Cheap Annotation of Complex Information: A Study on the Annotation of Information Status in German TEDx Talks

Carmen Schacht   Tobias Nischk   Oleksandra Yazdanfar   Stefanie Dipper
Sprachwissenschaftliches Institut / CRC Information Density and Linguistic Encoding
Fakultät für Philologie, Ruhr-Universität Bochum
firstname.lastname@rub.de

Abstract

We present an annotation experiment for the annotation of information status in German TEDx Talks with the main goal to reduce annotation costs in terms of time and personnel. We aim for maximizing efficiency while keeping annotation quality constant by testing various different annotation scenarios for an optimal ratio of annotation expenses to resulting quality of the annotations. We choose the RefLex scheme of Riester and Baumann (2017) as a basis for our annotations, refine their annotation guidelines for a more generalizable tagset and conduct the experiment on German Tedx talks, applying different constellations of annotators, curators and correctors to test for an optimal annotation scenario. Our results show that we can achieve equally good and possibly even better results with significantly less effort, by using correctors instead of additional annotators.

1 Introduction

Information status concerns the way in which referents are referenced in a text: e.g. as a newly introduced entity (*a nice picture*), as a generally known entity (*the sun*), as a previously mentioned entity (*she*), etc. In language, information status is mainly reflected in the form of referring expressions, e.g. personal pronouns for a pre-mentioned entity or indefinite article for a newly introduced entity.

Investigating information status is a complex endeavor, as there exist various competing terminologies and classifications. In our work, we follow Riester and Baumann (2017) in their approach to the annotation of information status, applying the *RefLex* scheme, an annotation scheme encoding detailed information on contextual and extra-textual givenness of referents. The scheme covers both the referential and lexical dimensions of information status. Only the referential level is relevant to the work described in this study.

This work is part of a larger project on word order in German, investigating the influence of information status and information-theoretical factors such as surprisal and information density (Shannon, 1948). In particular, we are interested in the relationship between information status and information density. We therefore annotate data according to the RefLex scheme. Since the annotation of such a complex phenomenon requires expert annotators, it is rather costly in terms of time and personnel. Hence, we aim to find a more economical solution to the commonly expensive annotation and curation of information status.

In this paper we present the results of an annotation experiment that we conducted by testing various annotation scenarios for time and personnel efficiency as well as accuracy of the annotations. Specifically, we compare the traditional approach – multiple annotation and subsequent curation, which is usually considered a guarantee of high annotation quality – with a simpler approach in which a single annotation is subsequently corrected. Our results show that we can achieve equally good and possibly even better results with significantly less effort, by using correctors instead of additional annotators.

2 Related Work

Linguistic annotation is a corpus-linguistic method with a long tradition, where quality control plays an important role. Traditionally, the quality of annotations is measured using chance-corrected measures of inter-annotator agreement (IAA), also called inter-rater reliability (IRR), such as Fleiss' kappa or Cohen's kappa (Fleiss, 1971; Cohen, 1960; Carletta, 1996). These measures assume that two or more annotators annotate the same text independently of each other.

Another type of quality control is when only one annotator annotates the text and subsequently an expert annotator goes over these annotations and corrects them if necessary. In this case, the two versions – before and after correction – can be compared with each other applying measures such as F-score, measuring the accuracy of one version with regard to the other.

It can be assumed that fewer errors will be detected with this method than with multiple annotations. For example, the two large German-language treebanks were annotated according to these two paradigms: The first method – double annotation – was applied to the annotation of the TIGER treebank, the second method – annotation plus subsequent correction – to the annotation of TüBa-D/Z (Dipper and Kübler, 2017).

Grouin et al. (2014) evaluate the effect of differently-annotated types of training data (with double annotations, with a curated gold version, with an automatic pre-annotation that has been manually corrected) on the performance of a CRF classifier. In contrast to our approach, the annotation quality as such is not compared and evaluated directly, but indirectly, based on the performance of the trained system. Furthermore, in contrast to our experiment, they deal with a simple annotation task (identification of personal information in clinical documents).

A number of papers compare the quality of annotation with vs. without automatic pre-annotation; for an overview see, e.g., Mikulová et al. (2022). In contrast, we do not use automatic pre-annotation in our study.

3 The Data

The fragments that we annotated are extracts from the transcriptions of a total of five TEDx Talks which were given in German on a range of different topics. The texts are subject to licenses that permit free redistribution.[1]

From each talk, we annotated 100 referential expressions from two different sections of the talk, resulting in 10 fragments with 1,000 annotated units in total.[2]

We chose TEDx Talks for the annotation experiment as we considered them an adequate cross-section of content, while keeping the genre of the data – semi-scripted oral talks – constant. We excluded talks that involved particularities as for example rap, for this would distort the homogeneity of the dataset too much.

All annotations, curations, and corrections were created and handled in the annotation tool INCEpTION (Klie et al., 2018). Details of the procedure are provided in the following sections.

4 Annotation Guidelines

We base our experiments on the annotation of information status according to the RefLex scheme proposed by Riester and Baumann (2017). RefLex is a comprehensive annotation scheme that provides a total of 12 different labels, which can be divided into 7 classes, see Table 1. In addition, the features '+generic' and '+predicative' can be added to each expression. Markables are nominal phrases (NPs, incl. pronouns) and specific adverbs (e.g. *here*). If the NP is directly embedded in a prepositional phrase (PP), the entire PP is annotated. Possessive pronouns are also annotated.

In the following we describe the modifications we have made to RefLex. Table 2 provides an overview of the tags used in our study.

Label names Among other things, we have shortened the label names for the annotation. First, we omit the prefix 'r-' from all labels.[3] Second, we replace some of the longer names by short ones, see Table 3, e.g. `displaced` instead of 'r-given-displaced' or `known` instead of 'r-unused-known'.

Markables We define admissible markables as follows: A markable is either an NP (or PP, as specified in RefLex), a possessive pronoun or a deictic adverbial (*hier* 'here', *jetzt* 'now').

For complex phrases with embedded phrases, relative clauses or appositions, we annotate (i) the entire phrase (i.e. its head) and (ii) each of the embedded phrase(s).

Idioms are annotated as an entire span. Foreign language material is not considered, except for when it is referred back to. Incomprehensible passages, e.g. due to spelling mistakes or transcription errors, are ignored.

[1] The TEDx Talks are part of this playlist: https://www.youtube.com/playlist?list=PLzPiBVgAHXijVDasy92X6lZkl0DvFgSEg, accessed 2024-02-26. Our annotations are based on the subtitles extracted from these videos.

[2] The annotation guidelines as well as the annotated data are made freely available: https://gitlab.ruhr-uni-bochum.de/comphist/law25_infstat.

[3] The prefix 'r-' marks tags from the referential dimension rather than the lexical dimension of the RefLex tagset. As mentioned above, we only annotate the referential dimension, so the prefix is redundant information.

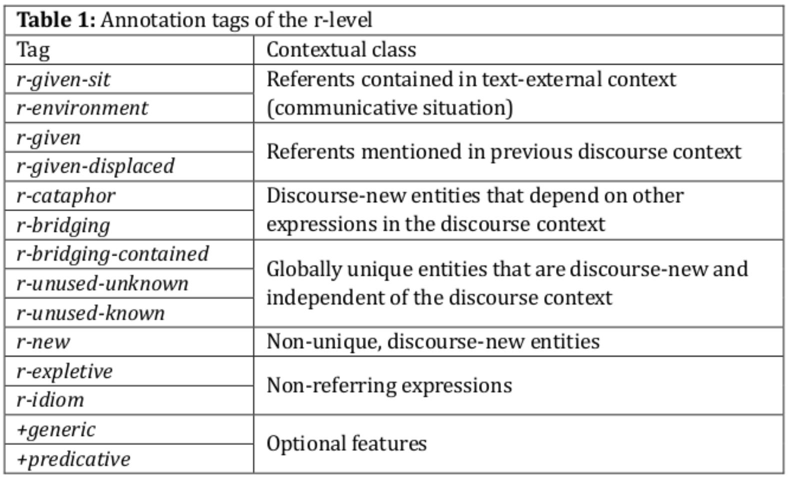


Table 1: Annotation tags of the r-level	
Tag	Contextual class
r-given-sit	Referents contained in text-external context (communicative situation)
r-environment	
r-given	Referents mentioned in previous discourse context
r-given-displaced	
r-cataphor	Discourse-new entities that depend on other expressions in the discourse context
r-bridging	
r-bridging-contained	Globally unique entities that are discourse-new and independent of the discourse context
r-unused-unknown	
r-unused-known	
r-new	Non-unique, discourse-new entities
r-expletive	Non-referring expressions
r-idiom	
+generic	Optional features
+predicative	

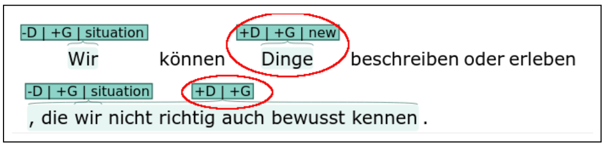
Table 1: Overview of the RefLex tagset (from Riester and Baumann, 2017, p. 9).

Label	**Form**	**Description**	**Examples**
new	indef, also complex	referent newly introduced; but may embed **given, known** etc.	***eine ganz andere Art der Freiheit***
given	def NP or pers/dem pron or pron adv or adv	referent mentioned before, possibly as text span	***sie***; ***da; dort; damals***; text span referent only in case of dem pron or pron adv: ***das*** *stimmt*; ***daran*** *denke ich oft*
bridging	not complex	referent mentioned before is a silent/implicit argument	
	1. def NP (no pron/adv)		***die Wohnung*** [(silent:) *in diesem Haus*]; ***diese Aussage*** [*nämlich dass …*]; ***das glücklichste Land*** [*von allen*]
	2. quantifying pron or NP		***alle/manche/niemand*** [*von denen*]; ***3l*** [*Milch*]
situation	1st or 2nd person, deictic	referent extratextual	***ich; dein; hier; jetzt***
cataphor	*es*; pron adv (only pron)	referent introduced subsequently	*denken* ***daran****, dass …*
known	def, not complex	1. encyclopedic knowledge	***der Papst***; locations; known persons
	def + indef	2. classes, always generic (**+G**)	***(die) Menschen*** *sind neugierig*; ***am Abend***; ***Löwen in Afrika***
unknown	def, complex	reference by description, everything **new** or **known**	***die Bilder von Vögeln***; unknown persons
contained	def, complex	containing embedded **given / bridging / situation / contained**	***seine Frau***; ***die Wohnung in diesem Haus***
displaced	def	referent mentioned more than 5 clauses ago	
expletive	*es; sich*	semantically empty expression	***es*** *gibt keinen Grund*; *ich erinnere* ***mich*** *an …*
idiom		does not introduce a referent, intransparent semantics	
noref		does not introduce a referent, transparent semantics	
	def + indef	1. formulaic incl. secondary prepositions	***zu Hause; vielen Dank; in jedem Fall*** ***an Hand; an Stelle; auf Grund; in Folge; mit Ausnahme***
	quantified	2. quantified adverbial expressions	***viel Zeit***
+generic (+G)		only in case of **new**, **given** and **known**	***ein Löwe*** *ist …*
+discontinuous (+D)		discontinuous constituent, incl. floating quantifier	***Dinge*** *machen,* ***die …*** ; ***das*** *ist auch* ***alles*** *sinnvoll*

Table 2: Overview and descriptions of the tags used in the annotation study.

RefLex Label	Our Label
r-given-sit, r-environment	situation
r-unused-known	known
r-ununsed-unknown	unknown
r-bridging-contained	contained
r-given-displaced	displaced
–	noref
+generic	+generic (+G)
–	+discontinuous (+D)

Table 3: Mapping between the original RefLex and our label names.



Wir können Dinge beschreiben oder erleben
, die wir nicht richtig auch bewusst kennen .

Figure 1: Annotation of example (1), featuring a discontinuous constituent (screenshot of INCEpTION).

Discontinuous constituents We added a special feature to mark constituents as discontinuous, as in (1). In the German original version, the relative clause is separated from its antecedent *Dinge* 'things'. In the annotation, the label `new`, which applies to the entire construction, is only annotated on the head noun *things*. In addition, the feature `+D` (for "discontinuous") is annotated at the head and at the relative clause, to mark them as one constituent.

(1) *Wir können Dinge beschreiben oder erleben, die wir nicht richtig auch bewusst kennen.*
'We can describe or experience things that we are not really aware of.'

Figure 1 shows the annotation for this example. The relevant annotations are highlighted in red (the second highlighted annotation `+D|+G` refers to the entire relative clause, whose words are marked in light green). The default value `-D` is automatically added by INCEpTION.

Generic In addition to the label `+/-D`, there is another special feature in Figure 1: `+/-G`, which stands for "+/-generic". Its default value is `-G`, but has been changed by the annotator for all the markables shown in the example, as *wir* 'we' refers to human beings in general in this example.

Note that we do not evaluate the annotations of these extra features `+/-D` and `+/-G` in our experiments.

Merging two labels RefLex distinguishes the two labels 'r-given-sit' and 'r-environment': Both refer to expressions for referents that are present in the immediate text-external context. 'r-environment' expressions additionally involve a deictic gesture (e.g. *this chair*), whereas 'r-given-sit' expressions do not (e.g. *I, we*). This distinction cannot always be made clearly without knowledge of the extra-textual context.

In (2), for example, it is conceivable that a picture or film of the supermarket and in particular of the fruit in the supermarket was shown during the TEDx Talk and the speaker pointed to the picture while uttering the phrase *this fruit* (highlighted in the English translation of the example). On the other hand, the phrase could also be understood as referring to the subsequent description.

(2) *Als erstes bin ich in einen Supermarkt gegangen und habe mir Obst angeschaut und dieses Obst gefunden: Obst, einzeln verpackt, weil Birnen und Äpfel sind ja tatsächlich schwer zu trennen.*
'The first thing I did was go to a supermarket to look at fruit and found *this fruit*: Fruit, individually wrapped, because pears and apples are actually difficult to separate.'

Hence, we abandon the distinction and keep one label `situation` for both RefLex labels.

New label We define a new label called `noref`, which is part of the class of non-referring expressions. Like idioms and expletives, such expressions do not introduce a referent. However, whereas the label `idiom` marks semantically intransparent spans, the new `noref`-label captures semantically transparent instances, such as *vielen Dank* 'thanks a lot', *zu Hause* 'at home', or so-called secondary prepositions like *auf Grund* 'due to; by reason of' or *mit Ausnahme* 'with the exception'.

Even though adding new labels always adds to the complexity of the tagset and thereby increases the risk of annotation errors, the addition of the `noref` label was judged to cover a relevant portion of information previously unaddressed and is therefore warranted.

Form-based characteristics We have enriched the definitions by consistently referring to possible

Form	Def	Examples
Articles		
Indefinite	indef	*ein Rad*
None	indef	*Räder*
Definite	def	*das Rad*
Demonstrative	def	*dieses Rad*
Possessor	def	*mein/Ottos Rad*
Quantifiers	def	*alle Räder; jedes Rad*
Quantifiers	indef	*keine/viele Räder*
Pronouns		
Demonstrative	def	*das; dieses*
Pronominal adv	def	*daran*
Indefinite	indef	*jemand*

Table 4: Forms of articles and pronouns and corresponding type of definiteness (column 'Def').

forms of the referring phrases, to facilitate annotation decisions and render them more robust against errors. In particular, the definitions have a strong focus on the form of the article, if any, or the type of pronoun or adverb, see Table 2, column 'Form'. Moreover, we added detailed definition of definiteness, see Table 4.

We also specified additional criteria for the labels `bridging`, `contained`, `unknown` and `known`, to allow for an easier distinction between those labels, see Table 2, column 'Description'.

Decision hierarchy There are often several options for annotating a phrase. For example, the second occurrence of *wir* 'we' in example (1) can be annotated either as `situation` or as `given` (because it has been mentioned previously). Similar cases often occur with referents labeled as `known` which are referenced multiple times.

Our guidelines specify that the label `given` (and `displaced`) should generally be annotated in preference, resulting, e.g., in coreference chains such as `unknown-given-given` or `known-displaced`. There are two exceptions to this rule: First, regarding the label `situation` as in (1), all coreferent occurrences are annotated as `situation`, cf. Figure 1. Secondly, generic *man* 'one/you/they' is always annotated as `known`.

Linguistic tests We define linguistic tests to aid the annotation decision process. These tests concern mainly the decision whether an expression is considered to refer to a class or to individuals. This is realized by testing whether the expression refers to every single member of the assumed class or to a subset of individuals.

For example, if we want to annotate the phrase *modernster Methoden* 'state-of-the-art methods' in example (3), we can ask the following test question: Does this apply to every single state-of-the-art method? In the example, however, we are dealing with a contextually restricted subset of methods (which are relevant for virtual worlds), so `known` (for a known class) is not used, but `new` for a newly introduced subset.

(3) *virtuelle Welten helfen uns, unsere Wahrnehmung, unsere menschliche Wahrnehmung, zu stärken mit Hilfe modernster Methoden und Techniken.*
'virtual worlds help us to strengthen our perception, our human perception, with the help of *state-of-the-art methods* and techniques.'

5 Experiments

Annotation and curation of linguistic resources is time consuming and costly, especially in the case of a complex phenomenon like information status and a detailed tagset such as the RefLex scheme. To keep annotation costs minimal, we conducted an annotation experiment to test for an optimized annotation mode, which allows for minimal costs in resources and maximal accuracy. We assumed that the expenses of the usual annotation and curation process, involving multiple annotators and curators, could be reduced significantly by installing different settings of annotation while maintaining a reasonable accuracy and therefore quality of the annotated data.

To test this, we set up various annotation scenarios in different personnel settings and tested for time and staff 'costs' in relation to the resulting annotation quality. There were four expert annotators (the authors) involved in the experiment. Before running the experiment, the annotators annotated and curated several passages in two training datasets for annotation training. All annotators were also involved in the fine-tuning of the annotation guidelines. After the training phase, the guidelines were finalized. Then the experiment was conducted. All annotations, curations, and corrections where created and handled in the annotation tool INCEpTION (Klie et al., 2018).

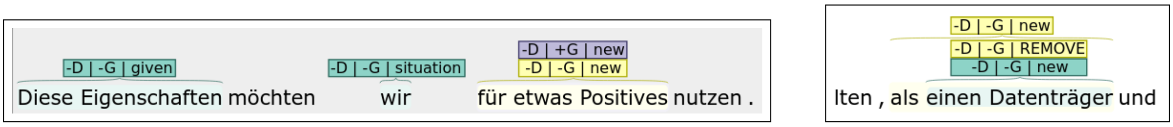


Figure 2: Original annotations and and corrections of example (4) (left), and a REMOVE correction, marking the erroneous span in example (5) (right).

Correcting annotations Figure 2 uses example (4) to show how we have implemented the correction steps in INCEpTION. The annotations shown in green are those of the annotator. The labels in yellow and purple come from two correctors.

For the correction steps, new layers (with new colors) were created in INCEpTION, with the same labels as the original annotation layer plus an additional label REMOVE (see below). The correctors could only see the original annotations of one annotator and not the corrections of the other corrector.

(4) *Diese Eigenschaften möchten wir für etwas Positives nutzen.*
'We want to use these qualities *for something positive*.'

Figure 2, left part, shows that the two existing annotations of example (4) were found to be correct by both correctors, so they didn't change anything. However, the phrase *for something positive* 'for something positive' was not considered by the annotator. Both correctors (shown in yellow and purple) have re-annotated this phrase.[4]

Removing an erroneous annotation of correcting the extent of an annotation span is a special case in the correction process. For this case, a new label REMOVE is employed, which is used to mark the incorrect span. A new correct span including a label is added, if needed. Figure 2, right part, shows the annotation of example (5). The original annotator did not include the preposition *als* 'as' in the span, which has been corrected accordingly by the corrector (shown in yellow).

(5) *als einen Datenträger*
'as a data storage medium'

Experimental settings The experiment included three different annotation settings (also see Table 10 in Appendix A for an overview of these settings):

Set 1 First, all four annotators annotated and collectively curated a gold version of 5 × 100 annotations.

Set 2 Secondly, only three of the annotators annotated and curated 2 × 100 annotations, and a single corrector corrected the annotations of one of the three annotators per batch.

Set 3 The last setting involved two annotators annotating and curating the gold version and the other two both correcting the same single annotation per batch, but independently from each other. In total, 3 × 100 annotations were annotated, curated and corrected in this setting.

The gold versions were created by the annotators themselves in a joint discussion round. This means that the gold versions are certainly influenced by the existing annotations, but this is trivially true for every gold version that is created on the basis of existing annotations.

The correctors did not participate in the curation. They only saw one of the annotations and corrected this annotation. They had no access to the other annotations or to the gold version.

So the relevant question is: Can the correctors arrive at a similarly high-quality "gold" result as the curators? Since a correction is significantly cheaper than a curation (requires less time and personnel), this would save a lot.

In order to make the two basic scenarios – multiple annotation followed by curation on the one hand vs. single annotation followed by correction on the other – as comparable as possible, the correction is based on one of the annotations that is also used to create the gold version (as one of several annotations).

6 Results

To evaluate the quality of the various annotation scenarios, we use two different measures: Fleiss' kappa as a measure of inter-annotator agreement and F_1-score as a measure of the annotators' and

[4] As already noted, we ignore differences regarding the labels +/-D and +/-G.

Set	Labels (κ)	Spans (%)
1	0.63	73.58
2	0.73	67.67
3	0.76	88.19

Table 5: Inter-annotator agreement: Fleiss' kappa for exact matching spans and proportion of matching spans across the different settings.

correctors' accuracy with regard to the gold standard and as a measure for the correctors' agreement among them.[5]

Agreement among the annotators We first analyzed agreement between the annotators, see Table 5. Only spans that were exact matches were included in the evaluation using Fleiss' kappa. The second column shows the proportion of these spans in all spans. The table already shows solid scores for the labels in the first phase, which increase continuously, indicating a robust baseline of inter-annotator scores for the further evaluation of the experiment.

Distance between annotations and gold Next, we examined how far the individual annotators were from the curated gold version. We calculated this distance in the form of aggregated F-scores across all annotated text fragments per annotator, see Table 6. Only exact matches were counted as correct. We distinguish between F-scores for spans and for labels, to differentiate between correctly identifying spans and subsequently labeling them correctly. The span scores were calculated as the harmonic mean of span precision and recall. The label scores are the micro-averaged harmonic mean of label precision and recall per person. As Table 6 shows, label F-scores range from 0.63 to 0.75 while span F-scores are considerably higher at 0.88 to 0.93, indicating a relatively robust span identification across annotators, while label identification seems to pose some challenges.

For us, a highly relevant question is how far away the results from the different tasks are from the optimal gold version. In other words, we want to compare two distances: (i) How far are the individual annotators from the curated gold version? (ii) How far are the corrected versions from the gold version? If the corrected versions are further away from the gold version, this would mean that the corrections have introduced additional errors and worsened the annotation overall. The expectation would therefore be that the corrected version is as close as possible to the gold version, so that a correction can serve as a substitute for an elaborate double annotation with subsequent curation.

Question (i) has been answered above (see Table 6). Question (ii) is addressed next.

Annotator	Labels (F_1)	Spans (F_1)
Person1	0.75	0.93
Person2	0.70	0.93
Person3	0.64	0.88
Person4	0.63	0.88

Table 6: Annotator vs. gold: F_1-scores for labels and spans between each annotator and the curated gold version.

Corrector	Labels (F_1)	Spans (F_1)
Person1	0.75	0.92
Person2	0.81	0.95
Person3	0.79	0.93
Person4	0.86	0.96

Table 7: Corrector vs. gold: F_1-scores for labels and spans between each corrector and the curated gold version.

Distance between corrections and gold For the evaluation of the corrected labels, we also used an absolute match heuristic, where only exact matches were counted as correct. However, to account for the fact that spans could be added or removed by the correctors, we introduce an additional label called `NONE`, which covers two possible scenarios: (i) A span was added by the corrector but does not exist in the gold standard (gold = `NONE`, correction = `foo`). (ii) A span in the gold standard was omitted by the corrector (gold = `foo`, correction = `NONE`).[6]

[5] In our view, chance-corrected measures such as Fleiss' kappa are not applicable to the other scenarios because it is to be expected that the gold version as well as the corrector's version are biased by the given annotations and therefore the assumptions concerning chance agreement are no longer correct.

[6] This approach also allows us to also account for cases in which the extent of a span has been corrected (as shown in Figure 2, right part), in that `REMOVE` annotations are treated as

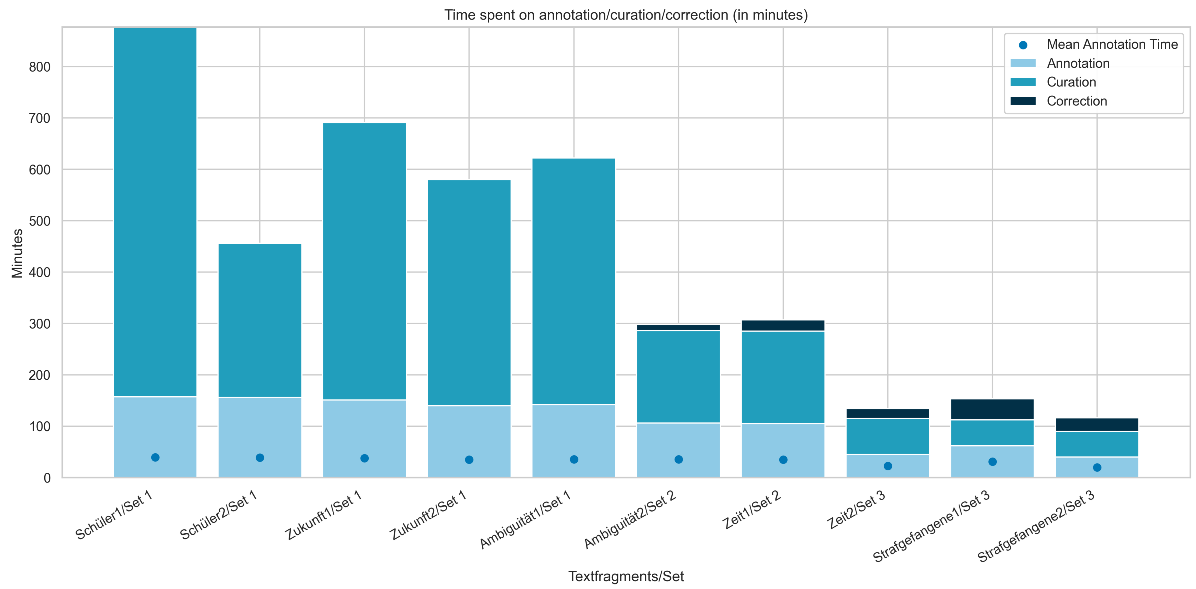


Figure 3: Accumulated annotation, curation and correction times per text fragment. Note that total annotation time represented in the bars decreases substantially due to employing fewer annotators per scenario, but average annotation time stays relatively constant.

Task	**Labels (F_1)**	**Spans (F_1)**
Annotation	0.68	0.91
Correction	0.80	0.94

Table 8: Annotation vs. correction: macro-average of the annotation and correction F_1-scores for labels and spans.

	Labels (F_1)	**Spans (F_1)**
Correctors	0.95	0.97

Table 9: Corrector vs. corrector: F_1-scores for labels and spans between the correctors.

For comparing corrections with the gold version, we calculated span and label F-scores for each individual corrector across all corrections, see Table 7. The table shows that practically all F-scores are substantially higher than the F-scores of the original annotators in both span and label identification.

Table 8 shows the macro-averaged F-scores of both tasks. The F-scores of the correction task clearly outperform the overall annotation scores, indicating an increase in data quality for the correction scenario as compared to the usual annotation setting of multiple annotations and subsequent curation.

Agreement among the correctors Finally, we also compared the correctors with each other using the F_1-score, by considering one of the correctors as the “gold” version to which the other corrector is compared. As above, the span scores were calculated as the harmonic mean of span precision and recall and the label scores as the micro-averaged harmonic mean of label precision and recall, see Table 9 for the results. Both label and exact span agreement are exceptionally high, indicating highly consistent identification of relevant text spans and similar interpretive strategies.

Comparing time and personnel across the scenarios To evaluate the influence of the various annotation settings on time and personnel spent on the annotation process, all annotation, curation and correction times were tracked, see Figure 3 for the respective settings and measured times.

The bars encode the accumulated time required per text. The different settings include either annotation plus curation (Set 1), or annotation, curation plus correction in different weightings (Sets 2 and 3). Average annotation time is marked by a blue dot within the columns.

NONE annotations.

The first five bars represent the accumulated time requirements for annotating (light blue) and curating (azure) the text fragments in Set 1, by four annotators and curators. That is, the lower part of these bars shows the sum of the four individual annotation times and the upper part of the bars shows the curation time multiplied by four (because four curators were involved). The time requirements shown therefore correspond to the personnel costs that would have to be invested.

The next two bars show the total time of Set 2, comprising three annotators and curators plus one corrector (midnight blue). The final three bars represent Set 3, with only two annotators/curators and two correctors. Note that this is the minimal amount of annotators/curators necessary to realize traditional annotation and curation.

As expected, the overall time is trivially reduced significantly from setting to setting (as fewer people are involved in the annotation and curation per setting). In addition, a training effect can be observed during curation: every second text fragment from the same text is curated faster than the first (e.g., compare the curation time of the first and second bar or of the third and fourth bar). The curation time also appears to be decreasing in general, although this may also be an effect of the respective texts.

However, Figure 3 also shows that the average annotation time (the blue dots) stays relatively constant. This shows that, in contrast to curation, there is practically no training effect with annotation, or only a marginal one.

Set 3 is the setting in which the time required for the conventional annotation setting – involving 2 annotators + joint curation – can best be compared directly with the correction setting, involving 1 annotator + 1 corrector. Figure 4 relates the two alternatives directly to each other. The left column of each pair shows the accumulated time for two annotators (light blue) and the curation time multiplied by two (azure). The right column of each pair shows the sum of the average annotation time (blue) and the average curation time (midnight blue). The comparison clearly shows the drastic time gain due to the correction setting.

Considering that the F-scores for span and label identification in the correction setting not only stay constant between the conditions of annotation/curation and annotation/correction, but even increase, the annotation costs saved in terms of time and personnel are considerable.

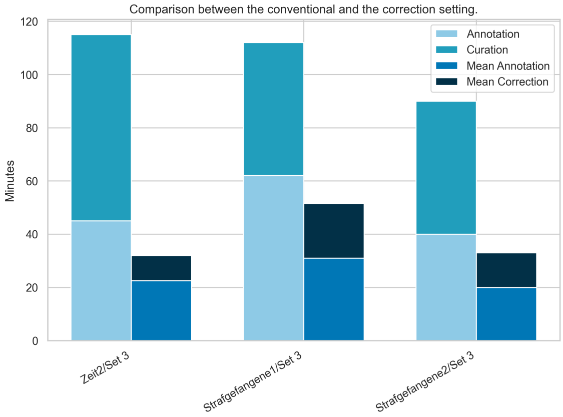


Figure 4: Comparison of accumulated time required by the conventional setting (left bars) and the correction setting (right bars).

7 Conclusion

We set out to investigate various annotation scenarios and their respective efficiency in terms of time and personnel employed and conducted an annotation mode experiment where we compared the scenarios of (i) four annotators and four curators, (ii) three annotators and three curators tested against a single corrector and finally (iii) two annotators and two curators tested against two correctors.

As has been shown in Section 6, the F-scores for span and label identification of the correctors not only stayed constant compared to the annotator F-scores, but even exceeded those annotators' values while reducing the total time of the entire annotation process approximately by half, even when considering the control curation condition in this calculation. We therefore argue that the third scenario of annotating and correcting is preferable to the conventional annotation and curation setting not only in terms of time and personnel, but also in terms of annotation quality, as the corrections closely match the gold version as can be inferred from the respective F-scores. We could thus show that time-efficient annotation – even in the case of highly complex tagsets such as the RefLex tagset – does not necessarily need to come at the traditionally high annotation cost.

Limitations

The study is based on data from only one type of text, TEDx Talks, and on only one type of annotation, information status. Overall, a rather small amount of data (1000 annotations from 5 different texts) was annotated. Whether the same or similar results can be obtained for other text and annotation types is an open question.

All annotators were involved in all parts of the study from the beginning and contributed to the development of the guidelines as well as annotating, curating and correcting data themselves. The significance of the study would have been stronger if these tasks had been carried out by different experts, for example if the developers of the guidelines had not annotated the data.

Since all annotators were directly involved in the development of the annotation guidelines as well as in the annotation, curation and correction processes, a marginal training effect may have positively influenced the overall annotation quality. Compared to a setup involving separate teams for annotation, curation, and correction, the resulting quality metrics may be slightly elevated. Nevertheless, the relatively stable mean annotation time across tasks highlights the substantial efficiency gains achieved through the integrated correction settings. These gains represent a notable improvement over conventional annotation workflows that rely on multiple independent annotations followed by subsequent curation – both in terms of time investment and the resulting data quality.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful comments. Funded by/gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 232722074 – SFB 1102 (Project C6).

References

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

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

Stefanie Dipper and Sandra Kübler. 2017. German treebanks: TIGER and TüBa-D/Z. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 595–639. Springer, Berlin.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):387–382.

Cyril Grouin, Thomas Lavergne, and Aurélie Névéol. 2014. Optimizing annotation efforts to build reliable annotated corpora for training statistical models. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 54–58, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Marie Mikulová, Milan Straka, Jan Štěpánek, Barbora Štěpánková, and Jan Hajic. 2022. Quality and efficiency of manual annotation: Pre-annotation bias. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2909–2918, Marseille, France. European Language Resources Association.

Arndt Riester and Stefan Baumann. 2017. *The RefLex Scheme — Annotation Guidelines*, volume 14 of *SinSpeC — Working Papers of the SFB 732 "Incremental Specification in Context"*. OPUS, Stuttgart.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

A Appendix

Setting / Text	**Person1**	**Person2**	**Person3**	**Person4**	**Distribution of Tasks**
Set 1					
Schüler-1	100 ann+cur	100 ann+cur	100 ann+cur	100 ann+cur	4 anno / 4 cur
Schüler-2	100 ann+cur	100 ann+cur	100 ann+cur	100 ann+cur	4 anno / 4 cur
Gesellschaft-1	100 ann+cur	100 ann+cur	100 ann+cur	100 ann+cur	4 anno / 4 cur
Gesellschaft-2	100 ann+cur	100 ann+cur	100 ann+cur	100 ann+cur	4 anno / 4 cur
Ambiguität-1	100 ann+cur	100 ann+cur	100 ann+cur	100 ann+cur	4 anno / 4 cur
Set 2					
Ambiguität-2	100 ann+cur	100 ann+cur	100 ann+cur	**100 corr**	3 anno / 3 cur / 1 corr
Zeit-1	100 ann+cur	100 ann+cur	100 ann+cur	**100 corr**	3 anno / 3 cur / 1 corr
Set 3					
Zeit-2	**100 corr**	100 ann+cur	100 ann+cur	**100 corr**	2 anno / 2 cur / 2 corr
Strafgefangene-1	100 ann+cur	**100 corr**	100 ann+cur	**100 corr**	2 anno / 2 cur / 2 corr
Strafgefangene-2	100 ann+cur	100 ann+cur	**100 corr**	**100 corr**	2 anno / 2 cur / 2 corr

Table 10: Detailed overview over annoation, curation and correction scenarios. ‘Person1’ to ‘Person4’ shows the tasks of the four expert annotators in the respective settings. ‘100 ann+cur’ means that this person created 100 annotations (independently of the others) and then curated the gold version together with the other annotators. This means that four people were involved in annotating and curating (‘4 anno / 4 cur’, column ‘Distribution of Tasks’). From Set 2 onwards, Person4 no longer annotated and curated, but instead corrected the 100 annotations of one of the annotators (‘100 corr’). From Set 3 onwards, two people corrected the same 100 annotations of one annotator, independently from each other.