Annotating Decomposition in Time: Three Approaches for *Again*

Martin Kopf and **Remus Gergel**
Saarland University
{martin.kopf, remus.gergel}@uni-saarland.de

Abstract

This paper reports on a three-part series of original methods geared towards producing semantic annotations for the decompositional marker *again*. The three methods are (i) exhaustive expert annotation based on a comprehensive set of guidelines, (ii) extension of expert annotation by predicting presuppositions with a Multinomial Naïve Bayes classifier in the context of a meta-analysis to optimize feature selection and (iii) quality-controlled crowdsourcing with ensuing evaluation and KMeans clustering of annotation vectors.

1 Introduction

The goal of this paper is to present a series of three original methods in the context of, and first hands-on results for, ascertaining theoretically relevant ambiguities in readings of historical data on decomposition. Decompositional adverbs (e.g., *again* and its relatives in many languages) have attracted attention not only in the context of formal analyses (say, structural vs. lexicalist) since they are insightful, if not uncontroversial, in their own right. They also touch on the representation of events, presuppositions, and more generally the way the structural and the meaning components of particular languages are to be related (cf. Rapp and Stechow, 1999; Beck, 2005; Zwarts, 2019; Ausensi et al., 2021, among many others). Moreover, recent inquiries into diachronic formal semantics have crucially shown that diachronic data can not only receive motivated theoretical analyses but are also able to elucidate synchronic debates that could not be solved otherwise thus far (Beck and Gergel, 2015; Degano and Aloni, 2022). However, major practical issues with much needed diachronic data are the costly process of extraction w.r.t. high-quality data, their reliable annotation, stronger validation (than, say, the intuitions of individual researchers), and, when possible, partially automatic amplification/replication. The structure of this paper is as follows: In section 2, we start off with a discussion of the English adverb *again* and its main readings – as relevant to the discussion at hand. Next, we discuss the three methods for producing semantic annotations for *again*: In section 3, we go into detail regarding the procedure behind exhaustively annotating its various readings with a team of expert annotators based on syntactically parsed diachronic corpora of English (ranging throughout recorded history, from Old to Modern English; our concrete focus here lies ‘only’ on the last two to four centuries). The first slice of this semantic annotation, i.e., all 1,901 uses of *again* in the Penn Parsed Corpus of Modern British English (2nd ed., ‘PPCMBE2’, cf. Kroch et al., 2016), is ready to be shared with the community along with a tool to be merged with users’ own instances of the PPCMBE2. The second method discussed in this paper (section 4) seeks to tap into the semantically enriched data and extend the expert annotation: We discuss the performance of a Multinomial Naïve Bayes classifier in predicting the main readings of *again* in PPCMBE2. We do so in the context of a meta-analysis exploring the best-performing feature combinations based on a set of 16 different features of three different major types (features based on our semantic annotation, structural features drawn from the pre-existing syntactic parsing, and ‘naïve’ features based on the textual surface). We cover the third and final approach in section 5. It reports on what we call an ‘informed crowdsourcing experiment’, which we designed to explore crowd aptitude for providing nuanced semantic annotations on diachronic data – natural language data for which our (‘informed’) crowd workers can have no actual native speaker intuitions whatsoever (as the bearers of truly native intuition are dead). Here we report on the performance of KMeans clustering of the crowdsourcing data when compared to our gold standard of expert annotations. We close with a general discussion in section 6.

Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 129–135
July 13, 2023 ©2023 Association for Computational Linguistics

2 *Again* and its readings

The natural language phenomenon at the core of all annotation tasks discussed here is the English adverb *again* and its well-documented ambiguity. Consider the example corpus data (1) and (2):

(1) i. [A]ll the plants then must be examined, (token 345)
[...]
ii. and those which are planted in pots, should in the following year's bloom be **again examined** (349)
(FALLOWFIELD-1791-2,34.349, '*Gardening Calendar*')

(2) i. He sat really lost in thought for the first few minutes; (token 565)
[...]
ii. He [Mr. Knightley] hesitated, (618)
iii. got up. (619)
[...]
iv. and he **sat down again**; (633)
(AUSTEN-1815-2,169.633, '*Emma*')

The adverb *again* in (1) has a repetitive reading ('*rep*'): An event of the same kind (*examining plants*) is presupposed. The *again* in (2) has a restitutive/counterdirectional readings ('*res/ct*'), i.e., the *again* here does not presuppose a *sitting-down* event by *Mr. Knightley* but an event in the opposite direction. This presupposition is satisfied in (2-iii) where *[he] got up*. The result state of the *sitting-down* event restores a state that held at a time prior to reference time. Note, that in (2) we could naturally assume that Mr. Knightly must have sat down at some point prior to the reference time for (2-iv). In fact, we can infer as much from the context (2-i) but it is never asserted in the prior contexts. Thus, in the domain of relevant times (as far as available in the context) we don't find the repetitive presupposition satisfied in the context. While the result state is overtly spelled out in (2-iv), this need not always be the case for *res/ct* uses, cf. (3) where *again* – on a decompositional analysis – has access to the result state of its predicate:

(3) a. [T]ake them [the trees] up in the fall of the year, give the roots and heads a pruning, (token 391f)
b. and **plant** them **again** [...] (393)
(COBBETT-1838-2,156.393, '*English Gardener*')

These two main readings, *rep* (1) and *res/ct* (2)-(3), are the most frequent ones in the data discussed here and in line with the literature (cf. Gergel and Beck, 2015). A third relevant reading of *again* are discourse-marker uses, which have a discourse organizing function rather than operating on predicates ('*dm*'). Other smaller readings of again exist in the historical data but are not reported here for the sake of brevity (labeled '*other*' in the discussion below).

3 Expert annotation of *again* and its various readings in PPCMBE2

3.1 Method

Based on presupposition (PSP) satisfaction in the linguistic context, our multi-annotator team (i) classified any use of *again* according to its reading, (ii) marked the main verb of the *again*-predicate ('target verb'), and (iii) marked the main verb of the antecedent satisfying a relevant PSP. Other categories were marked in absence of a verb (e.g., *Rain again [...]* cf. RUSKIN-1882-2,3,1019.286). Contextual material was still marked as antecedent – and additionally labeled with an 'inference'-tag – if it 'only' allowed the inference of a relevant PSP but did not constitute a perfect antecedent in a narrow sense. Early stages of the annotation process were marked by iterative cycles of ongoing annotation work informing our annotation guidelines and vice versa. In later stages, our annotators worked on the basis of a detailed multi-page set of annotation guidelines. A crucial point, on a macro level, was to have a robust set of rules to yield uniform decisions for known uses of *again* and to allow for sensitivity for unknown/deviant uses of *again* while remaining general enough to capture the various types of predicates *again* can operate on. On a micro level, our annotation guidelines needed to be able to handle the intricacies in the linguistic representation of event structure not only of *again* events but especially the interaction with (competing) potential antecedent events. Every single use of *again* received (at least) two independent annotations by trained annotators. Disagreements after the first round of annotations were cleared up by repeated reviews and finally consolidated by either a third annotator or by a team consensus.

3.2 Results

To illustrate: Based on our expert annotations, we get the diachronic picture in Table 1 and Figure 1 for Late ModEng (L1-L6), i.e., the PPCMBE2 corpus. These two simplified graphs represent the entire set of 1,901 uses of *again* from the period and show the relative frequency of the two major readings 'repetitive' (*rep*) and 'restitutive/counterdirectional' (*res/ct*), as well as discourse marker uses (*dm*), and the above mentioned fourth class (*other*) (containing minor other readings and low-frequency occurrences of unresolvable ambiguity/unclear cases). In particular, the overall decrease of *res/ct* readings clarifies and certifies previous accounts on the diachronic development of *again* w.r.t. its two major readings (Beck et al., 2009; Gergel and Beck, 2015), which had been done on disparate corpora (i) (solely) based on correspondence and (ii) lacking the 18th century (currently the most general unified corpus is used, from which Tab. 1 is an example).

subperiod	*rep*	*res/ct*	*dm*	*other*
L1, 1700-1734	50.6	42.7	4.5	2.2
L2, 1735-1769	51.2	43.1	2.4	3.4
L3, 1770-1804	59.7	33.1	5.3	2.0
L4, 1805-1839	58.0	33.9	5.3	2.8
L5, 1840-1874	64.1	24.7	10.3	0.8
L6, 1875-1910	60.7	25.0	12.6	1.7

Table 1: Frequency of readings over time in %

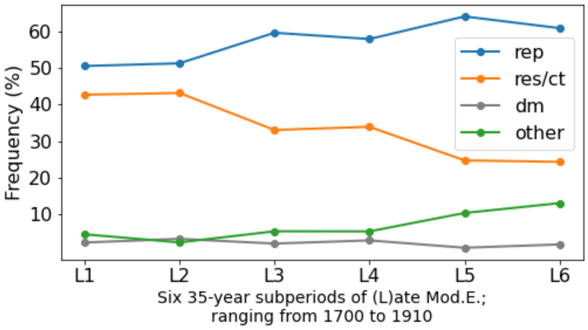

Figure 1: Frequency of readings over time in %

4 Classifying *again*s with a Multinomial Naïve Bayes classifier

4.1 Methods

Based on the expert annotations introduced in section 3 together with a variety of features, we carried out a meta-analysis to find the most promising features in predicting readings of *again* with a Naïve Bayes classifier. We reduced our data set of 1,901 annotations to the 1,722 uses that represent either *rep* (64.4%) or *res/ct* (35.6%) uses of *again*. For these 1,722 *again*s, we collected 16 different features of three major distinct types: (i) "Naïve" features that can be drawn from the linear surface of the text material, (ii) annotational features as per our semantic annotation (but crucially not including the classes of readings, i.e. the dependent variable), and (iii) structural features rooted in the pre-existing syntactic parsing of the data. These features we modeled as count vectors in separate feature matrices for which we computed all possible feature combinations. Over each of the resulting 65,535 different combinations of features, we ran 10 train-test-cycles of a Multinomial Naïve Bayes classifier (with a repeated and randomized 4:1 split between training and testing data for validation) as pretests and 100 train-test-cycles if the pretest gave an accuracy above 77.5%[1]. (Pedregosa et al., 2011; Pustejovsky and Stubbs, 2012)

4.2 Results

We achieve an average accuracy of up to 81.46% in classifying uses of *again* as either *rep* or *res/ct* (based on 100 cycles, standard deviation=2.18%). A set of core features is involved in most feature combinations that achieve average accuracies of 81% or higher: 1. antecedent verb, 2. target verb, 3. distance between antecedent material and *again*, 4. distance between *again* and target verb (also encodes precedence by including negative values), 5. word forms/unigrams in the *again*-clause (as delimited in the syntactic parse). For the average accuracy to go beyond 81% varying other features – often to the exclusion of one another – need to be included. The average accuracy of only the listed features (1.-5.) combined is 80.67% (based on 100 train-test cycles, std.=2.13%). Fig. 2 shows the average accuracies by the number of features. What this also shows is that an abundance of features seems to stunt the classifier and, while improving accuracy overall, also put a cap on it. For the 43 different feature combinations that achieve 81% or higher (purple line in Figs. 2 and 3), the average number of features is 8.58. Another important observation: If we remove all annotational features (especially those pertaining to antecedent material) and rely only on e.g. 3 features that can be gleaned from this corpus data with relative ease (from the preexisting part-of-speech and syntactic annotations): 1. target verb, 2. distance between target verb and *again*, and 3. the object language items

[1] The pretesting was necessary as a measure to reduce computational load. The threshold of 77.5% was informed by previous (shorter) runs in an attempt to strike a balance between expected computational load and desired robustness in the upper range of obtained average accuracies.

in the *again*-clause – with each having a single-feature accuracy of 73.7%, 63.2%, and 74.6%, respectively, – we get an average accuracy of 78.3% (std. 1,93% over 100 train-test cycles). The reported accuracies can be considered a promising first result and, especially since the classifier we used here is insensitive to order (e.g., word order) or weight of features, a result that might be improved upon, e.g. by expanding to *again*-clause bigrams.

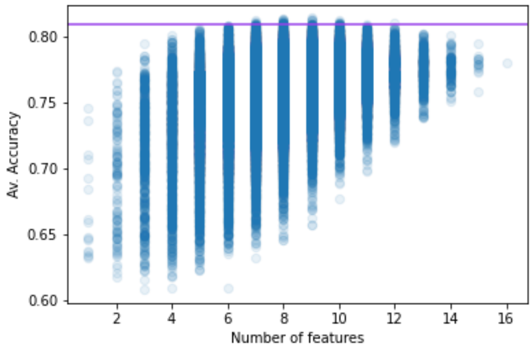

Figure 2: Average accuracy by number of features for 65,535 feature combinations; based on 10 or 100 train-test cycles respectively

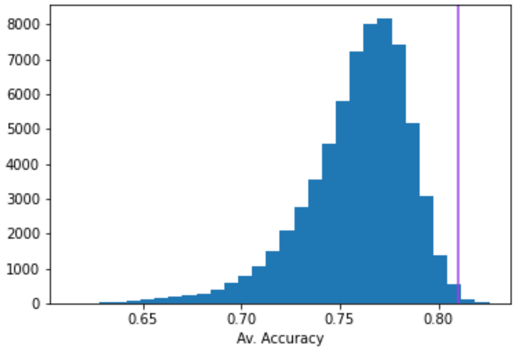

Figure 3: Distribution average accuracy for 65,535 feature combinations; based on 10 or 100 train-test cycles respectively

5 Informed crowdsourcing pilot

5.1 Methods

For this approach we recruited students as crowd workers from two consecutive lectures at the English department at Saarland University. The motivation for this course of action was owing to the intricate nature of the annotation task, i.e., heavily context-dependent semantic annotations on historical language data (with potential antecedent material at varying distances to the PSP trigger – at times significantly greater than, for instance, pronoun reference resolution tasks). Therefore, we needed to be able to communicate with our crowd members in order to quickly respond to uncertainties. We characterize the students who participated as ‘informed crowd’ because, on the one hand, they were not mere speakers of English providing intuitions but, on the other hand, they were not fully-trained as expert annotators. As students enrolled in an English program, our workers’ depths of formal commitment to linguistics is varied: To a large degree, their backgrounds include teachers in training, which means that English is one out of at least two subjects. In other cases, their English studies include a strong emphasis on literary and cultural studies. In next to none of the cases were the student crowd workers formally trained experts. Judging from participants’ place of birth – 83.6% out of the 128 participants who submitted annotations for this pilot study were born in Germany – they are overwhelmingly native speakers of German. In order to generate a return of investment for our students/crowd workers’ contributions, the lectures were drafted so that the crowdsourcing experiment would complement the lectures well. The first was a history-of-English lecture, the second a contrasting-grammars lecture. Both lectures featured a discussion of the diachrony and the semantics of *again* along with an exploration of the guiding research questions and, thus, a connection to the ongoing annotations tasks. Our crowd workers were given a heavily stripped and condensed version of our annotation guidelines, a practice data set, regular tutorial sessions and a recorded tutorial (i.e. a ‘how-to video’). We distributed individualized data sets, each containing five uses of again on a weekly basis directly to students’ inboxes (to minimize the possibility for teamwork). To avoid scarcity in the crowd-provided annotations, we only used a subset of the PPCMBE and the PPCEME (Kroch et al., 2004) data, i.e., 328 *again*s. Submissions were handled with the assignment functionality of our home institution’s online learning platform. Each student had to perform and submit a minimum of three sets of annotations over the course of a semester as part of their minimum grading requirement. An important note here is that submissions were graded exclusively based on formal criteria of the annotation scheme and not on any notion of ‘correctness/incorrectness’ of annotations as such (e.g., relative to a gold standard or the rest of the crowd). After the elicitation phase which yielded 3,319 valid annotations, we prepared the crowd-provided data for analysis by vectorizing the crowdsourced annotations. For a toy example of this conversion, consider Table 2 (pre-) and Table 3 (post-conversion). Moreover, see Table 4 where the sums of the toy data point vectors are combined into the unit vector u1 (along with another toy unit vector u2):

data point	factor	unit	annotator	...
dp1	lev_1	u1	a9	...
dp1	lev_1	u1	a2	...
dp3	lev_2	u1	a4	...
dp4	lev_3	u1	a7	...
...	...	...	...	...

Table 2: Annotations as levels

data point	lev_1	lev_2	lev_3	unit	annotator	...
dp1	1	0	0	u1	a9	...
dp1	1	0	0	u1	a2	...
dp3	0	1	0	u1	a4	...
dp4	0	0	1	u1	a7	...
...	...	...	...	...		...

Table 3: Annotations as one-hot vectors

data points	lev_1	lev_2	lev_3	unit	...
dp1 – dp4	2	1	1	u1	...
dp5 – dp9	1	2	2	u2	...
...	...	...	...	...	...

Table 4: Unit vectors as total of one-hot vectors

5.2 Results

We tested three different approaches for eliciting a 'crowd winner' and evaluating the crowd annotations in contrast to our gold standard provided by our team of expert annotators. The first was a simple majority vote approach[2] – with lev_1 coming out as the winner for unit u1 in the toy example in Tab. 4. In the second approach, we adjusted the bare data point vectors by crowd quality metrics ("CrowdTruth"; cf. Aroyo and Welty, 2013a,b, 2015; Dumitrache et al., 2018). Similar to simple majority vote, the highest value for a unit vector yielded the 'crowd winner'. The third approach was also based on crowd quality adjusted annotation vectors but relied on a KMeans algorithm for unsupervised classification of unit vectors (Pedregosa et al., 2011). We chose the number of clusters ('K') with the 'within-cluster-sum-of-squares' heuristic (WCSS, 'elbow method'; cf. Fig. 4).

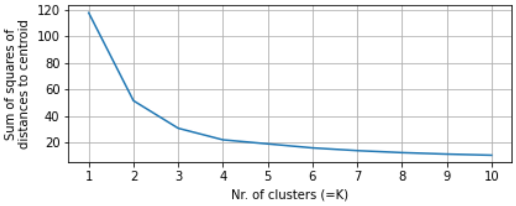

Figure 4: Within Cluster Variation by Ks

Out of the three different approaches, KMeans clustering proved to yield the highest accuracy rates. The detailed results are given in Table 5 where the rows show the gold-standard based readings (*other* were excluded in this pilot). The absolute numbers ('N') represent the number of *again*s available respectively per class and/or period. The corresponding percentages report the accuracies of the KMeans clustering. In addition to per-period, per-century, and overall accuracies, we report Cohen's Kappa in the bottom row. We get high accuracies for the repetitive readings ('*rep*') consistently throughout all periods. The lowest percentage accuracy we get for the restitutive/counterdirectional *again*s ('*res/ct*') – especially in the older data (75.0%). It is predominantly the *res/ct*-reading that is responsible for a decreased overall accuracy of older data.

	17th c.		18th c.		19th c.		all	
	N	%	N	%	N	%	N	%
rep	51	94.1	56	87.5	69	88.4	176	89.8
res/ct	56	75.0	36	80.6	29	89.7	121	80.2
dm	1	100.0	8	87.5	11	90.9	20	90.0
all	112	81.2	102	83.8	114	87.3	328	84.1
C's κ	112	0.65	102	0.7	114	0.73	328	0.7

Table 5: GS units (N) & CS-acc. (%), KMnCl.

Table 6 reports a confusion matrix and shows where the crowd inaccuracies lie. For instance, while Tab. 5 shows that 80.2% out of 121 *res/ct again*s were correctly identified as such (by the crowd and KMeans clustering), Tab. 6 reports on the comple-

[2] In order to avoid ties (u2 in Tab. 4), all data point vectors were adjusted for meta-features of the respective data point:

- experience$_{dp}$ stands for the experience the worker had when providing the data point at hand (ranging from 0 to 11),
- average evaluation$_{dp}$ stands for the average evaluation (i.e. the point system for grading purposes) a student received for the submission of the data set the data point originates from (from 0.0 to 1.0),
- semester progress$_{dp}$ stands for how far into the semester (i.e. ordinal number of weekly data roll-outs) the data point was produced (from 1 to 12), and
- motivation$_{dp}$ gives the total number of data sets the worker submitted who provided the data point at hand (from 2 to 12).

The features were ranked based on our intuition for respective relevance and scaled to such small weights that they could not tip the scale over the number of available crowd votes:

$$\begin{array}{lcl} (1 + (10^{-3} & * & \texttt{experience}_{dp})) * \\ (1 + (10^{-6} & * & \texttt{average evaluation}_{dp})) * \\ (1 + (10^{-9} & * & \texttt{semester progress}_{dp})) * \\ (1 + (10^{-12} & * & \texttt{motivation}_{dp})) \\ & = & \texttt{tie breaker}_{dp} \end{array}$$

mentary 19.8% inaccurate cases. 23 of these were classified as *repetitive* and only one as *discourse marker* (‘*dm*’). The ratio of true to false hits for the two main readings (*rep* vs. *res/ct*) is 9.3:1 for the *rep*-data (gold standard) and 4.2:1 for the *res/ct* data. Thus, if the goal is to reduce costly workload for expert annotators, a review of crowdsourced annotations ought to focus on the data that comes out as *res/ct* since it is here that we find a higher confusion rate (97:17 in contrast to 158:25, true to false positives, respectively).

	CS-*rep*	CS-*res/ct*	CS-*dm*
GS-*rep*	158	17	1
GS-*res/ct*	23	97	1
GS-*dm*	2	0	18

Table 6: Confusion matrix, crowd sourcing by gold std.

The strategy to focus on *res/ct* data for an expert review of crowd sourced data is also supported by the distribution of unit quality scores. Unit quality scores (UQS) are computed for each unit (= use of *again*): We calculated it as the average of all pairwise cosine similarities for all possible distinct worker$_i$ and worker$_j$ pairings (such that worker$_i$ $\neq$ worker$_j$) (Aroyo and Welty, 2013a,b, 2015; Dumitrache et al., 2018). Interpreting the UQS as a measure of crowd confidence, we conclude that the crowd decisions for true *rep*-readings came about with higher confidence than the true *res/ct*-readings, cf. top-left vs. bottom-left subplots in Fig. 5. Thus, focusing on the crowd-provided *res/ct*-labels in a review by expert annotators would also increase robustness of the annotated data in the ‘right places’.

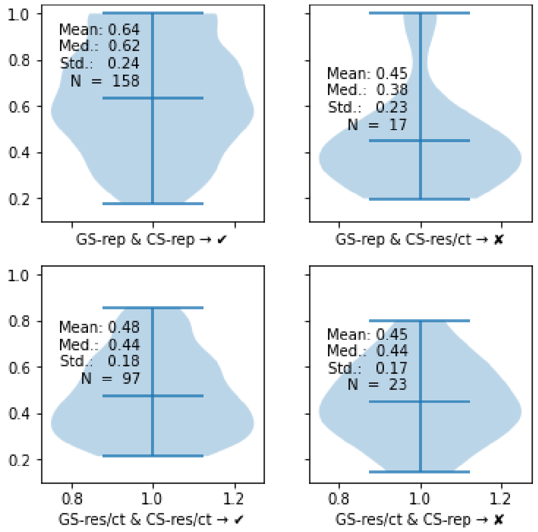

Figure 5: Unit Quality Score (UQS) for GS-CS matches & mismatches; as kernel density plots

6 Conclusion

At the current state of the technical possibilities explored and as far as the natural language phenomenon at hand is concerned, a gold standard cannot be substituted wholesale by either machine learning-based predictions or experimental data. The first upshot is that the gold standard itself must be as solid as possible (we sketched our detailed approach above, and we are open to constantly improving it). At the same time, we think that our two additional case studies are quite telling even if their performance was expectedly lower. The significance of such extensions is obvious when it comes to the annotation of larger amounts of data (be it for decompositional markers or other annotational tasks; of course, for low-frequency phenomena, the use of larger corpora or alternative methods becomes a necessity). The feature-based approach (section 4) then becomes relevant, also for cases in which the syntactic annotation is missing such as the EEBO type of corpora in our object-language English. In such a case, some of the syntactic features we have used in our approximations can be translated, e.g., in terms of precedence (an instance of again that precedes its modifying predicate is typically also higher in structure etc.). Overall, however, we believe that the human approach, i.e., the type of informed crowdsourcing we have utilized, is the most promising variant of annotational support when one strives to cover more data than one’s team can handle or for gaining more certainty empirically. The straightforward advantage is that the relatedness in the languages at hand can be used even if the ‘nativeness’ of the actual participants is not available. Some of our results have indicated that more distant periods in time do not necessarily become worse in the annotational performance. On a conceptual level, there is also initial evidence from independent areas of semantic change (cf. Gergel et al., 2021, 2023) that speakers adapt astonishingly well in simulated situations of change. Finally, even if certain targeted readings are comparatively low performing, one can still place a crowdsourcing approach at the start of an annotation pipeline. By validating crowd annotations with a gold standard for a subset of the data, one can learn which data (i) needs a closer review, (ii) which data needs less attention in a review, and (iii) which data could benefit from a thorough review due to inherent indecisiveness of the crowd.

Acknowledgements

We thank our expert annotators – especially Mohammad Bablli, Maryam Rajestari and Dana Rebar – and our students for their participation in this work, which has been funded by the German Research Foundation through the project Decomposing Decomposition in Time awarded to Remus Gergel.

References

Lora Aroyo and Chris Welty. 2013a. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *Web Science 2013*, New York: Association for Computing Machinery.

Lora Aroyo and Chris Welty. 2013b. Measuring crowd truth for medical relation extraction. In *Semantics for Big Data: Papers from the AAAI Fall Symposium*, AAAI Technical Report FS-13-04, Palo Alto, CA.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Josep Ausensi, Jianrong Yu, and Ryan Walter Smith. 2021. Agent entailments and the division of labor between functional structure and roots. *Glossa: A journal of General Linguistics*, 6(1):53.

Sigrid Beck. 2005. There and back again: A semantic analysis. *Journal of Semantics*, 22:3–51.

Sigrid Beck, Polina Berezovskaya, and Katja Pflugfelder. 2009. The use of *again* in 19th-century English versus Present-Day English. *Syntax*, 12(3):193–214.

Sigrid Beck and Remus Gergel. 2015. The diachronic semantics of English *again*. *Natural Language Semantics*, 23(3):157–203.

Marco Degano and Maria Aloni. 2022. Indefinite and free choice: When the past matters. *Natural Language and Linguistic Theory*, 40:447–484.

Anca Dumitrache, Inel Oana, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement.

Remus Gergel and Sigrid Beck. 2015. Early Modern English *again*: a corpus study and semantic analysis. *English Language and Linguistics*, 19(1):27–47.

Remus Gergel, Martin Kopf, and Maike Puhl. 2021. Simulating semantic change: a methodological note. In *Proceedings of Experiments in Linguistic Meaning (ELM)*, pages 184–196, University of Pennsylvania: LSA.

Remus Gergel, Maike Puhl, Simon Dampfhofer, and Edgar Onea. 2023. The rise and particularly fall of presuppositions: Evidence from duality in universals. In *Proceedings of Proceedings of Experiments in Linguistic Meaning (ELM) 2*, pages 72–82, University of Pennsylvania: LSA.

Anthony Kroch, Beatrice Santorini, and Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*, first edition. Department of Linguistics, University of Pennsylvania. Release 3.

Anthony Kroch, Beatrice Santorini, and Ariel Diertani. 2016. *The Penn Parsed Corpus of Modern British English (PPCMBE2)*, second edition. Department of Linguistics, University of Pennsylvania. Release 1.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly, Sebastopol, CA.

Irene Rapp and Arnim von Stechow. 1999. *Fast* 'almost' and the visibility parameter for functional adverbs. *Journal of Semantics*, 16:149–204.

Joost Zwarts. 2019. From 'back' to 'again' in Dutch: The structure of the 're' domain. *Journal of Semantics*, 36:211–240.