Decomposing decomposition in time: a methodological investigation*

Martin Kopf[1] and Remus Gergel[1]

Saarland University, Saarbrücken, Germany
{martin.kopf|remus.gergel}@uni-saarland.de

Abstract. This empirical submission reports on two original methods geared towards producing semantic annotations for the decompositional marker *again*. The two methods are (i) expert annotation based on a comprehensive set of guidelines and (ii) quality-controlled crowdsourcing with ensuing evaluation on the basis of the expert annotation. We report on a number of strategies for yielding a 'crowd winner' and present as the most promising candidate KMeans clustering of annotation vectors which are supplemented with corpus and annotational features. We report an observed accuracy of 85.54% with Cohen's κ at 0.73.

Keywords: Annotation · Decomposition · Diachronic semantics · Corpus Linguistics · Classification · Crowdsourcing

1 Background and introduction

The goal of this empirical contribution is to present recent output generated by a DFG-funded project concerned with the diachrony of decomposition. We focus on the expert annotation of theoretically relevant ambiguities in readings of historical data on decomposition [5, 6, 11–13]. We present relevant results of expert annotation, and discuss two further methods for non-expert-based generation of semantic annotations in line with the project's secondary goal of seeking additional means of validating data annotated by linguists themselves (see also [14], for a more detailed discussion of these methods). Decompositional adverbs (e.g., *again* and its relatives in many languages) have received attention since they are insightful in a number of respects: They have been the subject of competing formal analyses (typically: structural vs. lexicalist). They also touch on the representation of events, presuppositions, and more generally, the way the structural and the meaning components of languages interface (cf. [22, 4, 24], among others). Moreover, recent inquiries into diachronic formal semantics indicate that diachronic data are also able to elucidate synchronic debates that

* Supported by DFG (Deutsche Forschungsgemeinschaft, German Research Foundation). We would like to thank our team of expert annotators Mohammad Bablli, Lucie Gend, Maryam Rajestari, and Dana Rebar, and Nergis Schäfer. Furthermore, we want to thank the students at Saarland University who provided the crowd-based annotations.

could not be solved otherwise thus far (cf. e.g., [8] for a recent installment of this idea). However, reliable diachronic data have remained a desideratum and come with major practical issues due to their resource intensive process of extraction, annotation, and stronger validation, as well as, when possible, partially automatic amplification/replication.

The structure of this submission is as follows. In Section 2, we start off with a brief introduction to the English adverb *again* and the ambiguity associated with it before we discuss two approaches to closing the empirical gap in diachronic data. Here, we detail the procedure behind exhaustively annotating its various readings with a team of expert annotators based on syntactically parsed diachronic corpora of English. These corpora are the PPCME2 [17] covering the Middle English ('ME') period, the PPCEME [15] and PPCMBE2 [16], covering the Early and Late Modern Eng. periods, respectively ('EModE,' 'LModE'). All ca. 4,380 occurrences of *again* from these three corpora have received an exhaustive annotation. The first—and diachronically most recent—portion of this semantic annotation, i.e., all 1,901 LModE data, is ready to be shared with the community along with a Python-based alignment tool to merge our semantic annotations with users' own instances of the PPCMBE2. As such, the current state of the output of our project constitutes an update next to recent reports [14] and provides the diachronic overview in section 2, below. Crucially, while our earlier report essentially contained the expert annotation for LModE (1700–1910 CE), we now additionally include the data from both EModE and ME (ca. 1150 CE onward). The second major part of this submission, Section 3, reports and summarizes the findings of an 'informed crowdsourcing' experiment, which was designed to explore crowd aptitude for providing nuanced semantic annotations on diachronic data. Thus, the crowd workers had to work with natural language data for which they did not have any native speaker intuitions whatsoever. We evaluate the performance crowd annotators on the basis of our expert-provided gold standard. Furthermore, we report on a number of strategies for eliciting a 'crowd winner' and for enhancing these crowd-based annotations with corpus features. We conclude with Section 4.

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

(1) The next year many of them will begin to flower; all the plants then must be examined, and such as produce the largest flowers and have good colors, should be planted in pots for stage flowers; but all the plain flowers, that is, those which have but one color, should be planted in borders among other low flowering plants; and those which are planted in pots, should in the following year's bloom be again examined, and placed in pots or borders accordingly as they desire. (FALLOWFIELD-1791-2,33.349)

(2) He hesitated, got up. [. . .] and he sat down again; (AUSTEN-1815-2,169.633)

The *again* in (1) has a repetitive reading ('*rep*'): An event of the same kind (examining plants) is presupposed. The *again* in (2) is restitutive/counterdirectional ('*res/ct*'), i.e., the *again* here does not (necessarily) presuppose a sitting-down event by him but an event in the opposite direction, that is, the sitting-down event restores a state that held at a time prior to reference time. The readings in (1) and (2) are the most frequent ones in the data discussed here and in line with the literature (cf. [11]). The data in (3) has a reading of *again* not available anymore in Present Day English (PDE), the counterdirectional (= 'counterdirectional proper') reading. The two main ideas here are (i) that an eventuality in the opposite direction is presupposed and (ii) the '*again*-event' does not feasibly come with a result state. This kind of *again* can often be paraphrased with 'back'. In (3), the knight loving the queen in return expresses a stative without a result state. In other, similar uses such as responding to a letter (without ever before having had any correspondence with the sender, i.e. the *rep* reading is out), *to write again* does not conceivably bring about a result state (that held at a prior point in time) but constitutes an action directed 'back to where the previous action came from':

(3) quene Gwenyvere had hym in grete favoure aboven all other knyghtis,
queen Gwynevere had him in great favor above all other knighty
and so he loved the quene agayne aboven all other ladyes
and so he loved the queen again above all other ladies
'... he returned the queen's love ...' (CMMALORY-M4,180.2394, 15$^{\text{th}}$ c.)

A fourth relevant reading of *again* are discourse-marker uses. Rather than operating on predicates, they have a discourse organizing function ('*dm*,' discourse marker). 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).

Note, that with the somewhat impractical label 'restitutive/counterdirectional' (short '*res/ct*') we remain theory neutral. By convention, the terms 'restitutive' and 'counterdirectional' come with a commitment to either a structural analysis (cf. [23, 22]) or a lexical analysis (cf. [10]), respectively. On the one hand, including the separate label 'counterdirectional (proper)' (short '*ctd*'; both in the expert and crowd based annotations) has its merits in the observation that—in data like (3)—*again* predicates encode counterdirectionality in absence of a result state (cf. [6, 11] for a thorough discussion of the diachronic relevance of *ctd* again and *again*'s origins as a preposition). On the other hand, for practical reasons in the below (quantitative) discussion, the smaller class of *ctd* uses of *again* will occasionally be subsumed under the larger class of *res/ct again*s. We will be transparent as to the two strategies.

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

2.1 Method

Based on presupposition (PSP) satisfaction in the linguistic context, our annotators (i) classified any use of *again* according to its reading, (ii) marked the main verb of the *again*-predicate ('target verb'; e.g. *he *sat* down again* in (2)), and (iii) marked the main verb of the antecedent satisfying a relevant PSP (e.g. *he hesitated, _got_ up* in (2)). 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' ('INF')—if it 'only' allowed the inference of a relevant PSP but did not constitute a perfect antecedent in a narrow sense. During the initial phase of the annotational work (especially on the PPCMBE2 data) we fine-tuned our annotation guidelines and our annotations in an iterative process as a team of annotators. The resulting multi-page set of annotation guidelines has remained the basis for further annotational work. On a global level, our guidelines needed to be 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 potential antecedent events. For instance, our guidelines considered proximity between *again*-event and plausible antecedents as crucial. See (4) as an EModE example (from Robert Boyle's Experiments and considerations touching colours; 1664) where *again* operates on the predicate *reduce to whiteness*. At first glance, on a repetitive reading, the relevant PSP would be satisfied in the context with "the whole mixture will appear White" (marked with double slashes around the main verb). On a counterdirectional reading, "the Whiteness will presently disappear" is a viable antecedent. However, the material encoding the same counterdirectional PSP, which is marked with double underscores, is closer to where we find the *again*-event encoded. Therefore, on the one hand, this use of *again* is to be classified as *res/ct*, and, on the other hand, the place containing the relevant PSP closest to the *again*-predicate is annotated as antecedent.[1]

(4) [A]nd into a spoonfull or two thereof [filtered mix of 'Fair Water' and 'Common Sublimate'], (put into a clean glass vessel,) shake about four or five drops (according as you took more or less of this Solution) of good limpid Spirits of Urine, and immediately the whole mixture will //appear// White like Milk, to which mixture if you presently add a convenient proportion of Rectifi'd (Aqua Fortis) (for the number of drops is hard to determine, because of the Differing Strength of the liquor,

[1] As an anonymous reviewer points out, the repetitive reading would be favored from a Present Day English perspective which has preverbal *again* trigger an obligatorily repetitive interpretation. Precisely in order to remain consistent in face of diachronically varying word order facts, we rely on annotation guidelines independent from precedence and other structural considerations (cf. e.g. [10]).

> but easily found by trial) the Whiteness will presently //disappear//, and the whole mixture _become_ Transparent, which you may, if you please, again *reduce* to a good degree of Whiteness (though inferiour to the first) onely by a more copious affusion of fresh Spirit of Urine. (BOYLECOL-E3-P1,134.11)

As an example in the same vein, but with the twist that mere discursive proximity is not enough to reliably identify antecedents and to disambiguate, consider (5), from George Adams' Essays on the microscope (1787). At first glance, the material marked with double slashes seems to satisfy a repetitive PSP, thus, making for a putative antecedent. A closer look reveals that the parting-event (marked with double underscores) is the proper antecedent event to the *again*-event (in bold) satisfying a restitutive/counterdirectional PSP:

(5) The filaments of a cortical vessel are to be looked on (agreeable to what we have already observed) as so many little bundles placed near together, and at first growing parallel to each other; but soon quitting this direction, the filaments of one fascicle _parting_ from that to which they originally belonged, and inclining more or less obliquely towards another, sometimes //uniting// with it, at others bending backwards, and *uniting* again with that from which it proceeded, or with some one that it meets with. (ADAMS-1787-2,663.157)

Every single use of *again* received 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. In later phases, particularly for EModE and ME data, disagreements were reconciled with a third and sometimes fourth annotator review.

2.2 Results

Based on our expert annotations, we get the diachronic picture in Table 1 and Figure 1 for the time span from ca. 1150 CE to 1910 CE. These two simplified graphs represent a set of 4,377 uses of *again*: 945 from PPCME2, 1,532 from PPCEME, and 1,900 from PPCMBE2. Recall that the corpora represent the ME, EModE, and LModE periods respectively. They are further subdivided (M1 being the first ME subperiod, E1 the first EModE one, etc.). The charts show the relative frequencies of the two major readings 'repetitive' (*rep*) and 'restitutive/counter-directional' (*res/ct*)—with *ctd* included in *res/ct*. Note further, all other uses of *again*, i.e. unresolvable ambiguous uses, discourse marker uses, etc. are subsumed under the label *other*, cf. Figure 1 and Table 1. Moreover, Table 1 shows the frequency of adverbial *again* throughout the corpus data along with the number of available *again*s respectively. In particular, the overall decrease of *res/ct* readings and increase of *rep* readings clarifies and certifies previous accounts on the diachronic development of *again* [5, 6, 11], which had been done on disparate corpora (i) solely based on correspondence and (ii) lacking the 18th century (currently the most general unified corpora are used, from which

Table 1 are examples). In the next section, this expert annotation will serve as the gold standard for evaluation of crowdsourced data.

subperiod	major readings (%)			freq.	*again*
	rep	*res/ct*	*other*	#	%
M1 (1150-1250 CE)	15.8	78.9	5.3	57	0.023
M2 (1250-1350 CE)	9.3	85.2	5.6	54	0.037
M3 (1350-1420 CE)	12.6	84.9	2.5	484	0.099
M4 (1420-1500 CE)	17.4	81.7	0.9	350	0.133
E1 (1500-1569 CE)	31.2	62.2	6.7	526	0.088
E2 (1570-1639 CE)	41.9	46.8	11.3	613	0.103
E3 (1640-1720 CE)	40.7	54.2	5.1	393	0.073
L1 (1700-1769 CE)	50.8	43.6	5.6	486	0.060
L2 (1770-1839 CE)	59.0	33.8	7.2	639	0.070
L3 (1840-1910 CE)	62.6	24.9	12.5	775	0.076

Table 1: Frequency of *again* & its major readings, 1050–1910 CE

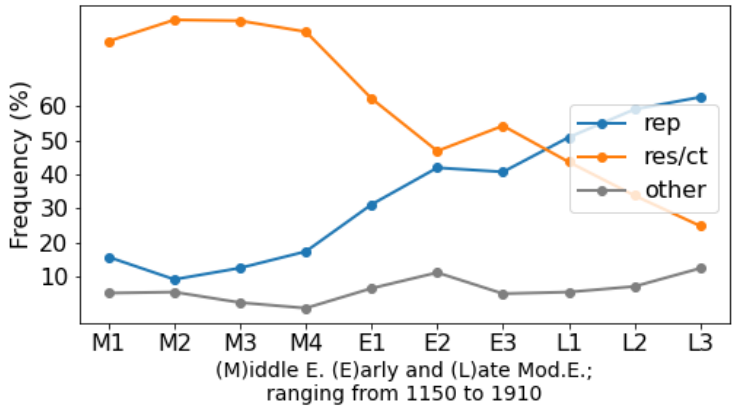

Fig. 1: Relative frequencies of major readings

3 Informed Crowdsourcing pilot

In this section, we report on an 'informed crowdsourcing' experiment. We are exploring the potential of crowdsourcing for the creation of semantic annotations for diachronic data. Note, that we remain committed to providing state-of-the-art expert annotations with our team of trained annotators (cf. Section 2). However, 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. 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: majority vote, quality-scores-adjusted votes (cf. e.g., [3]), and KMeans clustering [20]. Section 3.2 will introduce 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 3.3, 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 3.3.

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 make 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 1). The data was provided to the crowd workers in spreadsheet files, with various meta data (corpus ID, date of compositions, etc.) and dedicated, empty columns to be used by the annotators. In more detail and in condensed form, the task instructions included:

- Classify all uses of *again* according to their readings. In the column "reading," use the labels *rep*, *res/ct*, *ctd*, and *other* (i.e. crowd workers were instructed to pay attention to the *res/ct*-vs-*ctd* distinction mentioned in Section 1).
- Annotate the place in the context that helped disambiguate between possible readings: 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 in a clause/fragment allowing the inference that one PSP over another, competing PSP is satisfied in the context.
- In the separate column names "reasoning," briefly explain your reasoning as to why a particular decision was made.
- In the column "ant," indicate whether an antecedent was found; 'yes'/'no'.
- Try to avoid ambiguities. If, however, based on PSP satisfaction and historical context/world knowledge etc. an ambiguity cannot be resolved, 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. [2]). 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].

3.1 Diachronic data

Data for this crowdsourcing pilot were sourced from the PPCEME [15] and the PPCMBE [16]. Out of these two corpora, we picked corpus texts based on the following criteria: (i) For each of the 17$^{\text{th}}$, 18$^{\text{th}}$, and 19$^{\text{th}}$ centuries we selected about 100(+) uses of *again*. (ii) We prioritized the most abundant corpus texts (with respect to absolute frequency of *again*). (iii) We excluded biblical corpus texts since they tend to be more conservative and we wanted our crowd workers to face data representative of its period(s). Regarding (ii), we tried to keep worker fatigue to a minimum and, thus, wanted to avoid workers having to familiarize themselves with new texts/contexts over the course of the study too many times. This resulted in:

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

As introduced above in section 2, as a gold standard for evaluating the crowdsourcing data, 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: First, the classification task based on PSP satisfaction often required our workers to read long stretches of context in order to complete their task. Second, we served diachronic language data rather than present-day language data to the crowd workers. As a consequence, 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 `repetitive`/`restitutive` ambiguity), who made up the majority of our crowd workers.

[2] On occasion, external sources needed to be consulted for creating the gold standard. For instance, if a particular context was on historical events such as e.g. enduring armed conflict between two entities and the corpus text at hand did not disambiguate between competing readings, external sources were consulted in order to supplement the primary context.

3.2 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 crowd workers. We then continue with details on data handling, and data set design before closing with a brief outline of the elicited data.

Crowd recruitment: The crowdsourced annotations were collected over the course of two semesters at the English Department at Saarland University. 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, while the workers were not mere speakers of English providing intuitions, they were not fully-trained as expert annotators either. 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.

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

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

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).

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 spreadsheet files (with one use of *again* per row). Aside from various meta- and corpus-related information, the spreadsheet 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 spreadsheet files not suffice and to copy and paste the relevant antecedent sentence(s) into the spreadsheet file (with the corresponding IDs) in order to perform the annotations.

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:

17^{th} cent:	1,086 data points
18^{th} cent:	969 data points
19^{th} cent:	1,264 data points
	3,319 data points

3.3 Evaluation

We will begin this section with a general discussion of data processing and a brief overview of the crowdsourced annotation data. We then move on to introduce a

[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 3.3), and (ii) a more immediate vector representation of ambiguity if we consider the four basic classes as definitive dimensions in a vector space (cf. 3.3). 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.

number of approaches to deciding on 'crowd winners' among potentially divergent crowd annotations. For each approach we will provide the observed accuracy and, where applicable, Cohen's Kappa for overall inter-annotator agreement, as well as agreement by subperiods and by GS-readings [7].[5]

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 2 (pre-) and Table 3 (post-conversion; both page 12)[6]. Note, that in the following discussion, we excluded a small number *again*s that were annotated as ambiguous as per our gold standard, resulting in 325 *again*s. We also excluded data that the crowd workers labeled as ambiguous, reducing our initial 3,319 data points to 3,180. Moreover, this evaluation of our informed crowdsourcing is based on subsuming all *ctd* readings under the *res/ct* label. As above, the major readings (and corresponding labels) are *rep*, *res/ct*, and *other*—the latter containing all discourse related uses of *again*.

We will refer to the one-hot encoded data in the rows in Table 3 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 4 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).

Turning back to our crowd data, for visualization, we can fit our raw unit vectors into a principal component analysis (PCA) [20], see Figure 2, 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 Figure 2 correspond to gold standard readings. Only the main readings appear in the legend. However,

[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 [21]. Cohen suggests 0.61–0.8 as a range for "substantial" agreement [18]. 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 crowdsourced 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.

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

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 2: 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 3: Annotations as one-hot vectors

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

Table 4: Unit vectors as total of 1-hot vec.s

ambiguous data are represented in the chart (shaded in gray). Note, that in the discussion below, we subsume all *ctd* readings under the wider label *res/ct*. Further, we do not include two uses of *again* that were classified as ambiguous in our expert annotation. Consequentially, the same data are excluded from the crowdsourcing data for the purposes of the current discussion. Therefore, our CS dataset is reduced to 3,214 observations (instead of 3,319).

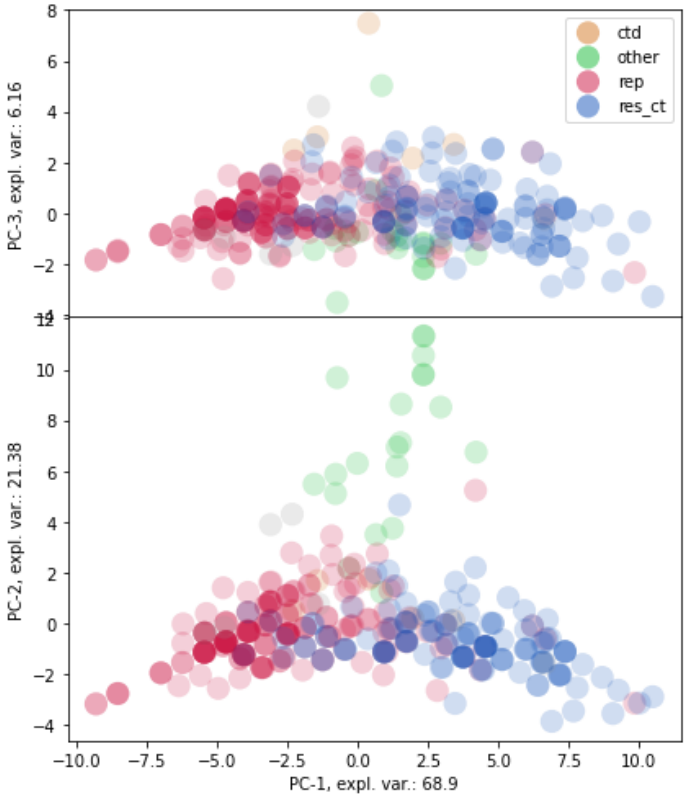

Fig. 2: PCs1–3, 'raw' unit vectors

1. Majority vote: With the majority vote approach, the maximum value of a reading per unit vector will decide the 'winning crowd label,' e.g. in Table 4 this would be 'lev_2' for u_vec '*u*1'. In order to avoid ties between two (or more) crowd classifications, we established a tie breaker system: Every data point vector was adjusted for meta-features of the respective data point. These features were ranked according to perceived relevance for providing reliable judgments and combined into a tie-breaker$_{dp}$ score; cf. below list for details. Note, that due to the

ranking associated with each feature, higher rank features take precedence over lower rank features. As an example, if $\texttt{Experience}_{\text{dp}}$ comes to break a tie, then none of the lower ranking features have any bearing. Similarly, if $\texttt{Experience}_{\text{dp}}$ is insufficient to break a tie but $\texttt{Average evaluation}_{\text{dp}}$ does, then $\texttt{Semester progress}_{\text{dp}}$ and $\texttt{Motivation}_{\text{dp}}$ have no bearing. Finally, $\texttt{Motivation}_{\text{dp}}$ will only break a tie if all higher ranking features are identical for a data point. The tie breaker system is relevant for 18 out of 325 uses of *again*. For remaining 307 uses of *again* (i.e. 94.5%), the bare counts suffice to yield a winner. Based on this relatively simple approach, we achieve an overall accuracy of 83.38% (with Cohen's Kappa at κ=0.69).

1. $\texttt{Experience}_{\text{dp}}$ stands for the experience the worker had when providing the data point at hand, i.e. the number of complete data sets the worker had submitted prior to providing a data point (ranging from 0 to 11; dp_vec adjusted by 10^{-3}).
2. $\texttt{Average evaluation}_{\text{dp}}$ stands for the average evaluation (i.e. the point system for grading purposes) a student received for the submission of the data set that the data point originates from (from 0.0 to 1.0; dp_vec adj. by 10^{-6}). The point system is based on formal criteria rather than a gold standard bases notion of correctness,
3. $\texttt{Semester progress}_{\text{dp}}$ stands for how far into the semester (i.e. ordinal number of weekly roll-outs) the data point was produced (from 1 to 12; 10^{-9}). This is distinct from experience since any student i might have submitted their first dataset in week 1, while a difference student j might have submitted their first data set in week 10. In such a scenario, data points from those respective first data sets by students i and j have the same $\texttt{Experience}_{\text{dp}}$ score, but different $\texttt{Semester progress}_{\text{dp}}$ scores (1 for i and 10 for j). The underlying intuition is that a student who has been part of a lecture—regardless of the intensity of their engagement—is better informed than a student at the beginning of the same lecture.
4. $\texttt{Motivation}_{\text{dp}}$ gives the total number of data sets the worker submitted in total (from 2 to 12; 10^{-12}). The underlying intuition here is the hypothesis that motivated students produce more reliable annotations.

	17^{th} c.		18^{th} c.		19^{th} c.		all	
	N	%	N	%	N	%	N	%
rep	51	90.20	58	84.48	72	88.89	181	87.85
res/ct	59	71.19	38	78.95	30	90.00	127	77.95
other	2	50.00	5	80.00	10	80.00	17	76.47
all	112	79.46	101	82.18	112	88.39	325	83.38
C's κ		0.61		0.67		0.77		0.69

Table 5: %-acc. CS data w/ KMeans) for gold-std. classes

Table 5 provides a detailed overview. The major points here are: The crowd performs relatively well on *rep-again*s (as per our gold standard), while the crowd workers had difficulties with historically older *res/ct-again*s. On a more general level, overall accuracies are consistently lower for historically older data (79.46% for 17^{th}c. data vs 88.39% for 19^{th}c. data), which matches our initial intuitions since the grammars that generated the 17^{th}-cent. data are expected to be more alien to present-day speakers of (L2) English.

2. Adjusting annotations with quality metrics: Drawing on the "CrowdTruth" approach proposed by Aroyo & Welty in [1–3], 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 [9] 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 similarity is conceived of as the (positive range of) cosine similarity between vectors.

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 workers i and j pairings:

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

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 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 [9], we weighted this score with the relevant *UQS(u)* (cf. Section 3.3). The idea here is to not 'punish' workers for the work they did on controversial or difficult uses of *again*.

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

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[7]:

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

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 & Welty's and Dumitrache et al.'s terms), we can adjust the one-hot data point vectors for both UQS and WQS. Adding these up into unit vectors and picking the maximum (i.e. strongest dimension) as the 'crowd-sourcing' winner, we get a slightly improved overall accuracy of 84.0% (κ=0.7).

3. KMeans clustering: This approach entailed unsupervised classification of our crowdsourcing data based on 'KMeans clustering' [20]. We again relied on the above quality metrics (section 3.3) derived from the CrowdTruth literature in combination with KMeans clustering. Further, we normalized our 325 quality-metrics adjusted unit vectors in order to facilitate a clustering of our data.

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. Figure 3 where—in our case—the gains in reduction of sum of squares (SS) start leveling out with three clusters.[8] Allowing our data points to be sorted into three clusters, we can calculate accuracy values by assuming the most frequently represented (i.e. modal) gold standard label in each cluster as the canonical class of the cluster.

This approach raises the overall accuracy of the crowdsourcing data to 84.31% (κ=0.71). For all subperiods, accuracies are above 80%: 1700s, 82.14%; 1800s, 83.33%; and 1900s, 85.09%.

4. Enhancing crowd-based annotations with corpus features: From previous evaluations of a Naïve Bayes (NB) classifier of the expert annotations, we conclude that a relatively small set of features can result in respectable accuracies in classifying diachronic uses of *again* [14]. Drawing on those NB-based evaluations, we rely on a number of features from the corpus data and enhance the crowd-sourced data. The corpus features we used for this approach are (*i*) word forms (uni-grams) in the clause of the *again*-predicate (as delimited in the

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

[8] The SS is maximal for *K=1* and *0* for *K*=number of data points.

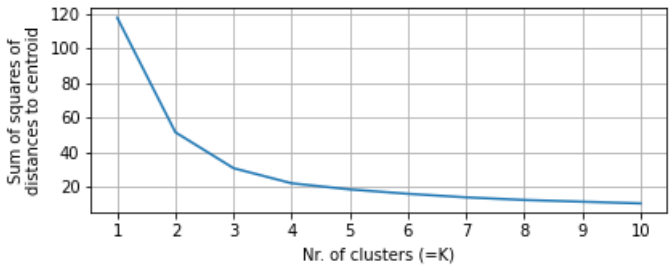

Fig. 3: Within Cluster Variation by K's

syntactic parse of the PPCHE corpora), (*ii*) Present-Day-English lemma, (*iii*) century, and (*iv*) 'result structures'. The last feature, 'result structures', amounts to a one-hot encoding of the presence (or absence) of two objects (one direct and one indirect, labeled 'NP-OB1' and 'NP-OB2' respectively in PPCHE), the presence of an adverbial particle ('RP' in PPCHE), the presence of a directional phrase ('XP-DIR'), and the presence of a goal PP (approximated by taking note of PPs headed by prepositions derived from the preposition *to*, as well as forms thereof). The motivation to include this feature is that these types of predicates overtly encode a result state, cf. [11, 6].

Where required, we separately encoded those features as one-hot vectors resulting in a matrix with 1115 features. By relying on a Principal Component Analysis [20], we applied dimensionality reduction down to 75 components (with the combined ratio of explained variance at 77.65%). We applied normalization to this resulting principal component matrix and concatenated it with the normalized and crowd-metrics-adjusted crowd-sourcing vectors. This matrix has the dimensions 325 (number of unit vector) by 78, i.e. 75 principal components derived from the corpus features plus three crowd-sourcing vectors (i.e. the number of major readings). Supplying this combined matrix to a KMeans clustering algorithm results in an overall accuracy of 85.54% ($Kappa$=0.73). For completeness' sake, note that the corpus-feature matrix on its own (i.e. only 75 principal components) performs at 58.46%. Consider Table 6 for a detailed overview of performance by century and gold standard readings.

While a 2-percentage-points increase in overall accuracy could be viewed as modest, the most valuable improvements have been achieved for those parts of the data that performed lowest with, for example, a majority vote approach, cf. Table 5 on p. 14. The *res/ct-again*s, in particular 17$^{\text{th}}$- and 18$^{\text{th}}$-century data, can be improved to accuracies at around 80% (or higher) when 'bare' crowd annotations are enhanced with crowd-quality metrics on the one hand and corpus features on the other hand.

	17^{th} c.		18^{th} c.		19^{th} c.		all	
	N	%	N	%	N	%	N	%
rep	51	88.24	58	84.48	72	90.28	181	87.85
res/ct	59	79.66	38	81.58	30	90.00	127	82.68
other	2	50.00	5	80.00	10	90.00	17	82.35
all	112	83.04	101	83.17	112	90.18	325	85.54
C's κ	112	0.67	101	0.69	112	0.81	325	0.73

Table 6: %-accuracies & Cohen's κ for corpus-feature enhanced CS data w/ KMeans) by gold-standard readings and centuries

Discussion: The declared goal of this contribution is to explore avenues to reduce the workload/cost involved in producing corpus data of a degree of quality allowing robust conclusions regarding possible correspondences between semantics and syntax from a diachronic perspective. Multiple candidates of anaphoric relations need to be evaluated next to one another in order to confirm a particular PSP as satisfied over its competitors. Moreover, the annotated corpus data are historical English—a language variety for which none our annotators (both experts and crowd workers) have native speaker intuitions. To our knowledge, the corpus our project has been able to produce is the first of its kind. Thus, it is important to stress that the accuracy rates discussed here should not be evaluated in light of benchmarks, nor do we aim to establish any such industry standards.[9]

While overall accuracy evidently incorporates performance on all data, the crux lies in the various sub-types of data. In particular, CS data that comes out as *rep* has higher accuracy rates while *res/ct*-CS is lower in accuracy. This picture remains by and large consistent over various modes of eliciting a crowd winner, cf. [14] for more details. Similarly, historically older corpus data is associated with lower accuracy rates than more recent data. Consequently and with respect to future work, if an expert based review of CS data needs to be selective due to limited resources, our pilot study allows the conclusion that focusing on the (older) CS-*res/ct* data is prima facie a valid recommendation.

However, considering the unit quality score corroborates this conclusion. As discussed in Section 3.3, the UQS is based on cosine similarity (thus, ranging from 0 to 1) and can be interpreted as a measure for agreement among the set of crowd workers with respect to one unit. Figure 4 shows the distribution of UQSs (as kernels density plots) for four types of the data, as discussed in (6), respectively:

(6) *Type A:* In the top-left subplot, there are 159 cases where the expert annotators (GS) and crowd workers (CS) agree on the label *rep*: 'GS-rep & CS-rep → ✓'.

[9] To clarify with respect to the concerns of an anonymous reviewer, our overall accuracies do not represent a meaningful threshold themselves.

Type B: In the bottom-left subplot, there are 105 cases where expert annotators (GS) and crowd workers (CS) agree on the label *res/ct*: 'GS-res/ct & CS-res/ct → ✓'.
Type C: The top-right subplot shows 17 cases where the expert annotators (GS) annotated as *rep* and the crowd workers (CS) decision was in favor of *res/ct*, resulting in the mismatch 'GS-*rep* & CS-*res/ct* → ✖'.
Type D: The bottom-right subplot shows 22 cases where the expert annotators (GS) annotated as *res/ct* and the crowd workers (CS) decision was in favor of *rep*, resulting the mismatch 'GS-*res/ct* & CS-*rep* → ✖'.

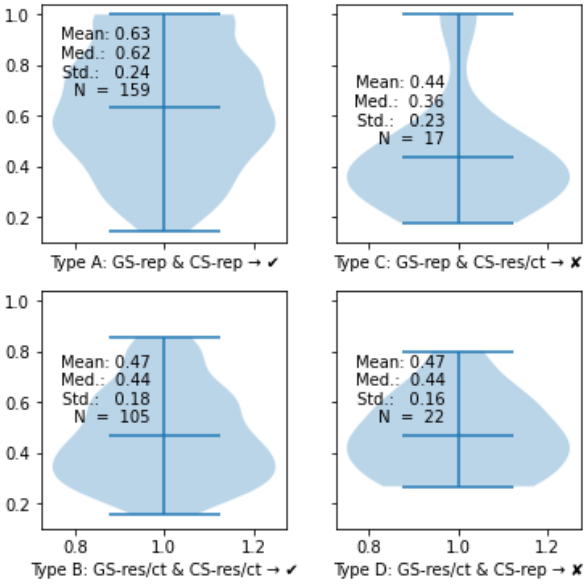

Fig. 4: UQS for GS-CS-matches & mismatches; as kernel density plots

For type-A data, we get a higher UQS average. Data types B, C, and D are consistently low. Furthermore, the distribution of UQSs is skewed towards the high end in type A data and in data types B, C, and D the distributions are skewed towards the low end. Thus, again with respect to future (or similar) work, the gains of an expert review of CS-*res/ct* data (types B+C) seem more promising than a review of CS-*rep* data (types A+D). However, the remaining mismatches in the type D instances remains a known risk. A complementary strategy could be to review crowd annotations that fall under a certain UQS threshold regardless of the CS decision.

While the crowd workers achieved respectable (overall) accuracy with a bare majority vote, the above discussion of multiple strategies for eliciting a 'crowd winner' shows that observed accuracy (and Cohen's κ) can be improved upon,

cf. Tables 5 and 6, above. Both crowd quality metrics and structural features (drawn from a syntactically annotated corpus) lend themselves to improve on the performance of crowd based annotations.[10] However, the separate moves towards ever improving accuracy rates discussed here are not to taken as a race for arbitrary benchmark ratings. Rather, each strategy is telling in its own right: For example, relying on crowd sourcing annotations produced for data that lack any syntactic annotations whatsoever (which in turn prevents enriching CS data with syntactic features as in Section 3.3), e.g. the vast EEBO collection [19], will result in accuracies as mentioned in Section 3.3 (cf. [14] for a more detailed discussion analogous to Tables 5 and 6). Similarly, aside from contributing improvements in accuracy, crowd-quality metrics, in particular UQS, can guide investment of scarce resources—especially with high-volume corpora as the EEBO.

4 Conclusion and outlook

In this article, we have reported on the major findings of an exhaustive expert annotation of the decompositional PSP marker *again*, and of a pilot study on informed crowdsourcing for annotating the decompositional *again*. We have evaluated the crowdsourced data on the basis of our expert-annotated gold standard. For the purpose of eliciting a 'winning annotation' among multiple crowd-based annotations, KMeans clustering transpired as the best-performing mode in terms of observed accuracy (84.31%) and in terms of Cohen's κ (0.71). We are able to improve on these results by combining the crowd-based annotations with structural features provided by the syntactic annotation of the PPCHE.

This pilot study 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. 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, there is an added benefit in working with existent syntactically annotated corpora as only a small number of features can bring about a boost in accuracy, as well as an improved κ value. It is important to stress that 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.

Our current enterprise can be seen, in a nutshell, as an exercise in empirical cross-validation of data which we aim to contribute to a future corpus that combines syntax and semantics in relevant ways. Decompositionality is only one area that helps linguistics understand correspondences (or sometimes the lack

[10] KMeans clustering of the 75-dimension PCA matrix derived from the corpus features combined with the *bare* crowd sourcing unit vectors (i.e. without crowd quality metrics) results in an observed accuracy of 81.23%

thereof) between syntax and semantics better. While the Penn corpora of historical English are already an established resource for historical syntacticians, there are still several parts that we take to be insightful add-ons when it comes to interpretation, e.g. quantificational, but also scopal readings more generally, modal flavors, etc. They can all throw significant light on developmental routes in the evolution of natural language semantics. What such areas have in common is that they show ambiguities that can (unlike lexical ambiguities) only be fully understood when both structure and a reliable sense of meaning have been established. Quite naturally, the further we go back in time for the purposes of diachronic semantics, the less applicable large data models may seem. Given this clear factual obstacle (at least currently so), we have overall sought to show that thoroughly cross-checked empirical work that is needed for any further theorizing on highly valuable corpora that already have structure incorporated can still be conducted in a meaningful way.

References

1. Aroyo, L., Welty, C.: Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In: Web Science 2013. New York: Association for Computing Machinery (2013)
2. Aroyo, L., Welty, C.: Measuring crowd truth for medical relation extraction. In: van Harmelen, F., Hendler, J.A., Hitzler, P., Janowicz, K. (eds.) Semantics for Big Data: Papers from the AAAI Fall Symposium. AAAI Technical Report FS-13-04, Palo Alto, CA (2013)
3. Aroyo, L., Welty, C.: Truth is a lie: Crowd truth and the seven myths of human annotation. AI Magazine **36**(1), 15–24 (2015). https://doi.org/"https://doi.org/10.1609/aimag.v36i1.2564"
4. Beck, S.: There and back again: A semantic analysis. Journal of Semantics **22**, 3–51 (2005)
5. Beck, S., Berezovskaya, P., Pflugfelder, K.: The use of *again* in 19th-century English versus Present-Day English. Syntax **12**(3), 193–214 (2009)
6. Beck, S., Gergel, R.: The diachronic semantics of English *again*. Natural Language Semantics **23**(3), 157–203 (2015)
7. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)
8. Degano, M., Aloni, M.: Indefinite and free choice: When the past matters. Natural Language and Linguistic Theory **40**, 447–484 (2022)
9. Dumitrache, A., Oana, I., Aroyo, L., Timmermans, B., Welty, C.: Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement (2018). https://doi.org/"https://arxiv.org/pdf/1808.06080.pdf"
10. Fabricius-Hansen, C.: Wi(e)der and again(st). In: Féry, C., Sternefeld, W. (eds.) Audiatur Vox Sapientiae. A Festschrift for Arnim von Stechow, pp. 101–130. De Gruyter, Berlin (2001)
11. Gergel, R., Beck, S.: Early Modern English *again*: a corpus study and semantic analysis. English Language and Linguistics **19**(1), 27–47 (2015)
12. Gergel, R., Blümel, A., Kopf, M.: Another heavy road of decompositionality: Notes from a dying adverb. In: Proceedings of PLC 39. vol. 22, pp. 109–118. UPenn, Pennsylvania (2016)

13. Gergel, R., Nickles, S.: Almost in early and late modern english: Turning on the parametric crew (but not tightly enough to change a parameter). In: Proceedings of Linguistic Evidence 2018. pp. 282–293. University of Tübingen, Tübingen (2019)
14. Kopf, M., Gergel, R.: Annotating decomposition in time: Three approaches for *again*. In: Prange, J., Friedrich, A. (eds.) Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII). pp. 129–135. Association for Computional Linguistics, Toronto, Canada (2023)
15. Kroch, A., Santorini, B., Delfs, L.: The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania, first edn. (2004), http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCEME-RELEASE-3, release 3
16. Kroch, A., Santorini, B., Diertani, A.: The Penn Parsed Corpus of Modern British English (PPCMBE2). Department of Linguistics, University of Pennsylvania, second edn. (2016), http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1, release 1
17. Kroch, A., Taylor, A., Santorini, B.: The Penn-Helsinki Parsed Corpus of Middle English (PPCMBE2). Department of Linguistics, University of Pennsylvania (2000), https://www.ling.upenn.edu/ppche/ppche-release-2016/PPCME2-RELEASE-4/, release 4
18. McHugh, M.L.: Interrater reliability: The kappa statistic. Biochemia Medica **22**(3), 276–282 (2012)
19. Oxford Text Archive: Early English Books Online (Phase 1) (EEBO). University of Oxford (2015), accessed on Sept 27, 2023, https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/5
20. 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., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
21. Poesio, M., Vieira, R.: A corpus-based investigation of definite description use. Computational Linguistics **24**, 183–216 (1998)
22. Rapp, I., Stechow, A.v.: *Fast* 'almost' and the visibility parameter for functional adverbs. Journal of Semantics **16**, 149–204 (1999)
23. Stechow, A.v.: The different readings of *wieder* "again"': A structural account. Journal of Semantics **13**, 87–138 (1996)
24. Zwarts, J.: From 'back' to 'again' in Dutch: The structure of the 're' domain. Journal of Semantics **36**, 211–240 (2019)