Graded and Word-Sense-Disambiguation Decisions in Corpus Pattern Analysis: a Pilot Study

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

Abstract

We present a pilot analysis of a new linguistic resource, VPS-GradeUp (available at `http://hdl.handle.net/11234/1-1585`). The resource contains 11,400 graded human decisions on usage patterns of 29 English lexical verbs, randomly selected from the Pattern Dictionary of English Verbs (Hanks, 2000 2014). The selection was random and based on their frequency and the number of senses their lemmas have in PDEV. This data set has been created to observe the interannotator agreement on PDEV patterns produced using the Corpus Pattern Analysis (Hanks, 2013). Apart from the graded decisions, the data set also contains traditional Word-Sense-Disambiguation (WSD) labels. We analyze the associations between the graded annotation and WSD annotation. The results of the respective annotations do not correlate with the size of the usage pattern inventory for the respective verbs lemmas, which makes the data set worth further linguistic analysis.

Keywords: CPA, graded decisions, English, verbs, usage patterns, annotation, Likert scales

1. Introduction

We investigate human graded decisions about the goodness of match between lexicographically defined usage patterns of verbs and random corpus concordances of those verbs in terms of interannotator agreement. The usage patterns originate from PDEV – the Pattern Dictionary of English Verbs (Hanks, 2000 2014) –, and the concordances come from the BNC (British National Corpus Consortium, 2007). PDEV is a lexical resource driven by the Corpus Pattern Analysis – an interesting method of lexical description (Hanks, 2013) based on manual syntactico-semantic clustering of random concordances of a verb into usage patterns. Lexical entries produced by this method are intuitively very appealing for human language learners. Besides, a number of attempts have been pursued to introduce PDEV as a resource or Corpus Pattern Analysis as an approach into the computational linguistics – for instance, several recent SemEval tasks targeted CPA (Baisa et al., 2015). While the computational tasks encompass classification of unseen sentences according to PDEV patterns already available as well as generating CPA-style lexical entries from large data, our investigation aims at what factors play a role in the *human* classification and clustering decisions. We hope that this research will help with a more nuanced evaluation of the classification and entry-building tasks.

As a first step, we built a toy lexical resource with 29 PDEV entries and 50-item batches of the corresponding concordances. It contains two types of annotation:

- graded decisions about the goodness of match between a concordance and *each* pattern listed in the corresponding PDEV entry;
- traditional WSD decisions (the best matching pattern for each concordance).

This lexical resource is called VPS-GradeUp ((Baisa et al., 2015) and is available through the LINDAT-CLARIN repository at `http://hdl.handle.net/11234/1-1585`).

It contains 11,400 graded human decisions as well as 1450 WSD decisions concerning these 29 verbs: *seal, sail, distinguish, adjust, cancel, need, approve, conceive, act, pack, embrace, see, abolish, advance cure, plan, manage, execute, answer, bid, point, cultivate, praise, talk, urge, last, hire, prescribe*, and *murder*.

The verb entries were selected randomly, based on their frequency and the number of patterns in PDEV. For more details on how the random verb sample was constructed and how the resource is structured see (Baisa et al., 2016), this volume. VPS-GradeUp draws on previous research associated with the VPS-30-En data set (Cinková et al., 2012)[1] and on (Erk et al., 2009; Erk et al., 2013).

In (Baisa et al., 2016), we have already reported that:

1. nothing suggests that the graded-decision annotation of usage patterns be less successful than graded-decision annotation of senses in the traditional lexicographic conception (cf. (Erk et al., 2013));
2. the correlation between annotators is slightly lower on the graded-decision task than on the traditional Word-Sense-Disambiguation setup, but still both significant and strong.

In this paper, we are making the first comparisons of both annotations to explore their mutual associations, and we look into the results of each annotation task at the level of individual verbs.

2. Methodology

Three annotators performed two tasks in an online survey (Google Forms): Graded decisions and WSD for 50-concordance batches per verb lemma. All were linguists,

[1] which in its turn contains 30 WSD-annotated verbs from PDEV revised to optimize the interannotator agreement

familiar with Corpus Pattern Analysis, and non-native speakers of English with high proficiency. Both tasks were processed simultaneously and in the same document (with each survey page containing one concordance from the BNC), with the graded decisions coming first.

The annotation form, as illustrated by Fig. 1, starts with the concordance. The annotator may indicate comprehension uncertainty (a). Each concordance is accompanied by its identifier (unique within one verb lemma) and the annotation question (c). Each PDEV pattern obtains a grade on a Likert scale (e) The Likert scales contain anchors, which lie on a continuum from syntactic to semantic criteria (d). During the analysis, the Likert scale was converted to a numerical scale with values from 7 ("exact match") to 1 ("irrelevant"). The next part contains the WSD decision (f). The annotation of VPS-GradeUp is relatively rich, containing more than just the 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 might be having.

The collected data was saved in separate spreadsheets. The final resource comes as one csv file with each row representing one Likert decision identified by the pattern number, verb lemma, and sentence ID. The WSD decisions associated with a given sentence ID are copied to all relevant rows. The annotation decisions of each annotator are located in separate columns.

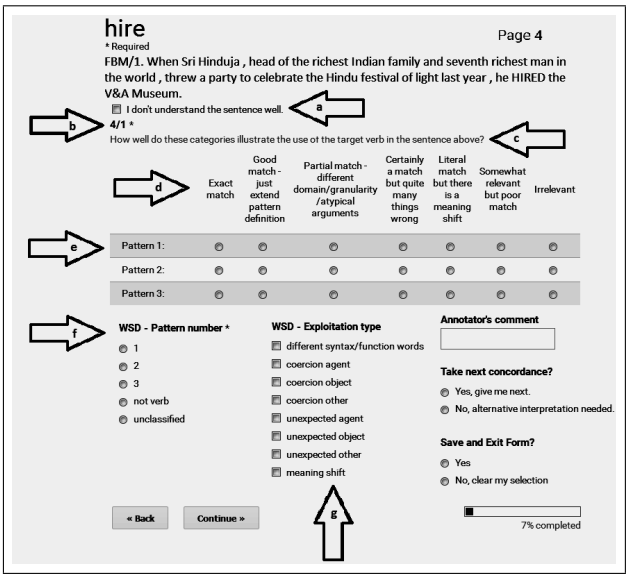

Figure 1: The annotation form using Google Forms.

3. Results

3.1. Interannotator Agreement on WSD

We were analyzing two dataset versions, which we call **Complete** and **VerbsOnly**. In the VerbsOnly dataset, we excluded concordances in which at least one annotator classified the target use of the verb lemma as "not verb", to particularly focus on disagreements between numbered patterns rather than on the well-known fuzzy borders between verbs and other parts of speech. The "not verb" situations typically occurred when the verb was used in a participle form; less frequently due to a tagging error in the BNC. The results presented in the coming chapters are drawing on the VerbsOnly dataset. The WSD setup has yielded Fleiss kappa of 0.76 and 0.78 on Complete and VerbsOnly, respectively. The mean percentual agreement rates lie at 0.79 and 0.81 with the standard deviation below 0.02 in both cases. The observed results display a very weak and also grossly non-significant correlation with the number of patterns (computed by Spearman's ρ with continuity correction) and are thus worth the effort of a more sophisticated analysis. Fig. 2 shows the inter-annotator agreement results for the individual verbs. The verbs are ordered according to the percentual agreement in VerbsOnly. Apart from the verbs *last, plan, cure, approve*, and, in particular, *murder*, the percentual agreement gives a picture similar to Fleiss kappa. *Murder* is a special case: it has only three numbered patterns, of which only one was used. For kappa, this means that the decisions must be easy to take, and each confusion results in a substantial score decrease – which, in this case, does not reflect the actual annotation success, as the entire 3 disagreement cases in the 50-concordance batch were concerning the part-of-speech ("Pattern 1" vs. "not verb") – hence the striking difference between its kappa scores on Complete compared to VerbsOnly and the high percentual agreements. With one exception (*last*), the Fleiss kappas are slightly higher or remain the same on VerbsOnly. The most problematic verbs, in terms of Fleiss kappa as well as percentages, were *approve, seal* and *sail* (under 0.6), along with *plan* and *need* (just around 0.6).

3.2. Interannotator Agreement on Graded Decisions

The estimation of the interannotator agreement on graded decision tasks is less straightforward and deserves to be approached from several different angles. A good first approximation is naturally the personal bias of each annotator. Fig. 3 reveals no dramatic shifts, with just one annotator slightly tending towards more positive marks, effectively within the range of two points from the median (middle value) of the decision triple. Given the fact that the most frequently assigned marks were 1, 2, and 7, as well as the evident dominance of zero-ranged decision triples (i.e. the judgments by all three annotators concerning one particular pattern of a particular verb used in a particular sentence), we observe interannotator agreement good enough to pursue more detailed investigations without normalization.

Pairwise annotator correlation rates were $\rho = 0.658$, $\rho = 0.656$, and $\rho = 0.675$. For the WSD decisions, pairwise correlations were $\rho = 0.785$, $\rho = 0.743$, and $\rho = 0.792$ (Spearman's ρ). All correlations are highly significant with $p < 2.2e^{-16}$. The observed correlations are higher than those of WSsim and USim reported in (Erk et al., 2009) as

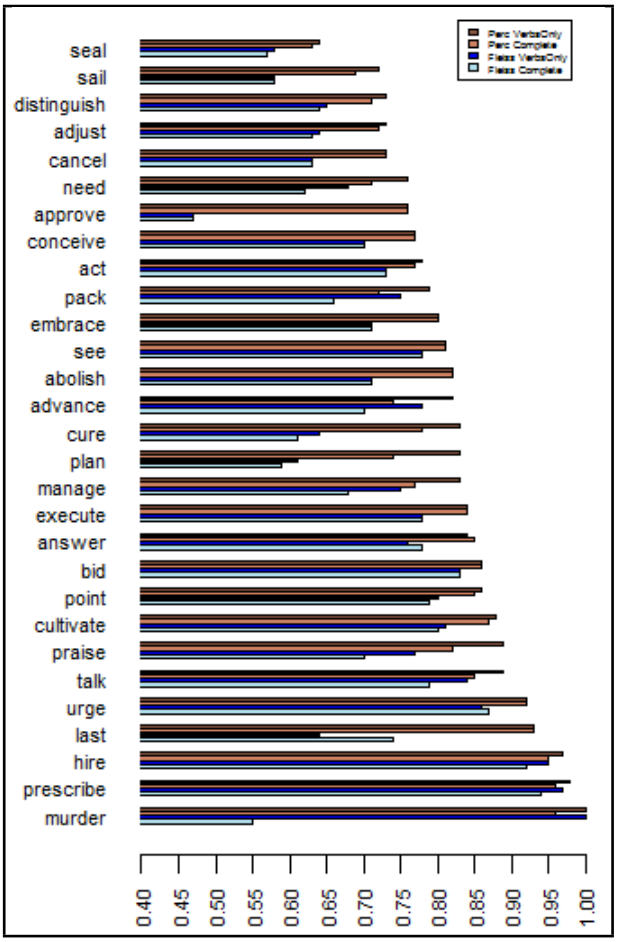

Figure 2: Lemma-wise values of Fleiss kappa and percentual agreements on Complete vs. VerbsOnly. The lemmas are presented in ascending order according to the percentual agreement on VerbsOnly.

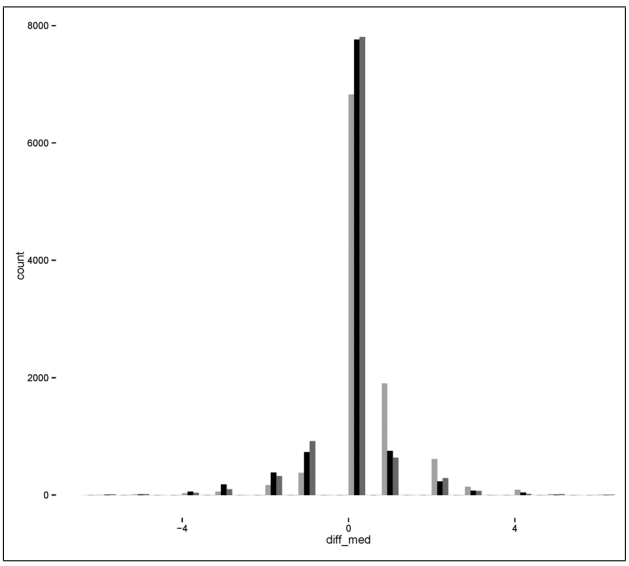

Figure 3: Annotator bias. On the x-axis is the difference between the chosen annotator's choice and the median (middle value) of the annotation triple; on the y-axis is the total number of times that this difference was achieved by the annotator.

between 0.466 and 0.504.

3.3. Distribution of Marks and Ranges

To analyze the interannotator agreement of the graded-decision setup for all annotators simultaneously, we were observing the median (i.e. the middle value) and range (i.e., the difference between the maximum value and the minimum value) of each decision triple. Fig. A4 represents the results for the individual verbs ordered according to the WSD percentual results on VerbsOnly (cf. Fig. 2). The 30th image in the bottom-right corner represents the overall results across the entire dataset. The heat map axes represent the decision medians (vertical) and range (horizontal). Each cell represents a percentual proportion of the given median-range combination within the particular verb, the color scale being dark (low) – bright (high).

The decision range shows the agreement of the three annotators on the median value – it is the difference between the lowest and the highest decision score. For instance, a pattern in a sentence can have been classified with 3, 6, and 7. The median is in this case 6 and the range is 4 as a result of 7 - 3. This decision is very positive, but on the other hand not particularly reliable because of the large range. The better the annotators agree on a particular pattern - concordance match, the smaller the range of their judgments is.[2]

In other words, Median and Range translate as "Goodness of match between a pattern and a concordance" and "Agreement on this goodness of match", respectively. High counts (indicated by bright color) in the top left corner mean high agreement on high goodness of match, while high counts in the bottom left corner mark high agreement on poor goodness of match. High counts further to the top right corner would indicate that there were cases some annotators considered good (as good as far high the cell lies), but one annotator considered them worse matches.

The brighter the left part of each individual image is, the better the agreement of the annotators. The vertical axis, on the other hand, indicates the goodness of match. The further up each individual image, the higher the median value; i.e., the brighter the upper part of the image, the more excellent matches between patterns and concordances occurred.

3.4. Hierarchical Cluster Analysis

We have also performed a hierarchical cluster analysis of the annotations of individual verbs, considering the cumulative proportional agreement within a given range (Fig. A5), to highlight annotation similarities between them. We see 16 clusters based on the similarity of annotation reflected in terms of mark medians and ranges. The curves representing individual verbs are compared to the average curve obtained from all ranges across all verbs, in descending order. The most substantial differences occur between ranges 0 and 2. As we know from Fig. A4, these have mostly occurred with the extreme marks (1-2 vs. 7). Note that, unlike the previous analysis in Fig. A4, this

[2] In terms of range, the interannotator agreement would be the same with the values 3,7, and 7: the fact that two values were identical is disregarded.

analysis does not consider the median marks with which the agreement cases occurred.

4. Discussion

Figure A4 offers some explanations of the "success" of the individual verbs in the WSD task, as it gives us an overview on which verbs are easy vs. difficult to annotate in the WSD setup, suggesting at least two factors that determine the level of difficulty: the syntactico-semantic distance between the patterns on the one hand and how well the entire entry fits the new data on the other hand. A verb can be **difficult to annotate in WSD terms** when:

1. **too many concordances fit several patterns very well with a range greater than 0** – when the annotators have to pick just one pattern from two or more that they classify as equally good, their choice and hence the interannotator agreement is a matter of coincidence;

2. **high numbers concentrate in the bottom part of the image** – in too many concordances, the annotators lack good options. Theoretically they can agree on classifying such cases as "unclassifiable", but in the practice we often see different combinations of disagreement between "unclassifiable", "not a verb", and various numbered patterns with various objections (which the annotators are supposed to record as "exploitations" according to the CPA).

While the former case can yield interesting insights regarding the syntactico-semantic distance between the patterns, the latter case signals a problem in the entry or in the data. An issue in the entry typically occurs when no pattern available is semantically close to the annotator's interpretation of the data. An issue in the data typically occurs when too many verbs are used in their participle forms in positions where they can easily be classified as nouns or adjectives or/and with their arguments underspecified.

The clusters in Fig. A5 are based on the graded-decision ranges unrelated to the medians of the decisions. This figure shows nicely that there is not necessarily an automatic mapping between "success" in WSD and "success" in graded decisions, judging by two verbs having ranked top on WSD but scoring particularly poorly in graded decisions: *talk* and, in particular, *urge*. The heatmap of *urge* in Fig. A4 (top row, fourth cell from the left) explains it, revealing that the most frequent cases of high range never really interfered with the best matching pattern. The second most frequent choice had median 5, which says it was not an ideal match. Nevertheless, if that pattern was the best-matching pattern for a (number of) concordance(s) in the WSD task (the image can't tell), it is likely to have been selected and agreed upon: a median of 5 is still high enough for the annotator not to consider a concordance "unclassifiable", and other considered patterns were seldom serious competitors, given the low frequency of other decisions on the upper part of the scale to the left (i.e., with reasonable range).

To name a counterexample to *urge*, *distinguish* (5th row, first left in A4) was very poor in WSD, but it scored very well in graded decisions, which is understandable, given all the brightness of the left bottom heatmap corner compared to the left top heatmap corner: the annotators often agreed that a pattern was a poor match, but they hardly ever reached a common positive decision about a match. This result suggests a problem in how well the entry covers the data encountered in the annotation experiment.

5. Conclusion and Future Work

We have performed pilot investigations of the results of two annotation tasks: graded similarity decisions and Word Sense Disambiguation on CPA verb patterns. In the first stage, which we are presenting here, we have mainly analyzed the interannotator agreement. Although the annotation tasks are both addressing the same general issue, namely sense matching, their results do not correlate in a straightforward way. In the near future, we are going to explore the data to the level of individual pattern disagreements in individual concordances to formulate hypotheses about the impact of different entry designs vs. morphosyntactic and semantic characteristics of the analyzed concordances on the graded decisions.

6. 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).

7. 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.

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

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.

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

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

8. Language Resource References

Baisa, Vít and Cinková, Silvie and Krejčová, Ema and Vernerová, Anna. (2015). *VPS-GradeUp*. Charles University in Prague, Faculty of Mathematics and Physics.

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.

Appendix A Images

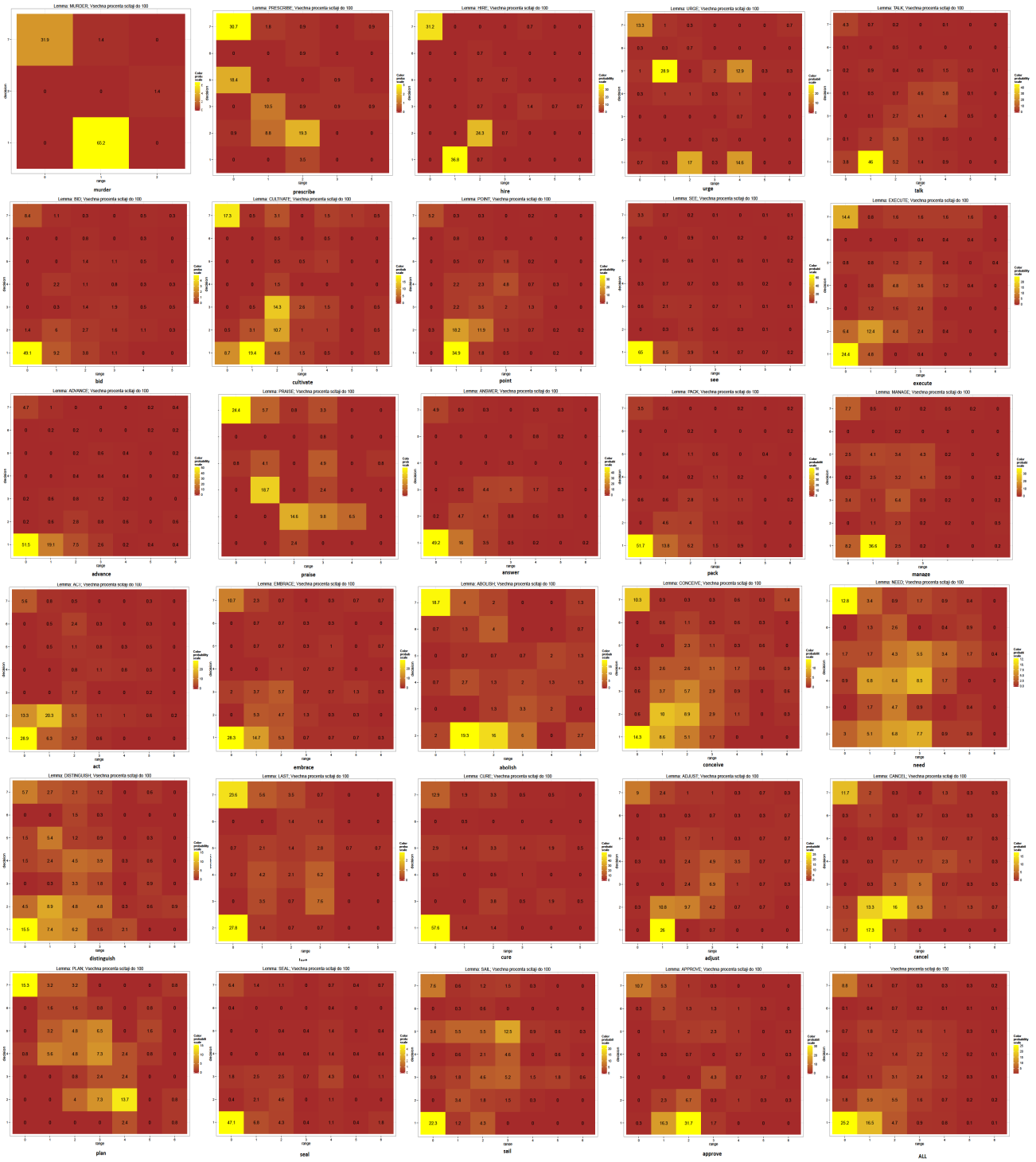

Figure A4: Decision ranges for each verb – the cells are listed in descending order according to their percentage agreement rates in the WSD task. The vertical and the horizontal axis represent the decision marks and the range respectively (with the cell corresponding to median value 1 and range value 0 in the bottom left corner of each heatmap). From top left to bottom right: *murder, prescribe, hire, last, urge/talk, praise, cultivate, point, bid/answer, execute, manage, plan, cure/advance, abolish, see, embrace, pack/act, conceive, approve, need, cancel/adjust, distinguish, sail, seal*, the overall result. The color scale of the cells renders the percentage of the total number of decisions for the given verb lemma. The brighter the color, the higher the percentage.

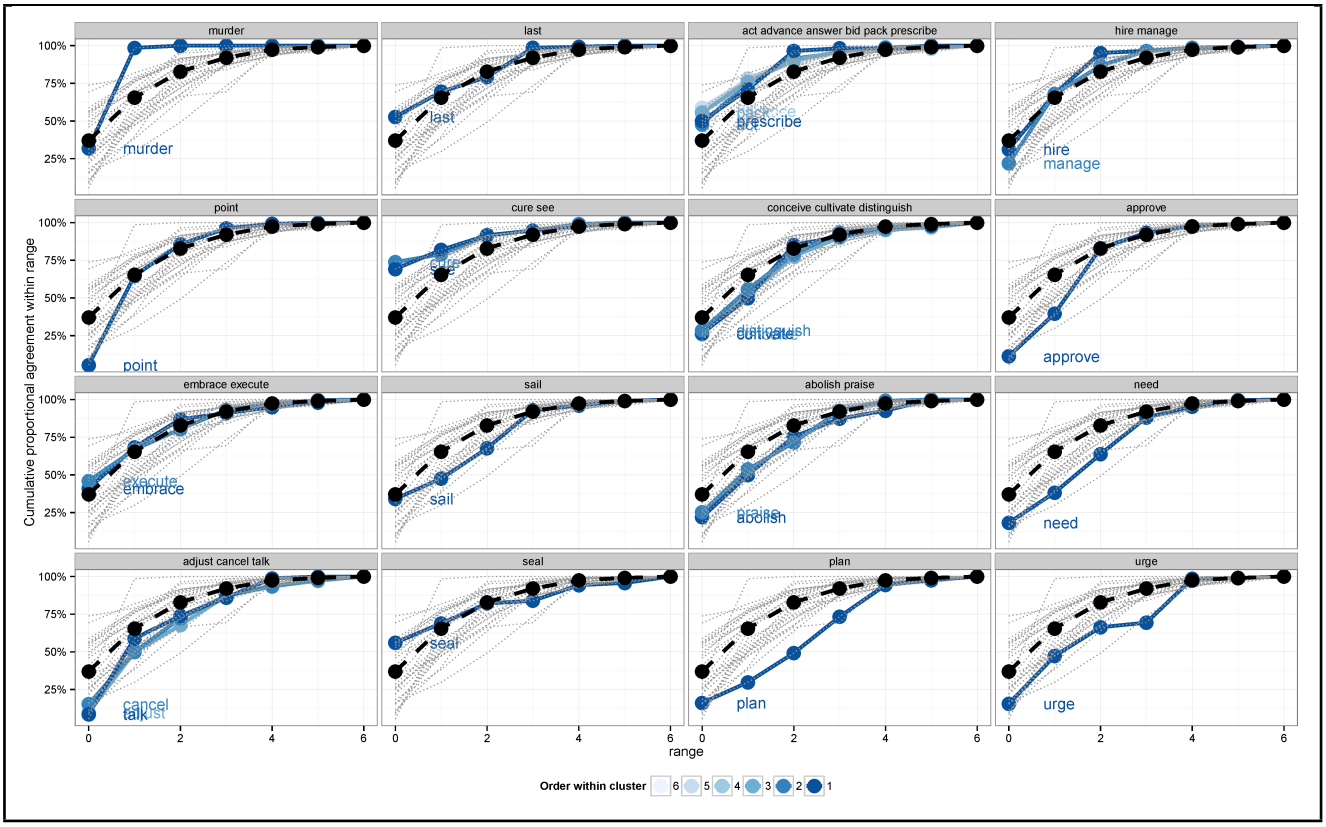

Figure A5: Hierarchical cluster analysis of agreement within a given range on individual verbs. The Y axis indicates the cummulative proportional agreement with range, i.e. how large a proportion of judgments on the given verb was passed with the range of zero, one, two, etc. For instance, in *murder*, some 30% of the judgments have zero range (i.e. are unanimously agreed on). When we include judgments with the range of one, we include other 70%. With *murder*, we only find ranges of zero or one, while in *urge* almost 15% of judgments have a range of four. The black curve is the average over the entire data set.