# VPS-GradeUp: Graded Decisions on Usage Patterns

**Vít Baisa♣, Silvie Cinková♠, Ema Krejčová♠, Anna Vernerová♠**
♣Masaryk University, Brno, Faculty of Informatics, NLP Centre
♠Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
xbaisa@fi.muni.cz, {cinkova,krejcova,vernerova}@ufal.mff.cuni.cz

**Abstract**

We present **VPS-GradeUp** – a set of 11,400 graded human decisions on usage patterns of 29 English lexical verbs from the Pattern Dictionary of English Verbs by Patrick Hanks. The annotation contains, for each verb lemma, a batch of 50 concordances with the given lemma as KWIC, and for each of these concordances we provide a graded human decision on how well the individual PDEV patterns for this particular lemma illustrate the given concordance, indicated on a 7-item Likert scale for each PDEV pattern. With our annotation, we were pursuing a pilot investigation of the foundations of human clustering and disambiguation decisions with respect to usage patterns of verbs in context. The data set is publicly available at `http://hdl.handle.net/11234/1-1585`.

**Keywords:** CPA, graded decisions, English, verbs, usage patterns, annotation

## 1. Introduction

We present **VPS-GradeUp** – a set of 11,400 graded human decisions on usage patterns of 29 English lexical verbs from the Pattern Dictionary of English Verbs[1](Hanks, 2000 2014). The annotation contains, for each verb lemma, a batch of 50 concordances with the given lemma as KWIC (key word in context), and for each of these concordances we provide a graded human decision on how well the individual PDEV patterns for this particular lemma illustrate the given concordance, indicated on a 7-item Likert scale for each PDEV pattern.

This data set has been created to observe interannotator agreement on PDEV patterns produced using the Corpus Pattern Analysis (Hanks, 2013). The manually annotated concordances of PDEV are sometimes regarded as the gold standard for an intuitively plausible clustering of verb uses, and they have already been used for statistical machine learning. Efforts have been laid into automating either directly the clustering of unseen concordances from scratch or at least their classification according to predefined patterns; e.g. (Popescu, 2013; Baisa et al., 2015) and (Materna, 2013). In this context, we have created a new data set to investigate the foundations of *human* clustering and disambiguation decisions. VPS-GradeUp is now publicly available at `http://hdl.handle.net/11234/1-1585`.

## 2. Related Work

VPS-GradeUp draws on previous research associated with the VPS-30-En data set (Cinková et al., 2012). Both data sets are similar in that they provide parallel human annotations of PDEV data and that they process 50-concordance batches for each selected verb entry. Nevertheless, the data sets differ in two important aspects:

1. While VPS-30-En allows revisions of PDEV entries to optimize the interannotator agreement, VPS-GradeUp sticks to the original PDEV patterns (referring to their versions from March through October 2015);

2. While VPS-30-En contains only simple word-sense-disambiguation (WSD) decisions[2], VPS-GradeUp contains graded decisions as well: for each concordance of a given verb and each pattern in the corresponding PDEV entry, the annotator decides how well the particular pattern explains the given concordance in both syntactic and semantic terms, using a 7-item Likert scale. This implies that, for a verb with $n$ patterns, each annotator provides $50 \times n$ graded decisions (50 being the batch size).

The design of VPS-GradeUp draws mainly on the WSsim and USim data sets first reported by (Erk et al., 2009) and later more extensively in (Erk et al., 2013). WSsim contains graded decisions on matching relations between WordNet senses and concordances of 11 selected lemmas; USim contains graded decisions on how well two different words in two different sentences paraphrase each other. Since PDEV patterns are formulated as clauses, the similarity relation between a concordance and a pattern closely resembles that of paraphrases studied by USim.

Due to its focus on CPA and PDEV, our annotation is also slightly comparable to (Rumshisky et al., 2009), but methodologically we are substantially closer to (Erk et al., 2009), so we will only refer to the latter work.

## 3. Verb Selection

The verbs for our data set were randomly selected from the list of complete PDEV entries narrowed down to verbs with at least 3 patterns, not contained in VPS-30-En, and having at least 100 BNC (British National Corpus Consortium, 2007) sentences not previously annotated by PDEV annotators. We excluded one pattern number outlier

[1]The entry structure of PDEV is presented in Fig. 1. In a nutshell, each category in a PDEV entry contains a pattern (formulated as a finite clause template with argument labels from PDEV's own ontology) and an implicature – another finite clause template with semantic types, explaining or paraphrasing the meaning of the pattern.

[2]That is, each annotator assigned each concordance exactly one pattern number.

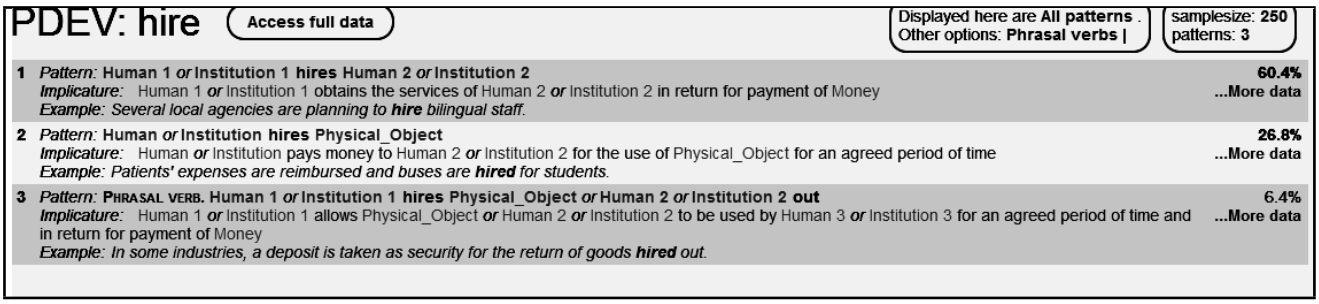



Figure 1: PDEV entry of *hire*–3 patterns

(*blow*, 62 patterns), making the pattern numbers range from 3 to 36. The resulting list was first divided into three frequency intervals. Each interval had the same number of members. Five verbs were randomly selected from each interval. The original list was subsequently reordered according to the number of patterns and was again divided into three equally large groups. The verbs previously selected due to the frequency criteria were then removed to ensure that no verb would be selected the second time. Again five verbs were randomly selected from each group. The final selection of 30 verbs contains *allocate, 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, say, seal, talk*, and *urge*. *Allocate* was used as a training verb and eventually removed from the set. The resulting set contains a noticeably large proportion of verbs starting with the initial letters of the alphabet, reflecting the population of complete PDEV entries. We decided not to randomize the candidate list with respect to the initial letters, since the verbs do not display any evident regularities associated with initial letters (frequency, number of patterns, identical prefixes/stems).

## 4. Annotation Scheme and Procedure

The annotation was carried out in online forms based on Google Forms (Google Forms, quoted 2016 02 16)[3], one form per verb. Each form contained one concordance analysis per page (Fig. 2). While annotating a verb, the annotators were referring to the corresponding PDEV entry displayed in a separate browser window (Fig. 1). By clicking the "Access full data" button, they could inspect concordances annotated with the best fitting pattern numbers (original PDEV annotation). In addition, the annotators were familiar with PDEV and its theoretical foundations, including the norms and exploitations. We had made sure that no concordances in our data set had been annotated in PDEV.

[3]Google Forms are respondent-friendly, free of charge, and relatively easy to generate automatically, but the verbs with the largest number of patterns seemed to be challenging their capacity: three forms failed to store the data in spreadsheets without any warning. In general, we were experiencing many crashes resulting in partial data loss. Therefore we cannot recommend using Google Forms as an annotation tool.

The annotation form, as illustrated by Fig. 2, starts with the concordance. The annotator may indicate comprehension uncertainty (a). Each concordance is accompanied by its identifier (unique within one verb lemma), the annotation question (c), and the grading of the Likert scales (d), with one Likert scale per pattern (e). The next part contains the WSD decision (f). The annotation of VPS-GradeUp is relatively rich, containing more than Likert and WSD-pattern number decisions. Conforming to the Theory of Norms and Exploitations (Hanks, 2013), the WSD number decision is complemented by exploitation markup (g); that is, when a concordance matched a given pattern with some reservations considering the syntax, lexical population of the arguments, or the overall meaning, the annotator ticked the corresponding multiple choice box for each type of reservation they were having.

## 5. Data Format

The data set comes in one csv file, where each row represents one Likert decision. It is identified by the pattern number, verb lemma, and sentence ID. The annotation decisions of each annotator – items a, e, f, g in Fig. 2 – are in separate columns (e.g. `Lik<AnnotatorName>` or `WSD<AnnotatorName>`), as illustrated in Fig. 3. The data set also contains annotators' comments and the concordances. We provide snapshots of the PDEV entries from the annotation period as images (cf. Fig. 1) and as structured text files.

## 6. Discussion

Before drawing any conclusions from the data, we measured the interannotator agreement. To compare our interannotator agreement with (Erk et al., 2009), we used Spearman's $\rho$. For the Likert decisions, the pairwise correlations were $\rho = 0.658$, $\rho = 0.656$ , and $\rho = 0.675$. For the WSD decisions, the pairwise correlations were $\rho = 0.785$, $\rho = 0.743$, and $\rho = 0.792$. The Fleiss' kappa for the WSD task was $0.76$. We are not considering the exploitation markup in either case.

All correlations are highly significant with $p < 2.2e^{-16}$. The observed correlations are even higher than those of WSsim and USim (between 0.466 and 0.504 in the 2009 paper), an outcome we had not expected. On the contrary, we had speculated that WSsim/USim data might have been slightly easier to annotate, given the lemmas processed: while VPS-GradeUp contains only

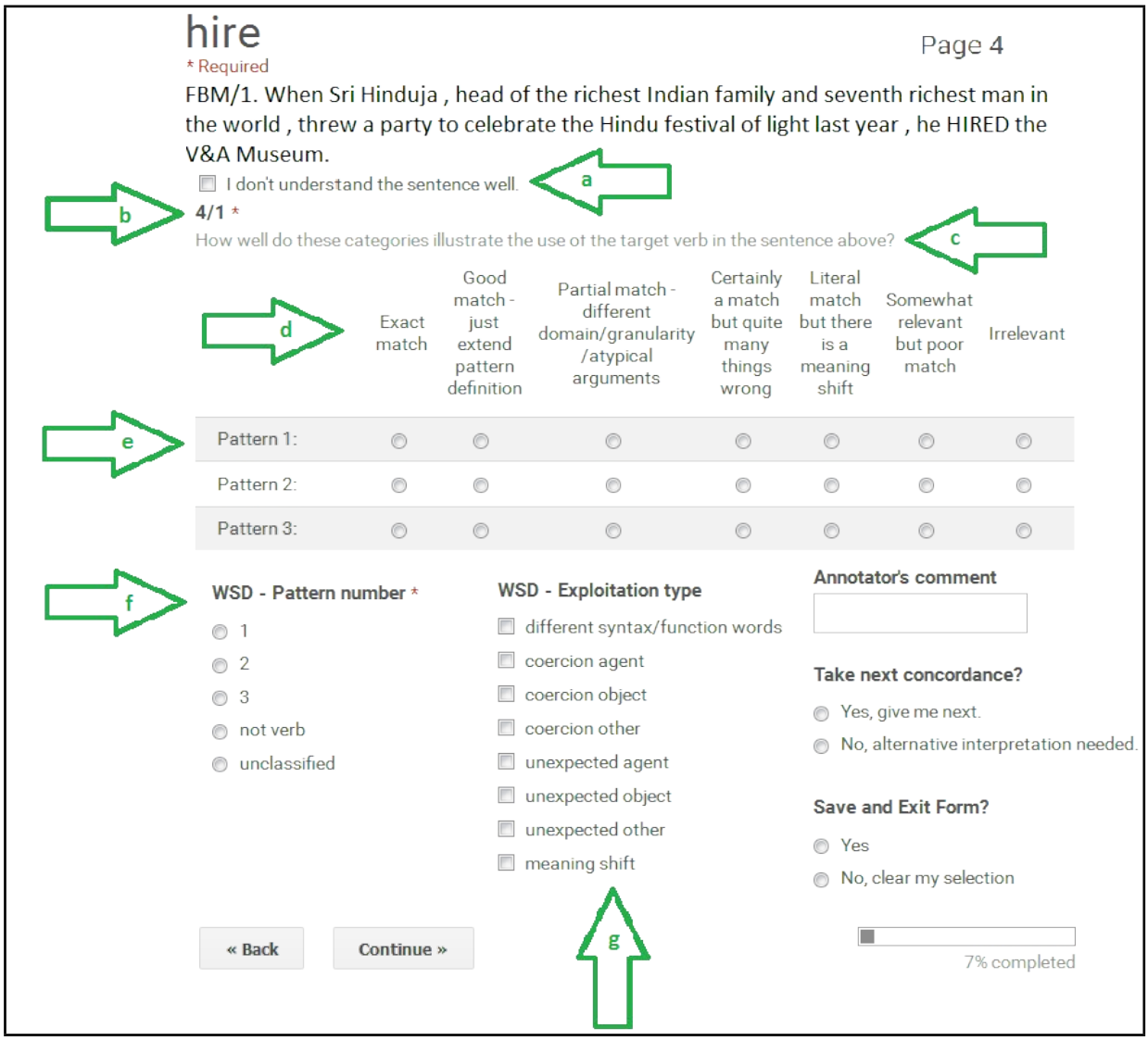



Figure 2: The annotation form using Google Forms

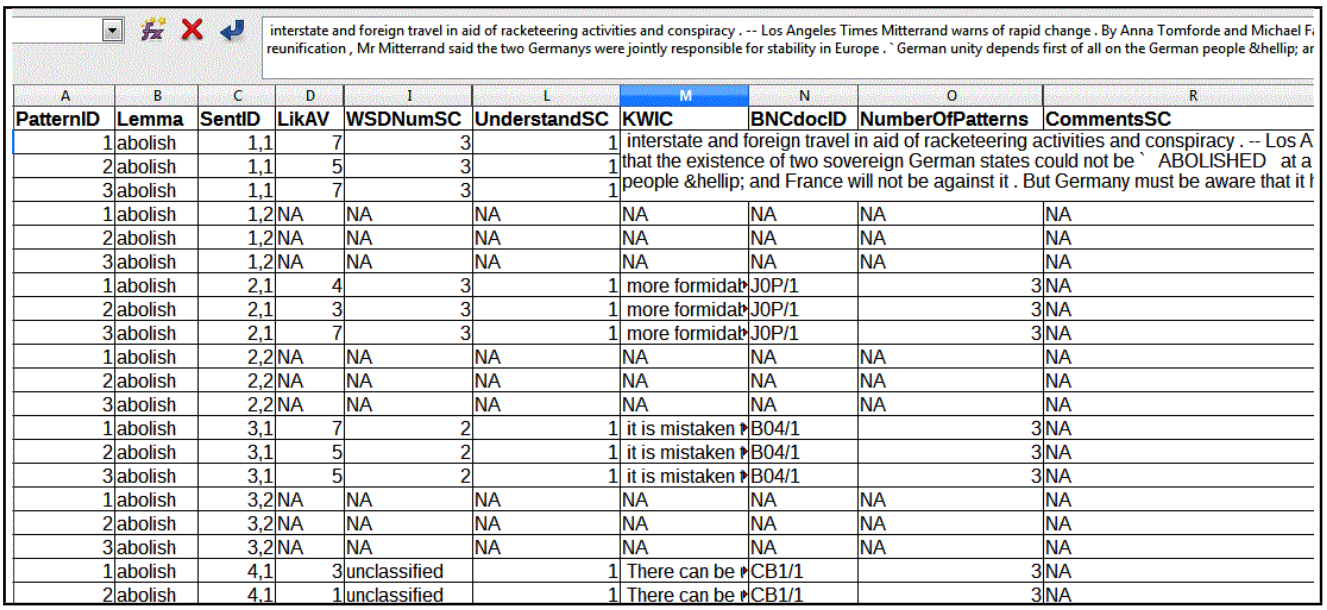
verb lemmas (29), WSsim contains 11 lemmas of various parts of speech (4 verbs: *add, ask, order*, and *win*, 5 nouns: *argument, function, interest, investigator*, and *paper*, and 2 adjectives: *important* and *different*). Verbs are generally expected to be more difficult to disambiguate than nouns, in particular when the reference lexicon is as fine-grained as PDEV. Also, we were using a 7-item Likert scale, whereas the WSsim/USim team used a 5-item Likert scale, and, intuitively, increased granularity ought to increase the risk of interannotator disagreement. On the other hand, WSsim/Usim were purposefully annotated by non-experts, whereas all our annotators are linguists (with high non-native proficiency in English) and had been working with PDEV before. Although they were not trained for Likert scales and were explicitly told to make spontaneous decisions, they are admittedly biased with respect to their previous WSD-annotation appointment and all accompanying discussions[4].

Even considering all conditions limiting the power of this comparison, our interannotator agreement results clearly show that a graded annotation of patterns is no less sensible than a graded annotation of senses.

Two remarks should be added to our discussion of Fig. 2 concerning potentially controversial or otherwise interesting issues: the description of the Likert items and the *"take next concordance"/"let me add an alternative annotation"* option.

The attentive reader might have noticed in Fig. 2 that the items on the Likert scales carry anchor descriptions instead of numbers, which could possibly make the sentence-pattern matching a categorical variable rather than an ordinal one, with no need to be displayed as a Likert scale.

[4]To provide a sounder comparison with (Erk et al., 2009), we should of course have addressed non-expert respondents. In this respect, our respondents only represent a convenience sample. We have not experimentally tested whether experts provide different results than non-experts. Nevertheless, the CPA as the underlying theory focuses on pattern creation more than on pattern matching and seems not to provide more specific criteria for the goodness of match between a pattern and a concordance than those we incorporated into the anchors on the scale, which makes us assume that non-expert annotators would not have significantly differed from expert annotators.



interstate and foreign travel in aid of racketeering activities and conspiracy . -- Los Angeles Times Mitterrand warns of rapid change . By Anna Tomforde and Michael F
reunification , Mr Mitterrand said the two Germanys were jointly responsible for stability in Europe . ` German unity depends first of all on the German people … ar

| A | B | C | D | I | L | M | N | O | R |
|---|---|---|---|---|---|---|---|---|---|
| PatternID | Lemma | SentID | LikAV | WSDNumSC | UnderstandSC | KWIC | BNCdocID | NumberOfPatterns | CommentsSC |
| 1 | abolish | 1,1 | 7 | 3 | 1 | interstate and foreign travel in aid of racketeering activities and conspiracy . -- Los A | | | |
| 2 | abolish | 1,1 | 5 | 3 | 1 | that the existence of two sovereign German states could not be ` ABOLISHED at a | | | |
| 3 | abolish | 1,1 | 7 | 3 | 1 | people … and France will not be against it . But Germany must be aware that it h | | | |
| 1 | abolish | 1,2 | NA | NA | NA | NA | NA | NA | NA |
| 2 | abolish | 1,2 | NA | NA | NA | NA | NA | NA | NA |
| 3 | abolish | 1,2 | NA | NA | NA | NA | NA | NA | NA |
| 1 | abolish | 2,1 | 4 | 3 | 1 | more formidab | J0P/1 | 3 | NA |
| 2 | abolish | 2,1 | 3 | 3 | 1 | more formidab | J0P/1 | 3 | NA |
| 3 | abolish | 2,1 | 7 | 3 | 1 | more formidab | J0P/1 | 3 | NA |
| 1 | abolish | 2,2 | NA | NA | NA | NA | NA | NA | NA |
| 2 | abolish | 2,2 | NA | NA | NA | NA | NA | NA | NA |
| 3 | abolish | 2,2 | NA | NA | NA | NA | NA | NA | NA |
| 1 | abolish | 3,1 | 7 | 2 | 1 | it is mistaken | B04/1 | 3 | NA |
| 2 | abolish | 3,1 | 5 | 2 | 1 | it is mistaken | B04/1 | 3 | NA |
| 3 | abolish | 3,1 | 5 | 2 | 1 | it is mistaken | B04/1 | 3 | NA |
| 1 | abolish | 3,2 | NA | NA | NA | NA | NA | NA | NA |
| 2 | abolish | 3,2 | NA | NA | NA | NA | NA | NA | NA |
| 3 | abolish | 3,2 | NA | NA | NA | NA | NA | NA | NA |
| 1 | abolish | 4,1 | 3 | unclassified | 1 | There can be | CB1/1 | 3 | NA |
| 2 | abolish | 4,1 | 1 | unclassified | 1 | There can be | CB1/1 | 3 | NA |

Figure 3: Selected columns of the resulting file: only one annotator judgment shown per research item ("AV" and "SC" are annotator initials). Exploitation judgements are entirely hidden. The regularly occurring NA values are the non-present values of an alternative interpretation, whose sentence ID ends with 2. Alternative interpretations are very rare.

Yet we argue that we have preserved the ordinal character of the categories and rightly captured them in the final data set as an ordinal variable spanning from 1 to 7 (*Irrelevant = 1, Exact match = 7*). The hints describing the respective categories have evolved from the previous experience with VPS-30-En, where we had substantially decreased the interannotator confusion by instructing annotators to prioritize meaning over form (i.e. implicature match over pattern conformity), and they represent a syntax-semantics continuum. When in doubt, the annotators were instructed to ignore the hints and proceed by their pure intuition of this continuum, though, since even minor syntactic or lexical deviations can cause a substantial semantic shift and make it impossible to exactly judge which type of non-conformity is predominant.

The alternative annotation option reflects the difference between vagueness and ambiguity. While vagueness is perceived as a state where several statements (here patterns) can apply simultaneously, ambiguity arises when the interpretations are so incompatible that we cannot reasonably assume that the speaker intended to convey both at the same time. The multiple Likert scales are supposed to capture vagueness, not ambiguity. Whenever a concordance is ambiguous, the annotator is supposed to treat each reading separately, since, if the readings are truly ambiguous, they ought to give different sets of concordance-pattern-matching judgments. The alternative readings have been reserved only for clear and well-understood ambiguity cases. They were meant to be used sparingly, as ambiguity rarely occurs in context and the annotators were given the largest context permitted by the Sketch Engine (Kilgarriff et al., 2004). The alternative reading could not be combined with the *"I do not understand the sentence"* option, and only one was allowed (assuming that a more complicated ambiguity always deservers the *"poor understanding"* choice or *"not verb"* in the WSD decisions). On the whole, there are so few alternative readings in the data set that they could hardly harm the interannotator agreement, so we have not processed them in any sophisticated way we had been considering before obtaining the results. When one annotator provided an alternative reading and another not, we simply observed more disagreement between them. We have not met a case where two annotators would both assign an alternative reading and disagree in which was the primary and which the secondary.

## 7. Conclusions and Future Work

According to (Erk et al., 2009), p. 17, data sets with graded lexical decisions "can hopefully be used for evaluating methods which relate usages without necessarily producing hard clusters". To the best of our knowledge, VPS-GradeUp is the only such data set in existence beside WSsim and USim.

Having built such a richly annotated resource, we are going to explore it further. We plan to investigate which features of dictionary entries on the one hand, and which features of concordances on the other hand, pose the greatest disambiguation obstacles, and whether our findings can be generalized to obtain a measure of disambiguation difficulty of a concordance relative to a given set of senses and shed more light on what is happening behind the scenes in human semantic clustering judgments.

We have ensured a fully random selection of verbs to annotate. Therefore we can generalize our PDEV-on-BNC annotation results beyond our sample, at least regarding PDEV entries with no less than three patterns.

## 8. Acknowledgements

This project was supported by the Czech Science Foundation grant GA-15-20031S, the Norwegian Financial

Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047, 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).

## 9. Bibliographical References

Baisa, V., Bradbury, J., Cinkova, S., El Maarouf, I., Kilgarriff, A., and Popescu, O. (2015). SemEval-2015 Task 15: A CPA dictionary-entry-building task. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 315–324, Denver, Colorado, June. Association for Computational Linguistics.

Cinková, S., Holub, M., Rambousek, A., and 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)*, pages 3176–3183, Istanbul, Turkey. European Language Resources Association.

Erk, K., McCarthy, D., and 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*, pages 10–18, Suntec, Singapore, August. Association for Computational Linguistics.

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

Google Forms. (quoted 2016-02-16). Forms Service | Apps Script | Google Developers.

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

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The Sketch Engine. In *Proceedings of EURALEX*.

Materna, J. (2013). Parameter Estimation for LDA-Frames. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 482–486, Atlanta, Georgia, June. Association for Computational Linguistics.

Popescu, O. (2013). Learning Corpus Patterns Using Finite State Automata. *FBK-irst, Trento*.

R Core Team, (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rumshisky, A., Verhagen, M., and 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.

## 10. Language Resource References

British National Corpus Consortium. (2007). *British National Corpus, version 3 (BNC XML edition)*. British National Corpus Consortium.

Hanks, Patrick. (2000-2014). *Pattern Dictionary of English Verbs*. Patrick Hanks, University of Wolverhampton, http://pdev.org.uk/about_cpa, quoted 2016–02–15.