Decomposing decomposition in time: a methodological investigation

Martin Kopf, Mohammad Bablli, Maryam Rajestari, Dana Rebar, Remus Gergel (Saarland University)
{ martin.kopf@ | mohammad.bablli@ | s8maraje@stud. | dana.rebar@szsb. | remus.gergel@ } uni-saarland.de

1. Background

The goal of this empirical contribution is to present recent output generated by the DFG-funded project 'Decomposing Decomposition in Time' (Gergel 2020). We focus on the expert annotation of theoretically relevant ambiguities in readings of historical data on decomposition (Beck et al. 2009, Gergel & Beck 2015, Gergel et al. 2016, Gergel & Nickles 2019). 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 Kopf & Gergel (2023), 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 their own right: 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. Rapp & von Stechow 1999, Beck 2005, Zwarts 2019, among others). Moreover, recent inquiries into diachronic formal semantics indicate that diachronic data are also able to elucidate synchronic debates that could not be solved otherwise thus far (cf. e.g., Degano & Aloni 2022 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. We start off with a brief introduction to the English adverb *again* before we discuss three approaches to closing the empirical gap in diachronic data. These approaches are as follows: First, we discuss the English adverb *again* and 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 (Kroch et al. 2000) covering the Middle English (ME) period, the PPCEME (Kroch et al. 2004) and PPCMBE2 (Kroch et al. 2016), covering the Early and Late Modern Eng. periods, respectively (EModE, LModE). Both the EModE and LModE data have received an exhaustive annotation (1,901 and 1,532 *again*s). For the ME data, this report is based on an 81.5% completion rate (771/946 *again*s). The first portion of this semantic annotation, i.e., all 1,901 EModE 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 (Kopf & Gergel 2023) 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 preview most of the data from both EModE and ME (ca. 1150 CE onward). As far as alternatives to an expert-based annotation are concerned, we will summarize the performance of the following approaches: The second method discussed in this abstract thus seeks to extend on the expert-based annotations from a computational perspective. From the perspective of a metanalysis aimed at identifying the best-performing feature-sets, we report on the performance of a Multinomial Naïve Bayes classifier in predicting the main readings of *again* in the LModE data. The third and final part of this submission summarizes a 'informed crowdsourcing' experiment, which was designed to explore aptitude for providing nuanced semantic annotations on diachronic data. Thus, the crowd workers were to work with natural language data for which they cannot have any native speaker intuitions whatsoever. We report on the performance of KMeans clustering of the crowdsourcing data in comparison to our expert-provided gold standard. 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, an event of the same kind (*examining plants*) is presupposed (*rep*). The again in (2) is restitutive/counterdirectional, 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 (*res/ct*). The readings in (1) and (2) are the most frequent ones in the data discussed here and in line with the literature (cf. Gergel & Beck 2015). A third relevant reading of *again* are discourse-marker uses: Rather than operating on predicates, they have a discourse organizing function ('*dm*'). Other smaller readings of *again* exist in the historical data but are not reported here for the sake of brevity (labelled '*other*' in the discussion below).

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'), and (iii) marked the main verb of the antecedent satisfying a relevant PSP. Other categories were marked in absence of a verb (e.g., *Rain again [...]* cf. RUSKIN-1882-2,3,1019.286). Contextual material was still marked as antecedent – and additionally labelled 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 finetuned 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*-predicate/event and plausible antecedents as crucial. See (3) as EModE example (from Robert Boyle's *Experiments and considerations touching colours*; 1664) where *again* operates on the predicate *reduce to whiteness*. On a repetitive reading, the relevant presupposition would be satisfied in the context with *the mixture will appear white* (marked with double slashes around the main verb). However, the material encoding the counterdirectional presupposition (marked with double underscores) is closer to where we find the '*again*-event' encoded. Therefore, this use of *again* is to be classified as *res/ct*:

(3) [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, 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 disambiguate, consider (4), from George Adams' *Essays on the microscope* (1787). Prima facie, the material marked with double slashes seems to satisfy a repetitive presupposition, 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 presupposition:

(4) 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 3rd- and sometimes 4th-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 4,204 uses of *again*: 771 from PPCME2 (=81.5% of 946 total), 1,532 from PPCEME (100%), and 1,901 from PPCMBE2 (100%). 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*), as well as discourse marker uses (*dm*), and the above mentioned fourth class (*other*) (containing minor other readings and low-frequency occurrences of unresolvable ambiguity/unclear cases). Moreover, Tab 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* (Beck et al. 2009, Gergel & Beck 2015), which had been done on disparate corpora (i) solely based on correspondence and (ii) lacking the 18th century (currently the most general unified corpora are used, from which Tab 1 are examples).

subperiod	major readings				# *again*	
	rep	*res/ct*	*dm*	*other*	N	%
M1 (1150-1250 CE)	20.0	53.3	-	26.7	30	0.023
M2 (1250-1350 CE)	8.5	76.6	-	14.9	47	0.037
M3 (1350-1420 CE)	11.9	82.3	0.3	5.5	362	0.099
M4 (1420-1500 CE)	16.9	63.9	0.3	19.0	332	0.133
E1 (1500-1569 CE)	31.0	56.8	3.6	8.6	526	0.088
E2 (1570-1639 CE)	41.9	44.9	9.6	3.6	613	0.103
E3 (1640-1720 CE)	40.5	50.4	3.8	5.3	393	0.073
L1 (1700-1769 CE)	51.0	42.9	3.2	3.0	473	0.060
L2 (1770-1839 CE)	58.9	33.4	5.3	2.3	640	0.070
L3 (1840-1910 CE)	62.4	24.9	11.5	1.1	788	0.076

Table 1: Relative frequencies of major readings over time

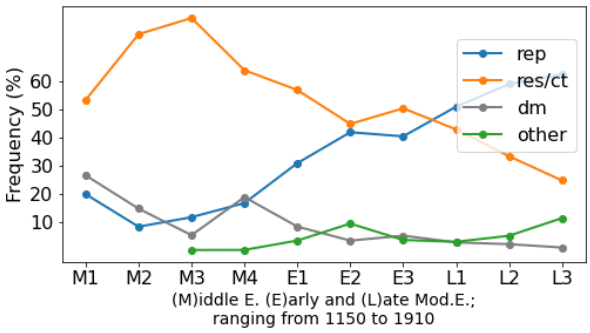

Figure 1: Relative frequencies of major readings

3. Classifying *again*s with a Multinomial Naïve Bayes classifier

3.1. Method: Based on a subset of the expert annotations introduced in section 2 (i.e., only LModE data) – together with a variety of features – Kopf & Gergel (2023) carried out a meta-analysis to find the most promising features in predicting

readings of *again* with a Naïve Bayes classifier. The data set of 1,901 LModE annotations was reduced to the 1,722 uses that represent either *rep* (64.4%) or *rec_ct* (35.6%) uses of *again*. 16 different features of three major distinct types were included: (i) "Naïve" features that can be drawn from the linear surface of the text material, (ii) annotational features as per our semantic annotation (modulo the classes of readings, i.e., the dependent variable), and (iii) structural features derived from the syntactic parsing of the treebank-formatted corpus data. The various features were modelled as count vectors in separate feature matrices for which all possible feature combinations were computed. As pretests, over each of the resulting 65,535 different combinations of features, 10 training-cycles of a Multinomial Naïve Bayes classifier were run (with a repeated and randomized 4:1 split between training and test data for validation) and extended to 100 train-test cycles if the pretest gave an accuracy above 77.5% (Pedregosa et al., 2011; Pustejovsky & Stubbs, 2012).

3.2. Results: An average accuracy of up to 81.46% was achieved in classifying uses of *again* as either *rep* or *res/ct*. A set of five core features is involved in most feature combinations performing at 81% or higher: *i*. antecedent verb, *ii*. target verb, *iii*. distance between antecedent material and *again*, *iv*. distance between *again* and target verb (encode precedence by including negative values), *v*. word forms in the again-clause (as delimited in the syntactic parse). For the performance to go beyond 81% accuracy, varying other features – often to the exclusion of one another – need to be included. The average accuracy of only the listed features combined is 80.67% (based on 100 train-test cycles, std.=2.13%).

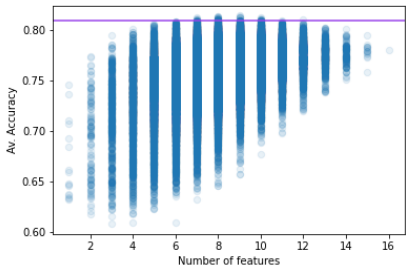

Figure 2: Average accuracy by number of features, over 10 or 100 train-test cycles respectively

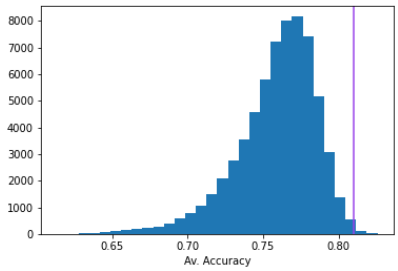

Figure 3: Histogram of average accuracies, over 10 or 100 train-test cycles respectively

4. Informed crowdsourcing pilot

4.1 Method: For this approach, suggested in Gergel (2020) and also explored in Kopf & Gergel (2023, English students were recruited as crowd workers from two lectures at Saarland university. The motivation for this course of action is due to the intricate nature of the annotation task, i.e., heavily context-dependent semantic annotations on historical language data (with potential antecedent material at varying distances to the PSP trigger – at times significantly greater than, for instance, reference resolution tasks for pronouns). The intention was to be able to receive direct feedback from the crowd and respond quickly to uncertainties. The informed crowd workers were provided with a heavily stripped and condensed version of our annotation guidelines, a practice data set, regular tutorial sessions and a recorded tutorial. Individualized data sets were distributed, each containing five uses of *again* on a weekly basis directly to students' inboxes (to minimize the possibility for teamwork). To avoid scarcity in the crowd-provided annotations, only a subset of the PPCMBE2 and the PPCEME data were used for this pilot, i.e., 328 *again*s. After the elicitation phase the crowd-provided annotations were prepared for analysis by vectorizing them (cf. Kopf and Gergel 2023 for further details).

4.2 Results: Different approaches for evaluating the crowd annotations were tested in contrast to the gold standard provided by expert annotators. The most successful approach, adjusting the crowd-provided annotation vectors with crowd quality metrics ("CrowdTruth"; cf. Aroyo and Welty, 2013a, 2013b, 2015) and relying on a on a KMeans algorithm for unsupervised classification (Pedregosa et al., 2011) resulted in 84.1% overall accuracy. The detailed results are in Table 2: The rows hold the gold-standard based readings; the absolute numbers ('N') represent the number of *again*s available respectively per reading and/or period. The corresponding percentages report the accuracies. In addition to per-period, per-century, and overall accuracies, Cohen's Kappa is in the bottom row. High accuracies for the *rep*-readings consistently obtain throughout all periods (around 90%). The lowest percentage accuracy obtains for the *res/ct again*s – especially in the older data (75.0%). It is predominantly the *res/ct*-reading that is responsible for a decreased overall accuracy of older data. See Kopf & Gergel (2023) for a more detailed discussion of the crowd sourced annotations and how this approach could be utilized to reduce workload for a team of expert annotators while at the same time ensuring a robust classification of the overall data.

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

Table 2: %-acc. CS data (w/ KMeans) for gold-std. classes

5. Conclusion

While providing an exhaustive annotation of decompositional items based on multiply reviewed expert judgments ('gold standard') has been our main intent, this costly process warrants an exploration of alternatives. At the current state of the technical possibilities explored, a gold standard cannot be substituted by either machine learning-based predictions or experimental data. The first upshot is that the gold standard itself must be as solid as possible (we sketched our detailed

approach above, and we are constantly open to improving it) for reliable training and evaluation of the approached sketched above and/or related. At the same time, we think that two additional case studies we have summarized are quite telling even if their performance had been expectedly lower. The significance of such extensions is obvious when it comes to the annotation of larger amounts of data. The feature-based approach (section 3) then becomes relevant, also for cases in which the syntactic annotation is missing such as the EEBO type of corpora in our object-language English. In such a case, some of the syntactic features that have been used in the approximations can be translated, e.g., in terms of precedence (an instance of *again* that precedes its modifying predicate is typically also higher in structure etc.) Overall, however, we believe that the human approach, i.e., the type of informed crowdsourcing presented is the most promising variant of annotational support when one strives to cover more data than one's team can handle or for gaining more certainty empirically. The straightforward advantage is that the relatedness in the languages at hand can be used even if the 'nativeness' of the actual participants is not available. Some of our results have indicated that more distant periods in time do not necessarily become worse in the annotational performance. On a conceptual level, there is also initial evidence from independent areas of semantic change (cf. Gergel, Kopf-Giammanco & Puhl 2021, Gergel, Puhl, Dampfhofer & Onea 2023) that speakers adapt astonishingly well in simulated situations of change. Finally, even if certain targeted readings are comparatively low performing, one can still place a crowdsourcing approach at the start of an annotation pipeline. By validating crowd annotations with a gold standard for a subset of the data, one can learn which data (i) need a closer review, (ii) which data need less attention in a review, and (iii) which data could benefit from a thorough review due to inherent indecisiveness of the crowd (section 4).

References:

Aroyo, L. and Welty, C. (2013a). Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In Web Science 2013, 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.

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. (2005). There and back again: A semantic analysis. *Journal of Semantics 22*: 3-51.

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

Degano, M., and Aloni, M. (2022). Indefinite and free choice: When the past matters. *Natural language and Linguistic Theory* 40: 447-484.

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

Gergel, R. 2020. Decomposing decomposition in time. DFG project proposal. Saarland University.

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.

Gergel, R., Kopf-Giammanco, M., and Puhl, M. (2021). Simulating semantic change: a methodological note. In *Proceedings of Experiments in Linguistic Meaning (ELM) 1*, Andrea Beltrama, Florian Schwarz, and Anna Papafragou (eds.). 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., Puhl, M., Dampfhofer, S., and Onea, E. (2023). The rise and particularly fall of presuppositions: Evidence from duality in universals. In *Proceedings of Experiments in Linguistic Meaning (ELM) 2*, Tyler Knowlton, Florian Schwarz, and Anna Papafragou (eds). 72-82. University of Pennsylvania: LSA.

Kopf, M., and R. Gergel. (2023). Annotating decomposition in time: Three approaches for again. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII*), pages 129–135, Toronto, Canada. Association for Computational Linguistics.

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.

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.

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

Pustejovsky, J., Stubbs, A. (2012). *Natural Language Annotation for Machine Learning*. O'Reilly.

Rapp, I. and von Stechow, A. (1999). *Fast* 'almost' and the Visibility Parameter for functional adverbs. *Journal of Semantics 16*: 149-204.

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