Normalization in Context: Inter-Annotator Agreement for Meaning-Based Target Hypothesis Annotation

Adriane Boyd
Department of Linguistics
University of Tübingen
adriane@sfs.uni-tuebingen.de

Abstract

We explore the contribution of explicit task contexts in the annotation of word-level and sentence-level normalizations for learner language. We present the annotation schemes and tools used to annotate both word- and sentence-level target hypotheses given an explicit task context for the Corpus of Reading Exercises in German (Ott et al., 2012) and discuss a range of inter-annotator agreement measures appropriate for evaluating target hypothesis and error annotation.

For learner answers to reading comprehension questions, we find that both the amount of task context and the correctness of the learner answer influence the inter-annotator agreement for word-level normalizations. For sentence-level normalizations, the teachers' detailed assessments of the learner answer meaning provided in the corpus give indications of the difficulty of the target hypothesis annotation task. We provide a thorough evaluation of inter-annotator agreement for multiple aspects of meaning-based target hypothesis annotation in context and explore measures beyond inter-annotator agreement that can potentially be used to evaluate the quality of normalization annotation.

1 Introduction

Learner language frequently contains non-canonical orthography and morphosyntactic constructions that present difficulties for natural language processing tools developed for standard language. Since manually annotated learner corpora are often small and the high degree of variation in learner productions leads to data sparsity issues even for larger learner corpora, it is useful to consider methods that normalize non-standard aspects of learner language. While normalization and applying standard language categories to learner language does not address the full spectrum of learner language analysis and fundamental concerns about analyzing learner language (cf. Meurers and Dickinson, 2017), it can facilitate access to learner language in applications such as corpus search tools and computer-aided language learning systems.

Normalizations such as the minimal target hypothesis from the Falko German learner corpus (Reznicek et al., 2012) have been developed in order to provide a version of a learner production that can be systematically searched and that is more appropriate for further manual or automatic analysis. The minimal target hypothesis contains a minimal number of modifications that convert the learner sentence into a locally grammatical sentence. As it may not be possible to determine exactly what the learner intended to say in an open-ended task such as an essay task, what constitutes a minimal change is based on grammatical properties, e.g., preserving a verb and modifying its arguments rather modifying the verb itself.

In terms of the difficulties an annotator may face while interpreting a learner utterance, consider the following learner utterance from the Hiroshima English Learners' Corpus (Miura, 1998):

(1) I don't know he live were.

It is possible to speculate about the intended meaning of this utterance, proposing multiple interpretations such as:

(2) a. I don't know if he was alive.
b. I don't know where he lives.

Then consider (1) again within the task context: a translation task from Japanese into English of a sentence with the meaning *I don't know where he*

Adriane Boyd 2018. Normalization in context: Inter-annotator agreement for meaning-based target hypothesis annotation. *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018 (NLP4CALL 2018)*. Linköping Electronic Conference Proceedings 152: 10–22.

lives. This task context makes it extremely likely that the intended meaning is that of (2b).

Without annotation guidelines based on detailed grammatical properties such as for Falko, target hypothesis annotation and likewise error annotation for learner language in open-ended tasks has been shown to be difficult to perform reliably (e.g., Fitzpatrick and Seegmiller, 2004; Lüdeling, 2008; Tetreault and Chodorow, 2008; Lee et al., 2009; Rosen et al., 2013; Dahlmeier et al., 2013). As an example, Dahlmeier et al. (2013) report Cohen's κ of 0.39 for the task of identifying which tokens should be edited in the NUCLE corpus of English student essays.

In contrast to open-ended tasks, a more explicit task context can provide more information about the potential meaning of a learner production (Meurers, 2015), thereby facilitating a more reliable interpretation of the form and meaning and thus more reliable annotation of target hypotheses, which preserve the intended meaning instead of prioritizing particular grammatical features. For the ComiGS corpus, which contains explicit task contexts in the form of comic strips used in picture description tasks, Köhn and Köhn (2018) report $\kappa = 0.86$ for the same task of identifying which tokens should be normalized.

In this paper, we systematically explore the dependence of normalization on task context through manual annotation studies, focusing on L2 learner responses in a reading comprehension task context. We explore inter-annotator agreement measures for normalization and error annotation, considering the use of related evaluation metrics beyond inter-annotator agreement for the direct evaluation of normalization annotation.

2 Background

Numerous manual annotation studies have shown that target hypothesis annotation is difficult to perform reliably, and since error annotation depends on the formulation of target hypotheses (cf. Hirschmann et al., 2007), inter-annotator agreement for error annotation has likewise had lower levels of reliability (e.g., Fitzpatrick and Seegmiller, 2004; Lüdeling, 2008; Tetreault and Chodorow, 2008; Lee et al., 2009; Rosen et al., 2013; Dahlmeier et al., 2013). For example, a detailed annotation study for the CzeSL corpus of L2 Czech shows a wide range of inter-annotator agreement results for the presence of different types of error tags (Rosen et al., 2013), from $\kappa > 0.6$ for *incorrect stem*, *incorrect inflection*, and *incorrect word boundary* to $\kappa < 0.2$ for *ill-formed complex verb forms* and *incorrect pronominal references*. The authors perform a detailed inspection of the *agreement* errors ($\kappa = 0.54$) that reveals that half of the disagreements correspond to differing target hypotheses, where the annotators provided the correct error tags for their respective target hypotheses, but since these differ, the error annotation is inconsistent.

2.1 Task Context

For the contribution of available task context with respect to inter-annotator reliability, several studies on normalization annotation for in both L1 and L2 task contexts report promising results. In a native language setting, Lee et al. (2009) investigate annotators' judgments of article/number selections for English nouns in a sentence containing noun phrase gap. Annotators choose which noun article (*a/an*, *the*, no article) and number combinations (*singular*, *plural*) are possible in this context. For the five possible categories, Cohen's κ increases from $\kappa = 0.55$ to $\kappa = 0.60$ when the available context increases from the current sentence with the gap to include five preceding sentences. In addition, κ increases when the noun has already been mentioned in the context and for those article/number combinations that are much more frequent overall, e.g., the article/number combination *the sun* is much more frequent than all other article/number combinations involving *sun*, so annotators are more consistent about their decisions for *the sun* than less frequent combinations.

In an L2 setting, promising interannotator agreement results are reported for the ComiGS (Comic Strips Retold by Learners of German) corpus (Köhn and Köhn, 2018), an L2 German learner corpus where learners write descriptions of stories presented without accompanying text in comic strips. The corpus is manually annotated with minimal and extended target hypotheses largely following the Falko guidelines (Reznicek et al., 2012), and in contrast to previous studies of target hypothesis annotation in learner corpora, the ComiGS corpus includes an explicit context in which to interpret the learner productions. For the identification of which tokens need to be modified in the minimal target hypothesis in ComiGS, they report $\kappa = 0.856$ and for the extended target

hypotheses $\kappa = 0.74$ (cf. $\kappa = 0.39$ for NUCLE (Dahlmeier et al., 2013), although clearly many differences between the annotation studies make a direct comparison difficult).

2.2 Inter-Annotator Agreement for Normalization Annotation

Evaluations of inter-annotator agreement for normalization annotation are typically performed for several perspectives on the manual annotation. As an example of some possible evaluations, the NUCLE corpus (Dahlmeier et al., 2013), which contains both normalizations and associated errors tags, presents inter-annotator agreement results for three aspects:

- **Normalization identification**: Do annotators agree on which tokens are normalized?
- **Error tag given norm. identification**: For those tokens where both annotators agree that a modification is needed, do they agree on the error tag assigned?
- **Error+norm. given norm. identification**: For those tokens where both annotators agree that a modification is needed, do they agree on both the error tag and the normalization?

As an alternative to examining only those cases where both annotators agree that a modification is necessary, which excludes many potentially interesting cases where annotators disagree about whether to make a modification in the first place, the CzeSL inter-annotator agreement evaluation (Rosen et al., 2013) considers each error tag separately as a binary annotation task:

- **Error tag identification**: For a given error category, do annotators agree on which tokens are annotated with this category?

Both Dahlmeier et al. (2013) and Rosen et al. (2013) report the agreement coefficient Cohen's κ (Cohen, 1960), which measures agreement for categorical annotation tasks for two annotators. Cohen's κ (Cohen, 1960) and Krippendorff's α (Krippendorff, 1980) are frequently used inter-annotator agreement measures for evaluating binary or categorical annotation decisions, e.g., *Is a token modified?* or *Is a token annotated with category X?*. Inter-annotator agreement measures estimate how likely it is that annotators agreed (for κ) or disagreed (for α) by chance and calculate the degree of agreement beyond the level expected by chance alone.

The values for both Cohen's κ and Krippendorff's α range from -1 (perfect disagreement) to 1 (perfect agreement) with 0 as chance agreement only. Cohen's κ is limited to nominal categories (all disagreements are counted equally) and only two annotators, while Krippendorff's α has the advantages that three or more annotators can be included and that not only nominal categories but also annotations on ordinal or interval scales or with sets of categorical tags can be compared more precisely. See Artstein and Poesio (2009) for a detailed overview of inter-annotator agreement for linguistic annotation.

As explored in Bollmann et al. (2016), Cohen's κ, Krippendorff's α, and other related measures of agreement are not appropriate for use with normalization annotation itself, as in the NUCLE evaluation of **error+norm. given norm. identification**. The difficulties lie in the fact that the possible values for normalizations are not a small, finite set of categories but the set of all possible tokens in the target language. Given a relatively small annotated corpus, it is not possible to estimate how likely a given token might be in order to estimate chance agreement and even if it were possible, it would still not take into account the fact that a target hypothesis is frequently a form closely related to the original token. Additionally, κ and α give a higher weight to less frequent annotations, which means that normalizations for infrequent words play a larger role in the agreement coefficient even though an annotator's performance typically does not depend directly on the frequency of the word to be normalized. In fact, the opposite is often true: a misspelled rare name provided in the task context may be simple to normalize while a frequent determiner may be more challenging.

As there is no consensus on suitable agreement measures for normalization or target hypothesis annotation, we will primarily report percentage agreement in the following studies. We return to the issue of inter-annotator agreement measures for full target hypotheses in section 4.2.4.

3 Data

The normalization annotation experiments presented in the next section are performed using the

Q: Was sah der Mann, als er die Tür aufmachte?
'What did the man see when he opened the door?'

SA: Er sahe seiner Frau.
'He saw his wife.'

TA: Als er die Tür aufmachte, sah der Mann seine Frau.
'When he opened the door, the man saw his wife.'

RT: Als er die Tür aufmachte (sie weinte dabei, die Tür), sahen ihm die blaßblauen Augen seiner Frau entgegn.
'When he opened the door (it creaked, the door), his wife's pale blue eyes awaited him.'

MA1: Binary: *appropriate*, Detailed: *correct*

MA2: Binary: *appropriate*, Detailed: *correct*

Figure 1: CREG Example

Corpus of Reading Exercises in German (CREG, Ott et al., 2012), a German L2 learner corpus containing learner answers to reading comprehension exercises, which was collected in to enable research into learner language in a task-based context. The learners are students in German classes at two American universities who completed reading comprehension exercises as part of their coursework. The corpus contains: 1) reading texts, 2) comprehension questions, 3) teacher-provided target answer(s), 4) student answers to the questions, and 5) teacher assessments of the student answer meaning.

An example student answer (SA) to a comprehension question (Q) is shown in Figure 1 along with the target answer (TA) provided by a teacher and an excerpt from the reading text (RT). The meaning of each student answer is assessed by two teachers (MA1/2), who provide a binary assessment of the meaning (*appropriate* or *inappropriate* as an answer to the question) without taking spelling or grammar into account and a detailed classification of how the student answer differs from the provided target answer using the categories: *correct*, *missing concept*, *extra concept*, *blend*, and *non-answer*. Student answers marked as *appropriate* in the binary assessment are most frequently *correct* in the detailed assessment, but *appropriate* answers may also contain missing concepts, extra concepts, or blends.

Our experiments will primarily use data from CREG-5K, a subcorpus of CREG that contains

Binary / **Detailed**	Approp. (%)	Inapprop. (%)
Correct	76.9	0.0
Missing Concept	14.5	43.7
Extra Concept	6.2	3.2
Blend	2.4	50.2
Non-Answer	0.0	2.9

Table 1: Meaning Assessments in CREG-5K

~5000 student answers with a balanced number of appropriate and inappropriate answers. In total, CREG-5K contains 5138 student answers to 877 questions for 98 reading texts. The reading texts vary greatly in length, with an average of 961 tokens and a standard deviation of 1271 tokens. The student answers have been selected to contain a minimum of four tokens and have an average length of 11.75 tokens with a standard deviation of 7.13 tokens. The distribution of binary and detailed meaning assessments for CREG-5K is shown in Figure 1.

4 Experiments

On the basis of the CREG corpus, we explore the extent to which *context* and *appropriateness* play a role in the normalization of learner language. We first perform two normalization annotation studies on non-words in CREG-5K. Our goal is to investigate whether the amount of task context plays a role in inter-annotator agreement and whether appropriate answers can be more reliably normalized than inappropriate ones. Next, in section 4.2 we will describe the annotation of full meaning-based target hypotheses for the appropriate answers in CREG-5K and explore the evaluation of inter-annotator agreement for full normalizations and error tags.

4.1 Non-Word Normalization

We focus initially on non-word normalization, which allows us to sample a range of cases across the corpus from typos to English translations provided within student answers. The texts are automatically tokenized using the OpenNLP tokenizer trained on the non-headline sections of TüBa-D/Z version 9.0 (Telljohann et al., 2004) and non-words are identified automatically for the annotators. A *non-word* is defined as a token that does not appear in the question or reading text (if available in the experimental condition) or in the

DEREWO list of the 100,000 most frequent inflected words in a large German reference corpus (Institut für Deutsche Sprache, 2009).[1]

In two related experiments, we investigate the roles of task context and answer appropriateness in non-word normalization. We hypothesize that it is easier to perform non-word normalization reliably given more task context and that appropriate answers are easier to normalize than inappropriate ones, since annotators know the intended meaning of an appropriate answer from the task context. We first describe the annotation scheme and annotation tool used in both experiments, then present the experimental results.

4.1.1 Non-Word: Annotation Scheme

Non-words are annotated with a normalization that would be part of a form-meaning target hypothesis (a target hypothesis that preserves the intended meaning of the student answer while taking the task context into account, see section 4.2) for the student's answer given the available task context. Each non-word is additionally annotated with the amount of context required for the annotator to be confident that the provided normalization is the intended token in this context. If the annotator cannot be confident of a single normalization, multiple normalizations can be provided along with the context category *Hard*. The annotators are instructed to consider each context category in order:

- Real Word: non-word is a real word
- No Context: umlaut spellings with *e*, *ss* vs. *ß*
- Answer: the student answer alone
- Question + Answer: the answer along with the question
- Reading Text + Question + Answer: the full task context
- Hard (ambiguous, English): a single normalization cannot be chosen with confidence

When the full context is not available (only in some conditions in Experiment 1), only the context categories for the provided context should be annotated.

[1]This process misses some non-words and misspellings in the corpus because the DEREWO word list contains both old and new German spellings and also some proper names such as *Fisher* that cause our automatic selection process to miss some tokens in CREG that require a word-level normalization. All non-word normalizations are reviewed and additional non-word annotations are added in the full form-meaning target hypotheses in section 4.2.

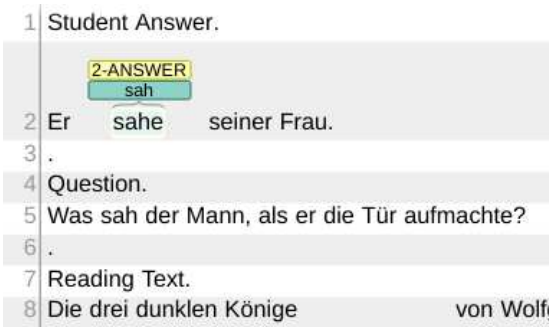

Figure 2: Non-Word Annotation in WebAnno

4.1.2 Non-Word: Annotation Tool

The non-words were annotated using custom layers in the tool WebAnno (Yimam et al., 2014). The student answers were preprocessed using a UIMA pipeline in order to tokenize them, identify non-words, and insert empty annotation spans to be filled in by the annotators. A screenshot of the WebAnno annotation environment is shown in Figure 2 for the student answer *Er sahe seiner Frau.* 'He saw his wife.' in response to the question *Was sah der Mann, als er die Tür aufmachte?* 'What did the man see as he opened the door?' The annotator has annotated the non-word *sahe* with the normalization *sah* 'saw' and specified the required context as the student answer alone.

4.1.3 Non-Word Experiment 1: Context

To evaluate the role of context in non-word normalization, the correct answers from CREG-5K (binary assessment: *appropriate*) were annotated. There were 1152 potential non-words in 2574 answers to 877 questions about 98 reading texts. The non-words were divided into four conditions by reading text, so that a reading text and its associated questions/answer are only included in one condition:

- training (10%)
- answer context only (15%)
- answer + question (15%)
- answer + question + reading text (60%)

Since we intend to annotate all non-words given the full context for the full form-meaning target hypotheses (see section 4.2), the non-words are not distributed equally between the conditions in order to reduce the reannotation burden in the next stage.

Two annotators annotated the training instances (10%) and met to discuss disagreements and to re-

	Norm.	Context	
	%	# Cats	α
Answer	74.8	4	0.696
A + Question	79.0	5	0.689
A + Q + Text	83.8	6	0.602

Table 2: IAA for Non-Words: Context

fine the annotation guidelines, then annotated the remaining instances (90%) independently. The results are shown in Table 2. As discussed in section 2.2, the agreement for the normalizations is presented as percentage agreement on the exact form provided and the agreement for the context category using Krippendorff's α.

As a result of the fact that the number of categories is not identical across conditions, the α values cannot be compared directly, however indicate moderate to substantial agreement on the context tags. When only the student answer is available, annotators agree 74.8% of the time on the normalization. This increases to 79.0% if the question is also available and to 83.8% if the question and reading text[2] are provided, showing that the presence of an explicit task context does enable a higher degree of reliability in normalization annotation.

For the annotations with the full context (60%, all six context tags are included), the confusion matrix for the context tags is shown in Figure 3. Some frequent sources of disagreement are rare inflections such as second person plural subjunctive forms (e.g., *stehet* 'would stand'), where one annotator annotated them as *Real Word* and the other normalized them to more frequent third person singular indicative forms (*steht* 'stand') with the category *Answer*, and instances where there are multiple, acceptable alternatives for prepositions in a particular context and one annotator consistently provided more alternatives, annotating such cases as *Hard* (vs. *Answer* for the other annotator).

4.1.4 Non-Word Experiment 2: Appropriateness

In the second non-word normalization experiment, the role of *appropriateness* is considered. The non-words consist of 529 non-words in 365 answers, presented to the annotator with the full reading text context. Since the appropriate answers from CREG-5K were annotated in the previous experiment, the appropriate answers come from CREG-1032 and other CREG subcorpora, while the inappropriate answers come from CREG-1032 and CREG-5K.

[2]As the students answering the reading comprehension questions do not have access to the teacher target answers while responding, the target answers are not presented to the annotations as part of this experiment.

	W	N	A	Q	R	H	Σ
W	41	0	26	1	2	2	72
N	0	71	4	0	1	3	79
A	5	8	321	4	7	18	363
Q	0	0	13	11	0	0	24
R	0	0	26	1	9	1	37
H	0	0	6	0	1	6	13
Σ	46	79	396	17	20	30	588

Table 3: Confusion Matrix: Non-Word with Full Context

	Norm.	Context
	%	α
Appropriate	83.3	0.678
Inappropriate	78.6	0.588

Table 4: IAA for Non-Words: Appropriateness

The two annotators from the previous experiment completed the annotation independently without any further training. The results are shown in Table 4. When the answer meaning has been assessed as *appropriate*, annotators agree on a single normalization in 83.3% of instances, nearly 5% higher than when the answer is *inappropriate*. Krippendorff's α, which is now comparable across both conditions since all six categories were used, is 0.678 for appropriate answers and drops to 0.588 for inappropriate answers, showing that annotators are more reliable in terms of the contribution of the task context for appropriate answers. This may be due to the fact that incorrect answers may include additional information that is not present in any part of the task context, so it may be more difficult to choose a context annotation.

4.2 Form-Meaning Target Hypothesis Annotation

Moving from non-word annotation to full target hypothesis annotation for the complete student answers, we present pilot results for the annotation of *form-meaning target hypotheses* on the *appropriate* answers from CREG-5K, the same subset of

CREG annotated in Experiment 1 (section 4.1.3) containing 2574 student answers.

A *form-meaning target hypothesis* (FMTH) is defined as a target hypothesis that provides a grammatical version of the student answer that:

- preserves as much of the meaning of the answer as possible
- respects the task context

If normalizations are necessary, these modifications should be as minimal as possible and align as closely as possible with material from the target answer, the question, and the reading text, e.g., if there is a missing concept, the inserted tokens should come directly from the task context.

After completing the non-word annotation experiments, one annotator reannotated the subset of non-words from Experiment 1 not presented in the full task context (30%) and the data was converted to Prague Markup Language[3] in preparation for use with the tool *feat* (see section 4.2.2). This annotator and a new second annotator performed the full target hypothesis and error annotation presented in the following sections.

4.2.1 FMTH: Error Annotation Scheme

The focus of the form-meaning target hypothesis annotation is on the normalization itself, however error annotation is also included to encourage a careful, reliable analysis of the student answers during the annotation process. The error annotation scheme attempts to parallel the CzeSL annotation scheme where possible, with non-words normalized in the first layer of annotation (word) and the full sentence normalized in the second layer of annotation (sentence). The top-level categories of the annotation scheme are presented in Table 5. For each error category, the table specifies whether a tag is possible on the word or sentence layers.

The top half of the table shows error tags similar to CzeSL, which are typical types of error tags seen in error-annotated learner corpora, and the bottom half of the table introduces new tags specific to the annotation of target hypotheses within a provided task context. In some instances, normalizations are necessary because of the question or reading text content, e.g., the tense of a student answer needs to be adjusted (tag: *Question*) or a proper name from the reading text is misspelled (tag: *Reading Text*). Students may have copied material from the reading text in a problematic way (e.g., copied 'not only' without the corresponding 'but also', *Copied - Problematic*), provided an answer that has a slightly incorrect meaning (*Answer Meaning*), or provided extra concepts that the annotators cannot normalize as consistently as material based on the task context (*Extraneous*). Problematic cases are discussed in further detail in section 4.2.3.

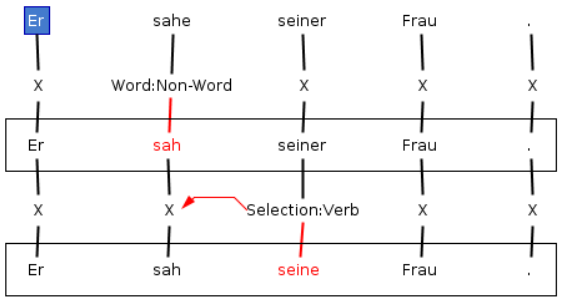

Figure 3: feat Annotation Tool

4.2.2 FMTH: Annotation Tool

The form-meaning target hypothesis annotation is performed using the tool *feat* (Flexible Error Annotation Tool), which was developed as part of the CzeSL project (Hana et al., 2012). We extend *feat* to support the CREG FMTH error scheme and to enable annotators to search for strings within long reading texts in order to make it easier to find the relevant sections and copied material.[4]

A screenshot of the *feat* annotation for the example from Figure 1 is shown in Figure 3. The top layer of tokens shows the original tokenized text, the middle layer shows the non-word normalizations, and the bottom layer shows the full form-meaning target hypothesis. In this example, the verb *sah* 'saw' selects the accusative case, so *seiner* 'his (DAT/GEN)' is normalized to *seine* 'his (NOM/ACC)' and the corresponding error tag *Selection:Verb* is chosen with a pointer identifying the head that selects this token.

4.2.3 Difficult Cases

Annotators encountered a range of difficult cases while annotating form-meaning target hypotheses, which relate to the nature of certain types of reading comprehension questions and aspects of annotating given a context provided by a written text.

[3] https://ufal.mff.cuni.cz/pml

[4] https://github.com/adrianeboyd/feat

Error Category	Word	Sent.	Description
Word	✓	✓	Capitalization, stem/inflection, word boundary, non-word error
Lexicon/Style	✓	✓	For lexical choice or style reasons the original token cannot be integrated into the target hypothesis
Selection		✓	Error in syntactic selection
Agreement		✓	Error in agreement
Order		✓	Error in word order
Modifier		✓	Error in a genitive modifier
Negation		✓	*kein* vs. *nicht*, double negatives, negative polarity items
Typo - POS		✓	Small spelling differences of 1-2 letters resulting in a different POS where a typo is more likely than a linguistically-motivated error
Secondary		✓	Annotator's normalizations require subsequent modifications to the student answer
Problem/Other	✓	✓	Problematic cases
Question	✓	✓	Target hypothesis chosen depends on the question content (providing a standalone answers, verb tense)
Reading Text	✓	✓	Target hypothesis chosen depends on the reading text content
Copied - Problematic		✓	Material lifted from the reading text that is not grammatical in the answer context
Answer Meaning		✓	Answer meaning does not correspond to target answer(s)
Extraneous		✓	Extra concepts in student answers

Table 5: Top-Level Categories in CREG FMTH Error Annotation Scheme

Q: Nennen Sie zwei Zimmer im Erdgeschoss.
'Name two rooms on the ground floor.'

SA: ein Wohnzimmer und ein Badzimmer
'a living room and a bathroom'

TA: Im Erdgeschoss gibt es ein Bad, Gäste WC, eine Küche und ein Wohn/Esszimmer.
'On the ground floor there is a bathroom, a guest bathroom, a kitchen, and a living/dining room.'

Figure 4: Difficult Cases: Enumerated Answers

Enumerated answers Enumerated answers present a particular problem for the reading comprehension task scenario. An example of a question with an enumerated answer is shown in Figure 4. When creating the CREG corpus, Ott et al. (2012) noticed a larger degree of disagreement in meaning assessment for enumerated answers, which appears to be due to the fact that is unclear how complete an enumeration needs to be to consider a student answer *appropriate*.

The target answers typically provide an exhaustive list of all items while an appropriate student answer provides only the number requested in the question. For form-meaning target hypothesis annotation, the annotators cannot rely on the target answers when evaluting the meaning of the student answer and when concepts are missing, there is also not a clear choice for which concept to insert into the student answer.

Extra concepts Students occasionally provide material in their responses that comes from their own world knowledge rather than the reading text. Figure 5 shows one instance where the student provides additional facts in an answer, which an annotator cannot evaluate within the task context.

Problematic copied material There are complicated annotation decisions to be made when the student has lifted material from the reading text in a problematic way. A few unnecessary words may be concatenated onto the end of a correct response or one half of a correlative conjunction pair may be missing. Such a case is shown in Figure 6, where the student has copied 'not only' from a sentence in the reading text without copying 'but also'. It is difficult for an annotator to decide whether to delete the first half of the correlative pair or insert the remainder of the sentence from the reading text, since neither choice would affect the meaning

Q: Wo leben die meisten Amischen heute?
'Where do most Amish live today?'

TA: Heute leben die meisten Amischen in Ohio, Pennsylvanien und Indiana.
'Today most Amish live in Ohio, Pennsylvania, and Indiana.'

SA: Die meisten Amischen leben in Ohio, Pennsylvania, und Indiana. Es gibt auch ein paar in Yoder, Kansas.
'Most Amish live in Ohio, Pennsylvania, and Indiana. There are also a few in Yoder, Kansas.'

MA: Binary: *appropriate*, Detailed: *extra concept*

Figure 5: Difficult Cases: Extra Concepts

Q: Was tat Herr Muschler, als seine Frau mit ihm zu sprechen versuchte?
'What was Herr Muschler doing while his wife was trying to talk to him?'

SA: Er sah nicht nur fern und die Zietung.
'He not only watched TV and the newspaper.'

TA: Er sah fern, las die Zeitung, rauchte eine Zigarette und trank ein Glas Bier.
'He watched TV, read the newspaper, smoked a cigarette, and drank a glass of beer.

RT: Herr Muschler sah nicht nur fern, sondern las außerdem noch die Zeitung.
'He was not only watching TV but also reading the newspaper.'

MA1: Binary: *appropriate*, Detailed: *extra concept*

MA2: Binary: *appropriate*, Detailed: *missing concept*

Figure 6: Difficult Cases: Problematic Copied Material

assessment for the response.

Reading text interpretation The least resolvable issues arise when two annotators disagree on the interpretation of the reading text itself. In Figure 7, the subject of an interview in a reading text states that he was unsure how many people might come to a demonstration and the student answer mentions 'force against not too many people', which potentially needs to be normalized under *Answer Meaning* to align with the target answer. One annotator interpreted the text to mean that the organizer was worried that not enough people would come and the other annotator thought that he was worried that too many people would come.

Q: Warum hatte Schorlemmer zu Beginn Angst?
'Why was Schorlemmer afraid at the beginning?'

TA: Er wusste nicht, wie viele Menschen kommen würden und ob die Polizei mit Gewalt gegen die Demonstration vorgeht.
'He did not know how many people would come and if the police would respond to the demonstration with force.'

SA: dass die Polizei mit Gewalt gegen nicht zu viele Menschen kommen
'that the police would come with force against not too many people'

RT: Ich hatte noch große Angst. Zum einen, weil ich nicht wusste, wie viele Menschen kommen würden. Zum anderen, weil ich Angst hatte, dass die Polizei mit Gewalt gegen die Demonstration vorgehen würde.
'I was still very scared. On the one hand, because I didn't know how many people would come. On the other hand, because I was scared that the police would respond to the demonstration with force.'

MA1: Binary: *appropriate*, Detailed: *correct*

MA2: Binary: *appropriate*, Detailed: *missing concept*

Figure 7: Difficult Cases: Reading Text Interpretation

With differing interpretations of the reading text, there is little hope for similar target hypotheses. Despite the explicit task context, such ambiguous statements may still be present in a reading text and lead to inter-annotator disagreement.

4.2.4 IAA for Meaning-Based Target Hypotheses

After annotating approximately 75% of the CREG-5K appropriate answers with meaning-based target hypotheses in a collaborative process including many discussions of difficult cases and refinements to the annotation manual, the two annotators annotated a subcorpus of 250 student answers independently in order to evaluate inter-annotator agreement. The subcorpus contains 3259 tokens in 250 appropriate student answers that have been sampled randomly from CREG-5K.

In order for our evaluation to be comparable to the evaluation of similar L2 German target hypotheses in Köhn and Köhn (2018), annotations on the word and sentence level are aligned with the original tokens by merging any inserted tokens into the annotation for the following token, with annotations at the end of a sentence merged into

the preceding token. In case there are multiple error tags on a single token or in merged annotations, these are treated as a set of error tags on the original token.

Cohen's κ[5] for **normalization identification** (see section 2.2 for detailed descriptions) is 0.68, which shows substantial agreement and falls in between results reported for NUCLE ($\kappa = 0.39$) and for ComiGS ($\kappa = 0.86$). For **error tag given normalization**, κ is 0.47, which is slightly lower than NUCLE ($\kappa = 0.55$) for a relatively similar set of error tags. However, our annotation allows annotators to annotate multiple error tags on a single word, resulting in 57 combinations of error tags (for 15 individual tags) which are treated as separate tags in κ's comparisons. Using the more appropriate MASI distance metric for set annotations (Passonneau, 2006), we obtain $\alpha_{MASI} = 0.50$ for 15 error tags, again given that both annotators normalized the token.

We find only small differences between **error tag given normalization** ($\kappa = 0.47$), which ignores cases where only one annotator annotated an error, and simply **error tag** for all tokens, with $\kappa = 0.45$. Although ~86% of the tokens are not annotated with error tags, chance-corrected agreement measures account for the high probability that an original token remains unmodified in a target hypothesis and that most tokens in the corpus are not annotated with error tags.

As with non-word normalizations, we calculate only the percentage agreement for the normalizations themselves. For cases where both annotators agreed that a token should be normalized, the same normalization is provided in 70% of instances. Given the fact that target hypothesis annotation can involve complicated edits and reordering, it is not surprising that the agreement is slightly lower than in the non-word experiments reported Table 2 and Table 4.

We perform a similar analysis of **error tag identification** to compare our results to those reported for CzeSL in Rosen et al. (2013). For the top-level error tags that appear at least ten times in our subcorpus, we evaluate whether annotators agreed about which tokens are annotated with a particular tag. These results are shown in Table 6. As in CzeSL, there is a wide range of agreement with some error tags being annotated fairly reliably (Punctuation, Order) and others with little agreement beyond chance (Reading Text, Answer Meaning).

Error Tag	κ	**Avg. Tags / Annotator**
Punctuation	0.65	58
Order	0.57	42
Selection	0.46	171
Typo	0.40	5
Agreement	0.38	60
Word	0.36	17
Lexicon	0.18	43
Secondary	0.17	24
Question	0.15	43
Reading Text	0.07	38
Answer Meaning	0.03	25

Table 6: IAA for Error Tag Identification

A common thread in the inspection of difficult cases throughout the annotation process is that difficulties frequently occur when the detailed meaning assessment is not *correct* for one or both teacher assessments. Since an answer with a *missing concept*, *extra concept*, or *blend* either does not supply the correct answer meaning or may include material from outside the task context, this is not surprising. To explore the relationship between difficulty as perceived by the annotators and inter-annotator agreement, we consider three partitions of the data: 1) both detailed meaning assessment are correct vs. all other combinations of assessments, 2) the two detailed meaning assessments are identical vs. different, and 3) the cases where at least one detailed assessment includes a particular detailed tag.

We calculate κ for **normalization identification**, κ for **error tag** for all error tags as shown in Table 7. Agreement measures for both drop slightly for *correct* vs. *other* but surprisingly increase slightly for answers where the teachers did not agree on the detailed assessment. Larger differences are seen for the individual detailed categories, with *blend* and *extra concept* instances showing much lower agreement, in particular for error tags related to *extra concepts*. In general, κ for **normalization identification** does not appear to reflect annotators' perception of overall difficulty, which can be explained by the fact that merely identifying problematic spans is only a

[5]All inter-annotator agreement measures are calculated using the scripts by Thomas Lippincott and Rebecca Passonneau: https://cswww.essex.ac.uk/Research/nle/arrau/Lippincott/agreement.tgz

	All	Both Correct	Other	MA1 = MA2	MA1 ≠ MA2	MA Includes			
						Correct	Blend	Missing	Extra
# Tokens	3259	2143	1116	2340	919	2914	157	652	455
# Answers	250	175	75	193	57	225	9	49	24
κ, Norm. Id.	0.68	0.69	0.66	0.68	0.70	0.69	0.60	0.70	0.62
κ, Error Tag	0.45	0.47	0.42	0.45	0.46	0.47	0.43	0.49	0.29
CharacTER	0.11	0.10	0.13	0.11	0.10	0.10	0.12	0.12	0.14

Table 7: IAA by Detailed Meaning Assessment

small part of the annotation task.

Since none of the inter-annotator agreement measures are suitable for comparing agreement between the normalization annotation, we turn to alternate metrics that have been proposed for the related tasks of machine translation evaluation and paraphrase detection. These metrics should ideally provide a more holistic evaluation of whether two target hypotheses are similar to each other on the sentence level rather than focusing on annotations for individual tokens. One recent metric from machine translation evaluation, CharacTER, seems particularly promising since it has been shown to correlate highly with human judgments for languages with richer morphology such as German and Russian (Wang et al., 2016).

CharacTER is adapted from the *translation edit rate* metric (TER, Olive, 2005), which calculates the number of edits required to convert one translation to a reference translation on the word level. CharacTER extends this to consider both *shifts* on the word level to align two sentences (counted as the average number of characters in the words shifted) and then further *character edits* required to transform the shifted sentence into the reference translation. This combination allows for variations in word order and small differences in morphological endings to be counted in a more fine-grained way than word-only edits. CharacTER is formally defined as:

$$\text{CharacTER} = \frac{\text{shift cost + edit distance}}{\text{\# characters in the hypothesis sentence}}$$

The CharacTER score is lower when two sentences are more similar, with a score of 0 for identical sentences. Since it is intended to compare a system translation to a reference translation, we extend CharacTER[6] to calculate scores with each annotator providing the reference translation once and average these scores on the sentence level. Although a translation metric does not account for the overlap between the original student answer and the target hypothesis (thus such low overall scores when compared to machine translation), the types of cases that teachers found difficult to assess and that annotators found difficult to normalize are reflected more accurately (with higher CharacTER scores) than with other measures.

[6] https://github.com/rwth-i6/CharacTER/

5 Conclusion / Outlook

In experiments on word-level and sentence-level normalization for an L2 German reading comprehension corpus, we show that inter-annotator agreement for normalization annotation increases when more of the task context is provided to the annotators and that *appropriate* answers can be normalized more reliably than inappropriate answers. In the evaluation of inter-annotator agreement for full form-meaning target hypotheses, which preserve the intended meaning while taking the task context into account, we explore a range of inter-annotator agreement metrics and how the CharacTER machine translation metric shows promise for the comparison of normalization annotations on the sentence level.

In future work on evaluating inter-annotator agreement for normalization annotation, we would like to explore the use of additional machine translation metrics and related metrics from paraphrase detection and plagiarism detection, since these could potentially capture many of the similarities in form and meaning while accounting for the fact that annotators' normalizations should come from the provided context as much as possible.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful feedbank. This work was supported by the German Research Foundation (DFG) under project ME 1447/2-1 and through the Collaborative Research Center 833.

References

Ron Artstein and Massimo Poesio. 2009. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):1–42.

Marcel Bollmann, Stefanie Dipper, and Florian Petran. 2016. Evaluating inter-annotator agreement on historical spelling normalization. *Proceedings of LAW X – The 10th Linguistic Annotation Workshop*, pages 89–98.

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

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31. Association for Computational Linguistics.

Eileen Fitzpatrick and M. S. Seegmiller. 2004. The Montclair electronic language database project. In U. Connor and T.A. Upton, editors, *Applied Corpus Linguistics: A Multidimensional Perspective*. Rodopi, Amsterdam.

Jirka Hana, Alexandr Rosen, Barbora Štindlová, and Petr Jäger. 2012. Building a learner corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Hagen Hirschmann, Seanna Doolittle, and Anke Lüdeling. 2007. Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*, Birmingham.

Institut für Deutsche Sprache. 2009. Korpusbasierte Wortformenliste DeReWo, v-100000t-2009-04-30-0.1, mit Benutzerdokumentation. Technical Report IDS-KL-2009-02, Institut für Deutsche Sprache, Programmbereich Korpuslinguistik.

Christine Köhn and Arne Köhn. 2018. An annotated corpus of picture stories retold by language learners. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 121–132. Association for Computational Linguistics.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA.

John Lee, Joel Tetreault, and Martin Chodorow. 2009. Human evaluation of article and noun number usage: Influences of context and construction variability. In *ACL 2009 Proceedings of the Linguistic Annotation Workshop III (LAW3)*. Association for Computational Linguistics.

Anke Lüdeling. 2008. Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In Maik Walter and Patrick Grommes, editors, *Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweitspracherwerbsforschung*, pages 119–140. Max Niemeyer Verlag, Tübingen.

Detmar Meurers. 2015. Learner corpora and natural language processing. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *The Cambridge Handbook of Learner Corpus Research*, pages 537–566. Cambridge University Press.

Detmar Meurers and Markus Dickinson. 2017. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning, Special Issue on Language learning research at the intersection of experimental, corpus-based and computational methods: Evidence and interpretation*. To appear.

Shogo Miura. 1998. Hiroshima English Learners' Corpus: English learner No. 2 (English I & English II). Department of English Language Education, Hiroshima University. `http://purl.org/icall/helc`.

Joseph Olive. 2005. Global autonomous language exploitation (gale). Technical report, DARPA/IPTO Proposer Information Pamphlet.

Niels Ott, Ramon Ziai, and Detmar Meurers. 2012. Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In Thomas Schmidt and Kai Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM), pages 47–69. Benjamins, Amsterdam.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evauation (LREC-06)*.

Marc Reznicek, Anke Lüdeling, Cedric Krummes, and Franziska Schwantuschke. 2012. *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.0*.

Alexandr Rosen, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2013. Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, pages 1–28.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lissabon.

Joel Tetreault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *Proceedings of the workshop on Human Judgments in Computational*

Linguistics at COLING-08, pages 24–32, Manchester, UK. Association for Computational Linguistics.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Seid Muhie Yimam, Chris Biemann, Richard Eckart de Castilho, and Iryna Gurevych. 2014. Automatic annotation suggestions and custom annotation layers in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 91–96, Baltimore, Maryland. Association for Computational Linguistics.