Crowdsourcing for generating diachronic semantic annotations in presupposition triggers: a pilot on *again*

The annotational evaluation of presuppositional data is hard to verify from theoretical corpus approaches (cf. e.g. Delin, 1992; Spenader, 2002). We present a pilot for a novel crowd-based approach to generating semantic annotations of diachronic corpus data. To account for the intricacies of a diachronic semantics annotation task, the 'crowd' of annotators was sourced in a 'controlled' manner, as the annotators were all participants of a History-of-English lecture (during the winter semester of 2021/22). The diachronic data was sourced from the Penn Parsed Corpora of Historical English, specifically the subcorpora for Modern British English (PPCMBE, Kroch et al., 2016) and Early Modern English (PPCEME, Kroch et al., 2004) respectively.

Our item of interest was the Engl. adverb *again* with its possible readings. We chose only non-religious texts such that at least ca. 100 tokens containing *again* would be available for each subperiod. In establishing a gold standard ('GS') for the data set, a small number of controversial/unclear uses of *again* were excluded from the data set that was intended for the crowdsourcing ('CS') experiment at hand. Thus, the data set for the current survey consists of 216 uses of *again*.

The task of the annotators was to classify uses of *again* according to its different readings:

(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 diachronic discussion

The *again* in (1) is ambiguous. On the 'repetitive' reading ('rep') (2-a), the *again* presupposes that an event of the same kind has occurred prior to reference time. On the second reading (2-b), the result state of the *again*-predicate is being restored and, thus, an event in the opposite direction is presupposed (i.e.'restitutive/counterdirectional', 'res_ct' for short). These two readings are the predominantly available uses of *again* in the data discussed here. Other readings include a 'counterdirectional-proper' ('ctd') reading (i.e. lacking a result state) and discourse-related uses where *again* has a discourse-organizing function rather than operating on predicates ('other' for short). Our annotators were provided with repeated tutorials and 'guided-annotation-sessions' and a rather trim one-page sheet of 'annotation guidelines' (in contrast to the multiple-page document for our expert annotators), which introduced them to the above readings and further detailed the annotations. Aside from classifying occurrences of *again*, the annotations needed to include whether the readings could be disambiguated based on presupposition satisfaction in the context (by means of antecedents or other textual material allowing relevant inferences) and to mark such places in the context. Additionally, annotators were asked to comment on their decision-making.

The student annotators submitted at least three annotation tasks (as a minimum grade requirement). Over the course of twelve, weekly rounds of annotations, 1,152 individualized annotation tasks were made available to 96 registered students. The data set for each annotation task consisted of five randomly assigned uses of *again*, distributed directly to the annotators' email inboxes. The motivation for this was to create a reasonably low possibility of some annotators ending up with identical data sets and any of them being aware of this fact. The goal was to keep annotators from coordinating, cooperating, and comparing notes about the annotations. In order to prevent uses of *again* being served to an annotator more than once, and in order to prevent gross imbalances in the number of annotations per token, a continuous record of previously assigned tokens and returned annotations informed data set compilation as the crowdsourcing study unfolded.

78 annotators provided 1,629 annotations for the 216 different uses of *again* from Late Modern English (670 annotations for 102 18th-century *again*'s, 959 annotations for 114 19th-century *again*'s).

These data were collected over a six-week period. While the crowdsourcing experiment spanned the entire semester (i.e. beyond six weeks) to include data from earlier stages of English, we do not have a viable amount of annotations yet for this subperiod of English.

For 39 (out of the 216) uses of *again*, we received a unanimous 'vote'. Out of the remaining 177 tokens, 15 needed a tie-breaker. To this end, the respective counts of available labels per token ('votes'), were adjusted based on (*i*) the degree of experience an annotator had when providing an annotation, (*ii*) the overall, formal quality of an annotation (based on the annotation guidelines), (*iii*) how far into the semester an annotation was performed, and (*iv*) an annotator's 'motivation' (i.e. how many annotations have been submitted overall). Each of these variables impacted the number of votes by a factor of 1.01 or 1.001 respectively. This stage of data processing allowed us to discern a final, crowdsourced reading for all 216 uses of *again*.

Contrasting these crowdsourced annotations with our gold standard, we calculated an overall (observed) accuracy of 83.33% (cf. table 1 for details). Due to our experimental design being crowd-based (with varying annotators & varying number of observations per item), we are not able to directly compute Cohen's kappa for inter-annotator agreement (see e.g. Poesio and Vieira, 1998). However, treating our crowd of annotators and our team of expert annotators as separate classifiers, we can calculate Cohen's kappa resulting in $\kappa = 0.70$ – indicating 'substantial' agreement at face value. Expected chance agreement, which Cohen's kappa is intended to correct for, is a problematic assumption regarding our data since considerable training and, thus, bias found its way into the annotations. This fact lends greater credibility to the observed agreement as a measure of accuracy. (McHugh, 2012) Consider Fig. 1 for the diachronic picture resulting from GS-data vs. CS-data. Both, Table 1 and Fig. 1 show only the numbers for the main readings (including 'other') and ignore a small number of ambiguous and 'ctd' cases. However, all relevant values shown have been calculated under inclusion of this small number of tokens (seven in total).

Contrary to our initial expectation, the factor 'experience' did not correlate with higher accuracy ratings. While this fact is somewhat discouraging as prolonged exposure to the task does not seem a viable means of triggering a training effect and consequently improving annotations, constant accuracies across all levels of experience also suggest that frequent annotator turnover should not affect reasonable accuracy rates too much – provided the annotators are equipped with a sufficient amount of training. The last point could be addressed with a tutorial video made available to an anonymous crowd of annotators in a later pilot for a production stage scenario.

While 83.33% overall accuracy is a reasonable degree of quality in crowd-based annotations, it is noteworthy that all 19th-century data has accuracies above this average whereas 18th-century averages are below 83.33%. This partial result suggests reduced confidence for diachronically older data. A possible avenue for increasing overall accuracy on the one hand and steering clear of the low-accuracy-slumps ('other' in 18th-century data w/ 62.5%) on the other hand may be to rely on distributional properties of *again* in order to filter out discourse-operator uses of *again*. Uses of *again* with this reading tend to occur in structurally higher positions, i.e. clause-initially. More targeted annotator selection and training could also lead to higher accuracy rates in older stages.

	18th c.		**19th c.**		**all**	
rep	56	87.50%	69	91.30%	125	89.60%
res_ct	36	75.00%	29	86.21%	65	80.00%
other	8	62.50%	11	90.91%	19	78.95%
all	102	80.39%	114	85.96%	216	83.33%

Table 1: GS-counts (N) & CS-accuracy (%)

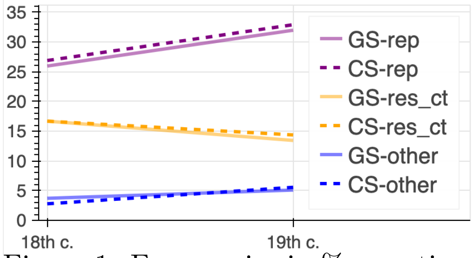

Figure 1: Frequencies in % over time, GS vs. CS

References

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

Delin, J. (1992). Properties of *It*-cleft presupposition. *Journal of Semantics*, 9(4):289–306.

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

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-Helsinki Parsed Corpus of Modern British English (PPCMBE2)*. Department of Linguistics, University of Pennsylvania, second edition. Release 1.

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

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

Spenader, J. (2002). *Presuppositions in spoken discourse*. Dissertation, University of Stockholm.