crude measurement, but it gives us a hint how important a detailed classification of disagreements is with respect to whether they result from a failure in the pattern design, from annotators' individual feed- back for the lexicographer, annotator errors or from the text-inherent vagueness or ambiguity, the aspects of which we want to explore more deeply.

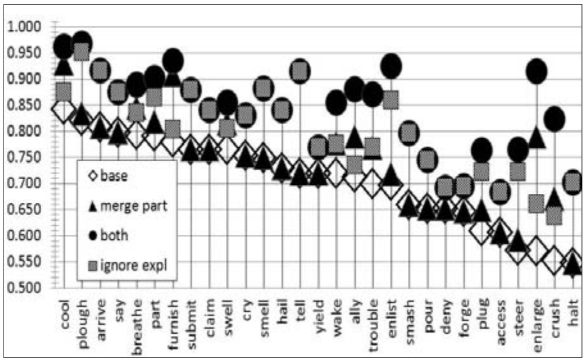


Fig. 8: Pairwise Cohen's kappa IAA for various settings

6 Ongoing and future work

We are still manually analyzing the interannotator disagreements in the way indicated above. We would like to explore the language-inherent factors that make concordances ambiguous or vague with respect to the patterns but do not blur the message of the text. We also seek to identify the interannotator disagreements that are caused exclusively by these factors.

Considering the coverage, we will have to focus on the verbs with more than 10,000 occurrences. We assume that the classical one-value annotation will be untenable for the frequent verbs that have extremely high collocability, and also, that a systematic solution must be found to capture their light verb uses.

We have been experimenting with a statistical pattern classifier. A bold next step would naturally be to mimic the human clustering process itself, based on larger amounts of data than a human lexicographer can oversee, and using it in real NLP tasks, such as machine translation, paraphrasing or recognizing textual entailment.

7 Conclusion

We have completed and released **VPS-30-En** (http://ufal.mff.cuni.cz/spr), a 30-verb lexical resource that draws on the Corpus Pattern Analysis method coined by P. Hanks. We have revised the entries as well as the annotated concordances on the basis of our experiments with multiple annotations.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013, and by the Czech Science Foundation under the projects P103/12/G084, P406/2010/0875 and P401/10/0792.

References

Artstein, R, Poesio, M. (2008). Inter-coder Agreement for Computational Linguistics. Computational Linguistics 34.4. pp. 555–596.

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Cinková, S., Hanks P. (2010). Validation of Corpus Pattern Analysis – Assigning pattern numbers to random verb samples. URL: http://ufal.mff.cuni.cz/spr/data/publications/annotation_manual.pdf .

Cruse, D.A. (1996). Lexical Semantics. Cambridge University Press.

Davies, M. (2008). The Corpus of Contemporary American English: 425 million words, 1990-present. Available at: http://corpus.byu.edu/coca/.

Hanks, P., Pustejovsky, J. (2005). A Pattern Dictionary for Natural Language Processing. In *Revue Francaise de linguistique appliquée*, 10:2.

Horák, A., Rambousek, A., Vossen, P. (2008). A Distributed Database System for Developing Ontological and Lexical Resources in Harmony. In *9th International Conference on Intelligent Text Processing and Computational Linguistics*. Berlin: Springer, pp. 1-15.

Kilgarriff, A. (1997). "I don't believe in word senses". *Computers and the Humanities*, 13 (2).

Kilgarriff, A., P. Smrz, P. Rychlý, D. Tugwell 2004. The Sketch Engine. Proc. Euralex, Lorient, France.

León A.P., Reimerink, A., Faber, P. (2009). Knowledge Extraction on Multidimensional Concepts: Corpus Pattern Analysis (CPA) and Concordances. In 8ème Conférence Internationale Terminologie Et Intelligence Artificielle. Toulouse, 2009.

Rumshisky, A., Pustejovsky, J. 2006. Inducing Sense-Discriminating Context Patterns from Sense-Tagged Corpora. In LREC 2006 Proceedings, Genoa,Italy.