Informed crowdsourcing for annotating semantic variation and change

Martin Kopf, Remus Gergel

Saarland University, V 0.1, September, 2022

Abstract

In this paper we explore crowdsourcing as a novel approach to generating semantic annotations on diachronic corpus data. We will contrast annotations performed by an informed crowd with an expert-annotated gold standard. In doing so, we will compare a number of different approaches to yielding a crowdsourcing winner, ranging from simple majority vote to KMeans clustering. At the same time we explore how 'quality scores' for the annotations as well as workers can improve results, especially for unsupervised clustering. We report an overall accuracy of 84.1% with a Cohen's κ=0.7 stressing that the added value we see in our approach lies in an extension of current techniques of crowdsourcing to diachronic concerns.

1 Introduction

We present and discuss data originating from an 'informed crowdsourcing' experiment geared towards creating semantic annotations for the presupposition trigger and decompositional adverb *again*. These semantic annotations are part of a wider effort to provide an exhaustive annotation of *again* (and other decompositional items) as an add-on layer to the Penn Parsed Corpora for Historical English.[1] In providing such data we hope to fill empirical gaps and continue efforts by Beck et al. (2009); Beck and Gergel (2015); Gergel and Beck (2015); Gergel and Stateva (2014); Gergel et al. (2016); Gergel and Nickles (2019) on strengthening the connection between linguistic theory and reality of attested patterns of potential decomposition and especially historical ones. The motivation behind creating an exhaustive semantic annotation of decompositional items on top of an existent, state-of-the-art syntactic annotation is to increase the detail of the empirical basis required for tracking language chance at the syntax-semantics interface (Eckardt, 2006; Beck and Gergel, 2015; Deo, 2015).

While we remain committed to providing state-of-the-art expert annotations with our team of trained annotators, the amount of resources required for producing a reliable gold standard is a considerable motivator for exploring other avenues for creating semantic annotations for diachronic data. We are thus investigating whether the potential of crowdsourcing can be harnessed to reduce the resources required for creating semantic annotations for diachronic data. In this pilot, we evaluate annotations sourced from an informed crowd on the basis of our own gold standard annotations. In order to facilitate this comparison (i.e. calculate accuracies), we rely on three different modes for coercing a crowd decision: simple majority vote, quality-scores-adjusted votes, and KMeans clustering (Pedregosa et al., 2011). Section 2 will briefly discuss English *again* and its ambiguity as the natural language phenomenon presented to our crowd workers, the corpus data the crowd workers had to cover, and the intricacies of the annotation task to be performed by the crowd. Section 3 will introduce

[1]

YCOE	Old English	1.5M	(Taylor et al., 2003)
PPCME2	Middle English	1.2M	(Kroch et al., 2000)
PPCEME	Early Mod. E.	1.7M	(Kroch et al., 2004)
PPCMBE2	Late Mod. E.	2.8M	(Kroch et al., 2016)

the recruitment of our crowd workers and detail what we mean by 'informed crowd'. We will touch on data distribution and data return, and end on a general overview of the annotations we were able to elicit. In section 4, we will start off with a note on data processing before moving on to assessing the agreement between crowdsourcing (CS) annotations and gold standard (GS) annotations. We will discuss our findings and conclude in section 5.

2 Background

In this section we want to introduce the ambiguity of *again*. We will elaborate on this ambiguity from an annotation point of view before moving on to introduce the diachronic data for this pilot study.

2.1 Again

The English adverb *again* has a well-documented ambiguity. Consider (1):

(1) Leo jumped up again.

(2) a. Leo jumped up, and he had done that before. (*repetitive*)
b. Leo jumped up, and he had been up before. (*restitutive/counterdirectional*)

cf. (Beck and Gergel, 2015; Gergel and Beck, 2015) for a diachronic discussion

The *again* in (1) has at least two readings. On the repetitive reading ('`rep`'), the *again* presupposes that an event of the same kind (as per the *again*-predicate) has occurred prior to reference time, cf. (2-a). On the second reading (2-b), the result state of the *again*-predicate is a restoration of a state that held and was departed from prior to reference time. Thus, an event in the opposite direction is presupposed (i.e. 'restitutive/counterdirectional', '`res_ct`' for short). In a situation where Leo has not jumped up, (1) can only be felicitous with the `res_ct`-interpretation in (2-b). The two readings in (2) are the most frequent ones in the data discussed here. A third reading relevant in the historical examples is the 'counterdirectional-proper' reading ('`ctd`'), which lacks a result state and merely presupposes an event in the opposite direction. The last type of *again* are discourse-related uses ('`other`'), which have a discourse organizing function rather than operating on predicates.

In order to disambiguate between the readings of *again*, hearers depend on context (especially for *again* operating on accomplishments and achievements (Dowty, 1979)). In other words, in order for a hearer to be able to interpret the *again* in (1), they need to have access to the context and whatever information it holds that can contribute to disambiguating among the various readings. This can be information encoded in a 'perfect antecedent' (e.g. *Leo sat down* for (2-b)) or packaged in any other way that allows drawing the inference that (2-b) holds. For simplicity, we will refer those places in the context as antecedents (cf. Delin (1992) and references there for a discussion of the theoretical implications of the notion of antecedent, but our goal here remains empirical). A further-going piece of motivation for our experimental situation is that with the help of context, speakers are able to discern the relevant meanings in related varieties even if their native grammars do not have access to the same intuitions as those of the produced text (Gergel, 2020; Gergel et al., 2021).

2.2 Annotation task

In this section we introduce and discuss the annotation task our crowd of workers had to perform. We pay special attention to the challenges in this pilot study which potentially makes this application of crowdsourcing stand out next to other NLP tasks performed by a crowd. In broad terms, the annotation task was to identify uses of *again* according to the ambiguities introduced above (section 2.1). In more detail, the task description included:

- Classify uses of *again* according to their readings; Use labels `rep`, `res_ct`, `ctd`, and `other`.
- Annotate the place in the context that helped disambiguate between possible readings; to operationalize this: Mark finite verb of clause containing antecedent with a pair of underscores ("_verb_"); if there is no finite verb, pick next

best word, i.e. any one-word item allowing the inference that one presupposition over another, competing presupposition is satisfied in the context.

- Briefly explain reasoning as to why a particular decision was made.
- Indicate whether an antecedent was found; 'yes'/'no'.
- We accounted for ambiguities by asking workers to separate the relevant labels with a comma (e.g. "`rep`, `ctd`").

The crowd workers were provided with a one-page sheet of annotation guidelines describing the relevant readings of the *again* and the above bullet points included in their task (which stands in stark contrast to the multi-page document that was created to establish an expert-annotated gold standard). The goal was to keep the effort and time spent on preparing for the core task to a minimum as crowd workers cannot be assumed to absorb lengthy manuals (cf. Aroyo and Welty, 2013b). Notice the apparent redundancy in marking antecedents and, additionally, noting whether an antecedent was found. Our motivation here was to elicit definitive and conscious responses for the entire width of the spectrum of contextual evidence (rather than having to guess whether a worker forgot to mark an antecedent, whether they did not find any, or they did not look (far enough) in the first place). Another reason to include a binary response as to the availability of an antecedent and its marking in the context was that the distance between *again* and its antecedent is potentially unbounded [2].

2.3 Diachronic data

Data for this crowdsourcing pilot were sourced from the Penn Parsed Corpus of Early Modern English ('PPCEME') (Kroch et al., 2004) and the Penn Parsed Corpus of Modern British English ('PPCMBE') (Kroch et al., 2016) (yielding 1,536 and 1,901 uses of *again* in total respectively). Out of these two corpora, we picked corpus texts based on the following criteria: (i) For each of the 17^{th}, 18^{th}, and 19^{th} centuries we selected about 100(+) uses of *again*. (ii) We prioritized the most abundant corpus texts (w.r.t. absolute frequency of *again*). (iii) We excluded biblical corpus texts. Re. (ii), we tried to keep worker fatigue to a minimum and, thus, wanted to avoid workers having to familiarize themselves with new texts over the course of the study too many times. Re. (iii), biblical texts tend to be more conservative and we wanted our crowd workers to face data representative of its period(s). This resulted in:

17^{th} cent:	112 *again*s
18^{th} cent:	102 *again*s
19^{th} cent:	114 *again*s
	328 *again*s

Gold standard: All 328 *again*s were annotated by our team of expert annotators (two annotations minimum with consecutive reviews of disagreements).

To sum up, there are a number of factors that make this pilot study stand out: 1. The classification task based on presupposition satisfaction often required our workers to read long stretches of context in order to complete their task. 2. We served diachronic language data rather than present-day language data to the crowd workers and as a consequence: 3. On the one hand, not all the relevant readings are covered by the grammars of native speakers of present-day English. On the other hand – and more importantly – not all the relevant readings are covered by the grammars of native speakers of present-day German (in the adverb *wieder*, with a similar `rep`/`res_ct` ambiguity), who made up the majority of our crowd workers.

3 Informed crowdsourcing

In this section, we will introduce the specifics of our 'informed crowdsourcing' approach. We will start off with crowd recruitment with special attention to the characteristics of our workers. We then continue with

[2] For creating the gold standard, on occasion external sources needed to be consulted on e.g. whether a certain army had invaded a country prior to reference time (`rep`) or 'only' left said country prior to reference time (`res_ct`).

details on data handling, and data set design before closing with a brief outline of the elicited data.

3.1 Crowd recruitment

The crowdsourced annotations were collected over the course of two semesters at the English Department at Saarland University (SU). The crowd workers were recruited as participants of two lectures (a history-of-English lecture in winter term 2021/22 and a contrasting-grammars lecture in summer term 2022). Some students participated in both lectures/semesters. In the context of the lectures, the annotations were referred to as 'empirical tasks' (in contrast to the 'summary tasks' geared towards the respective lecture's contents). 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. At the start of each term, an introductory session laid out the basic plan for the semester ahead along with a brief introduction to the annotation component.

We characterize our crowd workers as 'informed crowd' because, on the one hand, the workers 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 crowd workers formally trained experts. Judging from participants' place of birth – 83.6%[3] were born in Germany – they are overwhelmingly native speakers of German.

As far as training and preparation are concerned, in addition to the annotation guidelines, we offered a weekly tutorial dedicated to the annotation/empirical tasks. For both semesters of this tutorial, we did a 'practice round' of annotations on a curated set of data before we sent out proper data sets. In response to the results of the practice data, we provided another one-page sheet with generalized feedback. From there on out, in the context of the weekly tutorial meetings, we offered synchronous guided annotation sessions based on the practice data.

[3] That is, out of the 128 participants who submitted annotations for this pilot study

3.2 Data distribution and collection

Data sets were rolled out on a weekly basis to all students registered for the lecture(s). As a means of distribution of the single personalized data sets, we chose email. The goal was to keep the possibility of cooperation and coordination among peers to a minimum. Submission was handled via Microsoft Teams (central component of a MS software suite Saarland University is relying on for its digital environment). Only those submissions that matched the allocated data sets were accepted. Grading was based on formal criteria of the annotations, i.e. the degree of detail and consistency to which workers followed the annotation guidelines (grading was not based on the 'correctness' of the annotations/decisions involved in the annotations in any form).

3.3 Data sets

A single personalized data set included five uses of *again* pseudo-randomly picked from our larger pool of data. For each student(/crowd worker), a continuous record of previously assigned uses of *again* – identified with unit-IDs – progressively informed and limited the choice of data to be drawn from for the remainder of the two semesters. Weekly data sets were compiled as Excel files (with one use of *again* per row). Aside from various meta- and corpus-related information, the Excel table had (empty) columns for the required annotation. In addition to a column with the sentences containing the respective *again*s, there was a column labeled 'context' for each use of *again*, containing the ten sentences (='corpus tokens') preceding the *again*-sentence. The corpus texts that were used for this crowdsourcing project were accessible to all students in a (download-only) folder on MS Teams. The students were asked to rely on those files, should the amount of context provided in the Excel files not suffice and to copy and paste the rele-

vant antecedent sentence(s) into the Excel file (with the corresponding IDs) in order to perform the annotations.

3.4 Elicited data

We received 3,319 valid annotations (i.e. one of the above labels or a comma-separated combination thereof) from 128 different workers.[4] The diachronic distribution of these 3,319 data points is as follows:

– 17th cent:	1,086
– 18th cent:	969
– 19th cent:	1,264
	3,319 data points

4 Crowd vs. gold standard

We will start this section off with a general discussion of data processing and a brief overview of the crowdsourced annotation data. We then move on to introduce three approaches to deciding on winners among potentially divergent crowd annotations. For each approach we will provide the observed accuracy and, where applicable, Cohen's Kappa for inter-annotator agreement by subperiods, GS-readings and overall (Cohen, 1960).[5]

[4] If we split combined labels (due to perceived ambiguities) into separate labels (exclusively either `rep`, `res_ct`, `ctd`, or `other`), we end up with 3,425 data points. This approach allows for (i) slightly higher accuracy and agreement ratings (cf. section 4), and (ii) a more immediate vector representation of ambiguity if we consider the four basic classes as definitive dimensions in a vectors space (cf. 4.1). However, in a use case such as tracking semantic change (which is the ultimate goal here), ambiguities in diachronic data – based on linguistic evidence – need to be able to come out as the e.g. "winning annotation/final decision" in a majority vote rather than as diverging dimensions of identical magnitude.

[5] Cohen's Kappa is a statistic for agreement between raters that accounts for chance agreement. The maximum value – indicating complete agreement – is 1.0 while $\kappa = 0$ indicates no agreement other than chance agreement. As far the ranges of Kappa achieved in this study are concerned, $0.68 \leq \kappa < 0.8$ allow "tentative conclusions", while $\kappa > 0.8$ indicates good reliability (Poesio and Vieira, 1998). Cohen suggests 0.61–0,8 as a range for "substantial" agreement (McHugh, 2012). In a (2012) discussion contrasting the percentage and Kappa statistics, McHugh concludes that while Kappa is useful, as it accounts for guessing, one might favor the percentage statistic in contexts with relatively well-trained annotators, and suggests: When in doubt, provide both statistics. In our study, we use the κ-statistic to compare a gold standard to crowd annotations. For a gold standard – while not infallible – guessing is not an option. As far as our crowd workers are concerned, they did receive training and did not act as naïve native speakers. Thus, we are inclined to favor observed accuracy in percentages over the κ. Moreover, since κ is calculated globally, i.e. over all labels, and we are also interested in respective accuracies of our labels, we have to rely on the percentage statistic.

4.1 Data processing

For all approaches, the point of departure in terms of data processing constituted turning the unique levels of the factor (crowd-worker-provided) '`reading`' into vector dimensions with a one-hot encoding on the worker provided label. For a toy example of this conversion, consider Table 1 (pre-) and Table 2 (post-conversion):[6]

d.p.	`factor`	unit	...
dp_1	`lev_1`	u1	...
dp_2	`lev_1`	u1	...
dp_3	`lev_2`	u1	...
dp_4	`lev_3`	u1	...
...	...	...	...

Table 1: Annotations as levels

d.p.	`lev_1`	`lev_2`	`lev_3`	unit	...
dp_1	1	0	0	u1	...
dp_2	1	0	0	u1	...
dp_3	0	1	0	u1	...
dp_4	0	0	1	u1	...
...	...	...	...	...	...

Table 2: Annotations as one-hot vectors

We will refer to the one-hot encoded data in the rows in Table 2 as 'data point vectors' (*dp_vec* for short) and to vectors that combine all available data point vectors for one use of *again* as 'unit vectors' (*u_vec* for short). Consider Table 3 for a toy example that combines the sums of data point vectors from above into the unit vector u1 (along with another toy unit vector u2):

[6] 'd.p.' is short for 'data point'

(d.p.)	lev_1	lev_2	lev_3	unit	...
(1–4)	2	1	1	u1	...
(5–8)	1	2	1	u2	...
...	...	...	...	...	...

Table 3: Unit vectors as total of 1-hot vec.s

Turning back to our crowd data: For visualization, we can fit our raw unit vectors into a principal component analysis (PCA) (Pedregosa et al., 2011), see Fig. 1 where the bottom graph shows the principal components (PC) 1 and 2 (on axes x and y respectively) and the top shows PC 3 on the vertical, as-it-were z-axis (along with PC 1 on x). Note that these three principal components account for 96.4% of the variance in the annotation data. The hues in Fig. 1 correspond to gold standard readings. Only the main readings appear in the legend. However, ambiguous data are represented in the chart (shaded in gray).

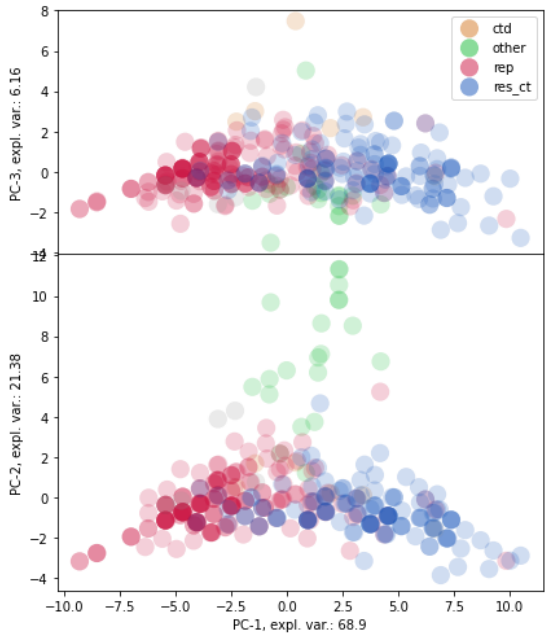

Figure 1: PCs1–3, 'raw' unit vectors

For further visualization and to show overall accuracy, we can calculate cosine similarities between each crowdsourced unit vector and the corresponding gold standard annotation.[7] Thus, we arrive at the distribution of cosine similarities in Fig. 2. The mean cosine similarity over all annotations is 0.84 and the median is at 0.94.[8]

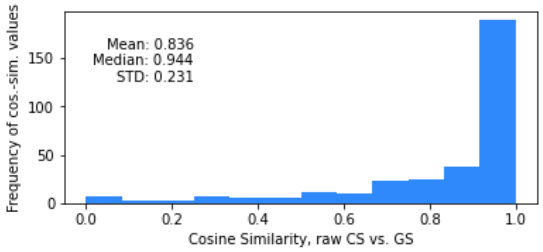

Figure 2: Distribution of cosine similarities, CS vs. GS (i.e. unit vectors vs. one-hot vectors)

4.2 Simple majority vote

Out of 328 different uses of *again*, 32 *again*s received a unanimous vote, 28 of which where were annotated as `rep`, 2 as `res_ct`, and 2 as `other` (by the crowd workers). Out of the remaining 296 *again*s, for 277 a majority was found on these 'bare votes'. For the remaining 19 *again*s, a tie breaker system needed to be established: Every data point vector was adjusted for meta-features of the respective data point; for details:

- $\texttt{experience}_{\text{dp}}$ stands for the experience the worker had when providing the data point at hand (ranging from 0 to 11),
- $\texttt{average evaluation}_{\text{dp}}$ stands for the average evaluation (i.e. the point system for grading pur-

[7] For instance, if we want to compare our toy example vector from Table 3, u1 = (2, 1, 1), with the 'toy gold standard vector', g1 = (1, 0, 0), we get a cosine similarity of 0.82:

$$cos\_sim(u1, g1) = \frac{u1 \cdot g1}{||u1|| * ||g1||} = \frac{\sum_{i=1}^{n} u1_i * g1_i}{\sqrt{\sum_{i=1}^{n} u1_i^2} * \sqrt{\sum_{i=1}^{n} g1_i^2}} = \frac{(2*1)+(1*0)+(1*0)}{\sqrt{2^2+1^2+1^2} * \sqrt{1^2+0^2+0^2}} \approx \frac{2}{2.45} \approx 0.82$$

[8] cf. section 4.3, p. 7 for details on cosine similarity

poses) a student received for the submission of the data set the data point originates from (from 0.0 to 1.0),

- `semester progress`$_{\text{dp}}$ 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`$_{\text{dp}}$ gives the total number of data sets the worker submitted who provided the data point at hand (from 2 to 12).

These features were ranked according to perceived relevance for providing reliable judgments and combined into a tie-breaker$_{\text{dp}}$ score:

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

For the results of this majority vote system, see Table 4, where we present observed accuracies[9] of the CS-winners relative to the gold standard (GS) (in %; along with the respective number of *again*s) by GS-label, by century, and overall – along with by Cohen's Kappa (bottom row).[10]

	17th c.		18th c.		19th c.		all	
	N	%	N	%	N	%	N	%
`rep`	51	94.1	56	89.3	69	92.8	176	92.0
`res_ct`	56	67.0	36	77.8	29	82.8	121	74.0
`other`	1	100.0	8	87.5	11	81.8	20	85.0
all	112	78.1	102	83.8	114	86.8	328	82.9
C's κ	112	0.6	102	0.7	114	0.72	328	0.68

Table 4: GS units (N) & CS-acc. (%), Maj.v.

Overall accuracies and Cohen's Kappa is consistently lower for historically older data, which matches our intuitions since the grammars that generated the 17th-cent. data are expected to be more alien to present-day speakers of (L2) English. What's striking in Table 4 is that *again*$_{\texttt{rep}}$ comes out with consistently high accuracies across all periods. It is the `res_ct` reading that is impacting overall per-century accuracies. With 67% in the 1600s, accuracy is reduced to roughly the half-way point between chance and the overall 82.9%.

4.3 Quality metrics

Drawing on the "CrowdTruth" approach proposed by Aroyo and Welty in (2013a; 2013b; 2015), we adjusted our vectors based on so-called unit quality scores (UQS) and worker quality scores (WQS), the latter being the product of worker-worker-agreement (WWA) and worker-unit-agreement (WUA). While Dumitrache et al. (2018) compute these scores iteratively (until convergence, i.e. until a minimum variation between iterations is achieved), we discuss a more linear approach here.

What all quality metrics discussed here have in common is that agreement is conceived of as the (positive range of) cosine similarity between vectors (cf. footnote 7, p. 6 for details).

4.3.1 Unit quality score (UQS)

Unit quality score is computed for each unit (= use of *again*). We calculated it as the average of all pairwise cosine similarities for every worker i and all other workers j that worked on this unit, s.t. $i \neq j$. In other words, for each use of *again* we are getting the average cosine similarity (*cos_sim*) for all possible worker$_i$ and worker$_j$ pairings:

$$UQS(u) = \frac{1}{n_i n_j} \sum_{i,j,i \neq j} ww\_cs(i,j,u), where$$

$$ww\_cs(i,j,u) = cos\_sim(dp\_vec_{i,u}, dp\_vec_{j,u})$$

4.3.2 Worker unit agreement (WUA)

WUA is computed for each worker i. For each unit u that worker i worked on, we have a 'data point

[9] Note, the observed accuracies are calculated accounting for 'partial matches' in ambiguous cases; e.g. if the annotation "`rep`, `res_ct`, `ctd`" is compared to "`rep`", then accuracy for this comparison will come out with a value of $0.\overline{3}$. This is of particular importance for the discerning reader who wants to double-check our accuracy values. Note also that any mention of Cohen's Kappa is not sensitive to ambiguous labels in the same way.

[10] Note, that 11 uses of *again* are missing from the by-class break-down in this table since their gold standard annotations form groups too small to reliably calculate accuracies. They are, however, included in the totals-row.

vector for u by i' ($dp\_vec_{i,u}$ for short). *WUA(i)* is the average cosine similarity between $dp\_vec_{i,u}$ and the relevant unit vector u_vec_u (minus $dp\_vec_{i,u}$). In line with Dumitrache et al. (2018), we weighted this score with the relevant *UQS(u)* (cf. section 4.3.1). The idea here is to not 'punish' workers for the work they did on controversial or difficult uses of *again*.

$$WUA(i) = \frac{\sum_{u \in units(i)} wu\_cs(u,i) * UQS(u)}{\sum_{u \in units(i)} UQS(u)},$$
$$where \;\; wu\_cs(u,i) = cos\_sim(dp\_vec_{i,u}, u\_vec_u - dp\_vec_{i,u})$$

4.3.3 Worker worker agreement (WWA)

WWA is computed for each worker i. Thus, for each i and for each u that i worked on, we get the $dp\_vec_{i,u}$ and calculate the pairwise cosine similarities between it and all the $dp\_vec_{j,u}$ from all other workers that worked on u[11]:

$$WWA(i) = \frac{\sum_{j,u} ww\_cs(i,j,u) * UQS(u)}{\sum_{j,u} UQS(u)},$$
$$\forall u \in units(i), j \in u, i \neq j.$$

The product of WUA and WWA form the worker quality score (WQS):

$$WQS(i) = WUA(i) * WWA(i)$$

Having computed the above quality metrics (i.e. 'CrowdTruth' metrics in Aroyo and Welty's and Dumitrache et al.'s terms), we can adjust the binary data point vectors for both UQS and WQS. Summing these into unit vectors and picking the maximum (i.e. strongest dimension) as the 'crowd-truth' winner, we get the improved accuracies in Table 5:

[11] *units(i)* – the set of all units that i worked on; *worker(u)* – the set of all workers that worked on u.

	17th c.		18th c.		19th c.		all	
	N	%	N	%	N	%	N	%
`rep`	51	94.1	56	91.1	69	92.8	176	92.6
`res_ct`	56	69.6	36	75.0	29	82.8	121	74.4
`other`	1	100.0	8	87.5	11	81.8	20	85.0
all	112	79.5	102	83.8	114	86.8	328	83.4
C's κ	112	0.62	102	0.7	114	0.72	328	0.68

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

While the overall accuracy is marginally improved (by 0.5% to 83.4%; κ=0.68), the most problematic data (`res_ct` in the 1600s) has improved more substantially (from 67.0% to 69.6%).

4.4 KMeans clustering

In this section we will discuss unsupervised classification of our crowdsourcing data based on 'KMeans clustering' (Pedregosa et al., 2011). We again relied on the above quality metrics (section 4.3) derived from the CrowdTruth literature in combination with KMeans clustering. In a first step, we normalized our 328 quality-metrics adjusted unit vectors in order to facilitate a clustering of our data, cf. Fig. 3, where hue corresponds to GS-readings (as before, only the major labels are in the key but other data are present shaded in gray).

KMeans clustering iteratively optimizes the mean distances of all data points to a K-number of 'centroids'. The 'K' is a hyper-parameter to be determined with the 'within-cluster-sum-of-squares' heuristic (WCSS, 'elbow method'): The idea here is to test for the reduction of the sum of squares (of distances to the centroids) as the number of clusters increases, cf. Fig. 4 where – in our case – the gains in reduction of sum of squares (SS) start leveling out with three clusters[12].

Allowing our data points (in Fig. 3) to be sorted into three clusters, we can calculate accuracy values by assuming the most frequently represented (modal) gold standard label in each cluster as the canonical class of the cluster. Thus, we arrive at the observed accuracies and κ-values in Table 6.

[12] The SS is maximal for $K = 1$ and 0 for $K = nr.\ of\ datapoints$

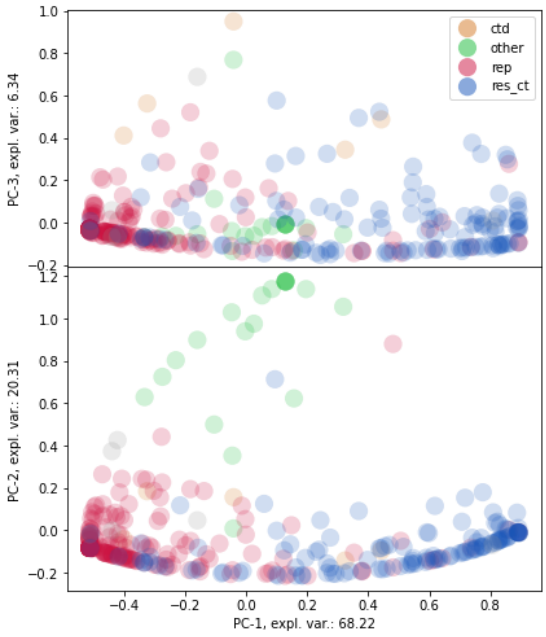

Figure 3: PCs1–3, CT-adj. & norm'd

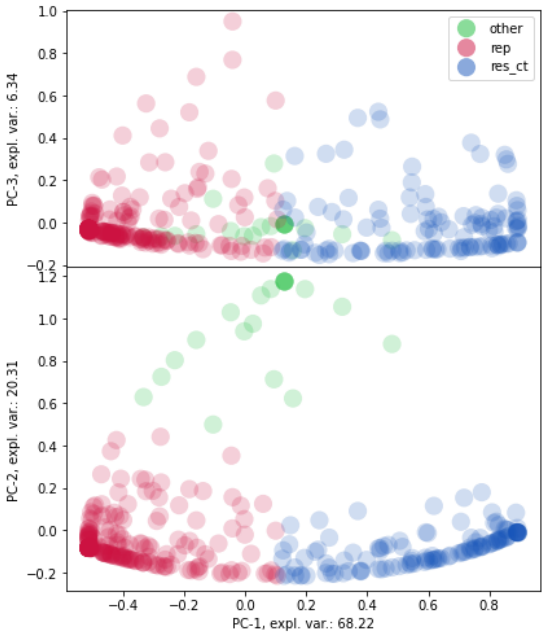

Figure 5: PCs1–3, CT-adj. & norm'd, KMnCl.

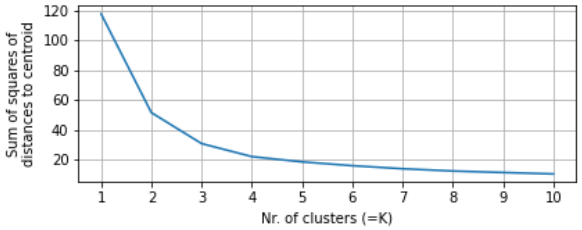

Figure 4: Within Cluster Variation by K's

	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
other	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 6: GS units (N) & CS-acc. (%), KMnCl.

Consider Fig. 5, where the crowdsourced data have been sorted into clusters with the labels in the key corresponding to the respective clusters.

This approach raises the overall accuracy of the crowdsourcing data to 84.1%. For the 'weakest' subset of our data we now have an accuracy of 75.0%. The clusters and these results remain stable over repeated random initializations.[13] This is a first, encouraging result: The crowd workers achieved respectable accuracy with a bare majority vote. Moreover, we were able to harness the inherent disagreement to raise accuracy across all subsets of data by applying a commonly used unsupervised classification approach. Most importantly, these results were achieved by our crowd workers annotating natural language data originating from as far back as Early Modern English – data for which our workers lack native-speaker intuitions.

[13] Along the alternative route mentioned above (splitting annotations with multiple labels into their respective single labels), we can get an overall accuracy of 84.5% (Cohen's κ = 0.7).

5 Discussion

Table 7 shows a confusion matrix for gold standard labels ('gs') by crowdsourcing labels – specifically, the KMeans-clustering labels ('cl') as the most successful means for getting a winning annotation.[14,15]

	(ctd_{cl})	other_{cl}	rep_{cl}	$\text{res\_ct}_{\text{cl}}$
ctd_{gs}	0	0	3	2
other_{gs}	0	18	2	0
rep_{gs}	0	1	158	17
$\text{res\_ct}_{\text{gs}}$	0	1	23	97

Table 7: Confusion matrix KMeans vs GS

The rows add up to the same values as in the 'all' column in Table 6. What this table shows (in absolute numbers) is the tendencies of inaccuracies in the crowdsourcing data. For example, while Table 6 informs us that 89.8% of the gold standard `rep`-cases have been identified as `rep` by the crowd, Table 7 shows us that the inaccuracy in the CS-data is owing to the crowd identifying the 18 false-`rep`'s (i.e. 1 $\texttt{other}_{\text{cl}}$ + 17 $\texttt{res\_ct}_{\text{cl}}$) as predominantly `res_ct` rather than `other`. Vice versa, the false hits for the GS-`res_ct`-data are mostly classified as `rep`-*again*s. The ratio of true hits to false hits for the two main readings (`rep` vs `res_ct`) is 9.3:1 for the $\texttt{rep}_{\text{gs}}$-data and 4.2:1 for $\texttt{res\_ct}_{\text{gs}}$-data – indicating higher confusion regarding the restitutive use of *again*. This is reflected in Fig. 6, where we plot the distributions of unit quality scores (UQS, cf. section 4.3.1) (as kernel density plots) for true/false positives for the two main GS-readings: `rep` and `res_ct`; cf. the two subplots in the left column where we have higher UQSs for true positives for repetitive *again*s and restitutive *again*s (i.e. '✓').

A noteworthy aspect of the crowdsourcing data is that in some instances the crowd annotations yielded very confident false positives. The most striking examples are those uses of *again* that achieved a unanimous vote but were in disagreement[16] with the GS

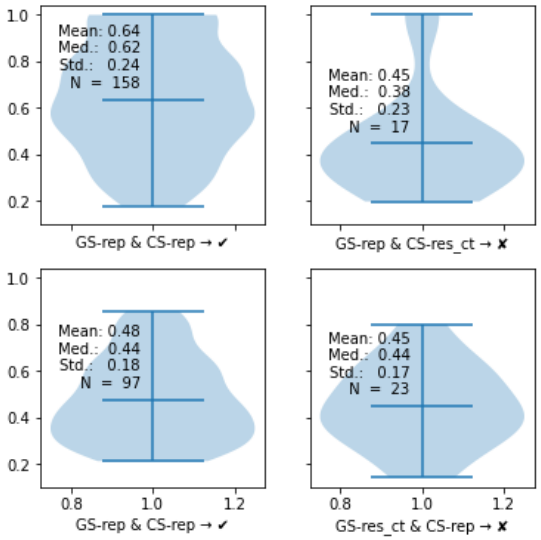

Figure 6: UQS by true/false positives for `rep`+`res`

annotations (UQS = 1, in top-right subplot, Fig. 6), to which we want to turn next. In (3) and (4), *again* does not require context to disambiguate. By virtue of the lexical aspect of the predicates both *again*s occur with, they license only the `rep`-reading: *to be safe* and *to be here* cannot restore a state as both encode states themselves. These uses of *again* express that the states they operate on hold at a time prior to reference time. The average (per-*again*) worker experience for (3) and (4) is 3.7 and 2.16 respectively[17] – indicating that lack of experience in our annotators should not be (fully) responsible for the confusion.[18] Focusing on (4) (from Charlotte Brontë's 'Wuthering Heights'), the annotators seem to have judged Catherine's (=*you*) return from Wuthering Heights in the middle of the night so salient that they coerced this return to be the relevant counterdirectional event to satisfy the PSP for a `res_ct`-reading of the *again* in (BRONTE-1848-2,2,283.219) (cf. Table 8 in Appendix, p. 13).

(3) She's here **again**!

[14] To be precise, Table 7 includes those 322 *again*s (out of 328 in total) that are not ambiguous according to the gold standard annotation.

[15] Notice how the column for $\texttt{ctd-data}_{\text{cl}}$ is empty because `ctd`-data is not represented in the KMeans model.

[16] based on all three modes discussed above

[17] the average experience over all data points is 2.42

[18] There is no detectable correlation between average experience and average cosine similarity with Pearson's r at -0.087

GS:`rep` vs. CS:`res_ct`
(CENTLIVRE-1723-2,53.571,
The Artifice)

(4) God be thanked, you are safe with us **again**!"

GS:`rep` vs. CS:`res_ct`
(BRONTE-1848-2,2,283.219,
Wuthering Heights)

As an explanation for why we get very confident crowd decisions in disagreement with our gold standard – particularly for the above cases – is that lexical aspect was not discussed in our annotation guidelines but only covered in our weekly tutorials.

5.1 Summary & conclusions

In this article, we have reported on the major findings of a pilot study on informed crowdsourcing for annotating the decompositional *again*. We have contrasted the crowdsourced data to an expert-annotated gold standard. In terms of yielding a 'winning annotation' among multiple crowd-based annotations, Kmeans clustering transpired as the best-performing mode in terms of observed accuracy (84.1%) and in terms of Cohen's κ (0.7). The pilot study discussed here allows the following main conclusions: First, given adequate data processing and relying on the inherent agreement and disagreement in crowdsourced data, we can achieve encouraging overall accuracies. This is especially true when relying on unsupervised classification – KMeans clustering in our case – which is performed on vectorized annotations sensitive to dimensions with lower magnitudes. And second, as a tentative conclusion, crowdsourcing data seems to bear the potential to reduce costs for 'manual' gold standard production by diverting resources to data identified as problematic/requiring more attention. Third, given, for example, the high-confidence false positives discussed in the previous section, we cannot assume that 'linguistic ambiguity' (e.g. the various readings of *again* due to its semantics – be that due to structural or lexical ambiguity) corresponds to 'crowd ambiguity' in a straight-forward way.

5.2 Outlook

Various avenues for (i) increasing the performance of the crowdsourcing approach and (ii) reducing the workload for expensive gold standard annotators remain to be explored.

With regard to the first point, one area that we have not drawn on with the above discussion is the available syntactic annotation of the data presented here. On the one hand, those features might be used for increasing accuracy in tandem with CS-data and, on the other hand, syntactic (and other corpus) features might at least partially explain confusion in the crowdsourcing data. An aspect of our data that we have left mostly untouched is the reasoning (and 'antecedent'-annotations) crowd workers provided along with the classification component, which might also be drawn on for understanding crowd confusion and improving accuracy.

As far as the second point is concerned at the moment, i.e. minimizing workload for creating a gold standard (but also for creating a quantitative, empirical basis to get an impression of the bigger diachronic picture), one strategy might be to focus on a review of CS-decisions (possibly aided by a condensate of crowd 'reasonings'): In a first step, coming back to the confusion matrix in Table 7, a review of the CS-`rep` data will allow us to confirm 158 uses of *again* as `rep`, and weed out 23 `res_ct` uses of *again*. Another major step toward utilizing crowdsourcing for annotating semantic variation and change is to recruit an anonymous and uninformed crowd. In terms of instructions, we hope to provide both text-based guidelines and how-to videos – both including at least a superficial discussion of lexical aspect. Especially with this last point, having an uninformed crowd perform the annotations discussed here, we hope to get a better simulated semantic change over multiple periods (Gergel et al., 2021). And finally, in the spirit of the notion of 'CrowdTruth' (Aroyo and Welty, 2015), we will add reviewing of our gold standard data to our agenda – especially for cases with high crowd confidence that disagrees with the gold standard.

References

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

Aroyo, L. and Welty, C. (2013b). Measuring crowd truth for medical relation extraction. In van Harmelen, F., Hendler, J. A., Hitzler, P., and Janowicz, K., editors, *Semantics for Big Data: Papers from the AAAI Fall Symposium*, AAAI Technical Report FS-13-04, Palo Alto, CA.

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

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

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Delin, J. (1992). Properties of it-cleft presupposition. *Journal of Semantics*, 9.

Deo, A. (2015). Diachronic semantics. *Annual Review of Linguistics*, 1:179–197.

Dowty, D. R. (1979). *Word Meaning and Montague Grammar. The Semantics of Verbs and Times in Generative Semantics and in Montague PTQ*. D. Reidel, Dordrecht, Holland.

Dumitrache, A., Oana, I., Aroyo, L., Benjamin, T., and Welty, C. (2018). Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement.

Eckardt, R. (2006). *Meaning Change in Grammaticalization. An Enquiry into Semantic Reanalysis*. OUP, Oxford.

Gergel, R. (2020). Sich ausgehen: Actuality entailments and further notes from the perspective of an Austrian German motion verb construction. In *Proceedings of the LSA Workshop "Formal Approaches to Grammaticalization"*, volume 5(2), pages 5–15. Linguistic Society of America, New Orleans.

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

Gergel, R., Blümel, A., and Kopf, M. (2016). Another heavy road of decompositionality: Notes from a dying adverb. In *Proceedings of PLC 39*, volume 22, pages 109–118, UPenn, Pennsylvania.

Gergel, R., Kopf, M., and Puhl, M. (2021). Simulating semantic change: a methodological note. In Beltrama, A., Schwarz, F., and Papafragou, A., editors, *Proceedings of Experiments in Linguistic Meaning (ELM)*, pages 184–196, University of Pennsylvania: LSA.

Gergel, R. and Nickles, S. (2019). Almost in Early and Late Modern English: Turning on the parametric screw (but not tightly enough to change a parameter). In Gattnar, A., Hörnig, R., and Featherston, S., editors, *Proceedings of Linguistic Evidence 2018*, pages 282–293, University of Tübingen, Tübingen.

Gergel, R. and Stateva, P. (2014). A decompositional analysis of *almost*: Bringing together diachronic and experimental comparative evidence. In *Proceedings of Linguistics Evidence*, pages 150–156, University of Tübingen, Tübingen.

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

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

Kroch, A., Taylor, A., and Santorini, B. (2000). *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*. Department of Linguistics, University of Pennsylvania, second edition. Release 4.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24:183–216.

Taylor, A., Warner, A., Pintzuk, S., and Beths, F. (2003). *The York-Toronto-Helsinki Parsed Corpus of Old English Prose, YCOE*. University of York. Release 3.

Appendix

Antr.	read.	annotator reasoning (unedited)
016	`res_ct`	"At first Catherine was ill, but now she is save again because she defeated the illness. We have a counterdirectional movement with the the result of her being save."
017	`res_ct`	"The mistress is named "Catherine". She is married to Master Heathcliff, whose father detests both. Master Heathcliff has been acting against Catherine. She was not "safe" staying with him, and now that she is back she is "safe". Res-ct: she came back + she is now safe"
036	`res_ct`	"The safeness of the father ist restored, because he is alive. "
039	`res_ct`	"Catherine has gone away and has now returned to them"
051	`res_ct`	"The father was not in a safe situation before. Now he has recovered."
052	`res_ct`	"The father is still alive so the event has not happened before."
064	`res_ct`	"Nelly is speaker; Nelly is happy to welcome Catherine in safety again, since they hoped for Catherine to come back (BRONTE-1848-2,2,282.184)"
065	`res_ct`	"I could not find an antecedent in the corpus (other than in the sentence itself), but the context hints at a restitutive use of again. "Are" is a stative verb and as such, the state of being safe was restored in the sentence. The counter-directional action was leaving and thus not being safe with the others in the household."
074	`res_ct`	"Her master is restored to his previous state of being safe with them. "
083	`res_ct`	"It seems like Ellen has been unsafe, probably it has something to do with her father being alive instead of dead. Hence Ellens state of being safe has been restored."
091	`res_ct`	"They weren't safe before but now, that they are reunited with their people, they are safe again. Counterdirectional and situation is the result state of a before different situation."
117	`res_ct`	"The event of her mistress being with them takes place in the opposite direction and is restored by her coming back again."

Table 8: Annotator (='antr.') decisions and reasonings for *again* in BRONTE-1848-2,2,283.219