An Annotated Corpus of Picture Stories Retold by Language Learners

Christine Köhn and Arne Köhn
Natural Lanuguage Systems Group
Department of Informatics
Universität Hamburg
{ckoehn,koehn}@informatik.uni-hamburg.de

Abstract

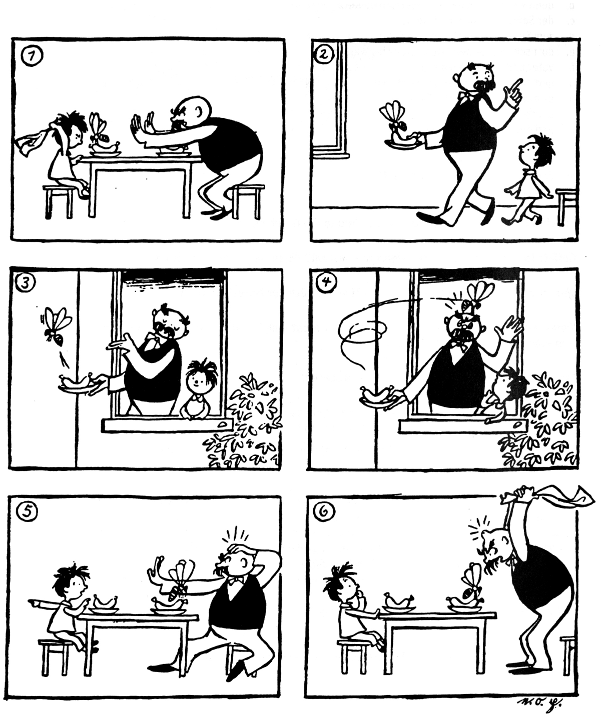
Corpora with language learner writing usually consist of essays, which are difficult to annotate reliably and to process automatically due to the high degree of freedom and the nature of learner language. We develop a task which mildly constrains learner utterances to facilitate consistent annotation and reliable automatic processing but at the same time does not prime learners with textual information. In this task, learners retell a comic strip. We present the resulting task-based corpus of stories written by learners of German. We designed the corpus to be able to serve multiple purposes: The corpus was manually annotated, including target hypotheses and syntactic structures. We achieve a very high inter-annotator agreement: $\kappa = 0.765$ for the annotation of minimal target hypotheses and $\kappa = 0.507$ for the extended target hypotheses. We attribute this to the design of our task and the annotation guidelines, which are based on those for the Falko corpus (Reznicek et al., 2012).

1 Introduction

Learner corpora are useful for a variety of tasks in natural language processing (NLP) and linguistics, including analyzing learner language, developing and evaluating NLP tools with respect to learner language and building grammatical error correction (GEC) systems. Many learner corpora are available today but they are far from covering the whole spectrum of learner utterances: First, they usually consist of essays, escpecially the large corpora such as ICLE (Granger et al., 2009), NUCLE (Dahlmeier et al., 2013), or Falko (Reznicek et al., 2012; Reznicek et al., 2013), and second, languages other than English are underrepresented.

Writing essays is a common task for students and it is therefore worthwhile to conduct research on student essays. However, the content and form of essays is only marginally constrained, which results in a low agreement between error annotations (e.g. Fitzpatrick and Seegmiller (2004)). To foster reliable interpretation of learner texts, Ott et al. (2012) argue for collecting learner corpora with explicit task contexts. Intuitively, knowing the context of an utterance facilitates interpreting it. Ott et al. (2012) chose reading comprehension as such a task and achieve reasonable inter-annotator agreement for meaning assessment of the responses. Reading comprehension exercises have the advantage that they include questions and texts as contextual information, which can be exploited for automating meaning assessment (e. g. Bailey and Meurers (2008), Hahn and Meurers (2012)). However, the provided texts and prompts influence the learner's choice of words and therefore pure interlanguage cannot be observed.

One task with explicit task context but without providing the students with texts or text fragments which encode the correct answer is picture description. Pictures showing a single activity in isolation constrain the answers to a sensible degree for further processing, e.g. King and Dickinson (2013) achieve 92.3% accuracy for extracting semantic triples of the form *verb(subj,obj)*. The identification of the verb, the subject and the object are crucial parts for content and grammatical assessment and work well for this type of picture description tasks because the resulting sentences are conceptually simple. We strive to move a step further and increase the variance of the learner writing and correspondingly the processing



(a) Moral mit Wespen "moral with wasps" by Erich Ohser

Der	*Sohn*	*will*	*die*	*Wespe*	*töten*	_	*aber*
The	son	wants to	the	wasp	kill		but
der	*Vater*	*haltet*	*er.*				
the	father	**hold**	**he**.				

'The son wants to kill the wasp, but the father **holds** him' (text 3nnn_1)

Der	*Sohn*	*sitzt*	*auf*	***ein***	*Hocker*	*neben*		*dem*
The	son	sits	on	a	stool	next to		the
Tisch	*und*	*schwingt*	*ein*	*Handtuch*	_	*um*	*die*	
table	and	swings	a	towel		to	the	
Wespe	*zu*	*töten.*						
wasp	to	kill.						

'The son sits on a stool next to the table and swings a towel in order to kill the wasp.' (text 2mVs_1)

Der	*Sohn*	*will*	***es***	*mit*	*einem*	*Tuch*	*töten.*
The	son	wants	it	with	a	cloth	kill.

'The son wants to kill it with a cloth.' (text ATwN_1)

(b) Examples for image 1, describing the son's actions. Errors in bold, underscores mark missing commas. Verb error in first sentence results in distorted meaning: The learner used *halten* "to hold" instead of *aufhalten* "to stop".

Figure 1: One of the picture stories from the retelling task with example sentences from the descriptions of the first panel of the story.

difficulty while sustaining a strong visual context. Simply increasing the number of items and situations that can be described in a picture is insufficient since this would probably result in descriptions about mainly unrelated actions happening in parallel as if concatenating the descriptions of single actions. The task we designed to obtain a language learner corpus meets the following criteria:

- It has a *strong visual context* which allows human annotators (and machines) to find a clear target hypothesis (in contrast to essay based corpora)
- It *captures real language use* and has no linguistic context in the form of texts, which would contain or influence the expected result (in contrast to question answering)
- It has *free-form answers* (in contrast to fill-the-gap exercises)
- It elicits a *variety of sentence structures*, e.g. encourages to establish causal or temporal links between sentences (in contrast to simple activity descriptions)

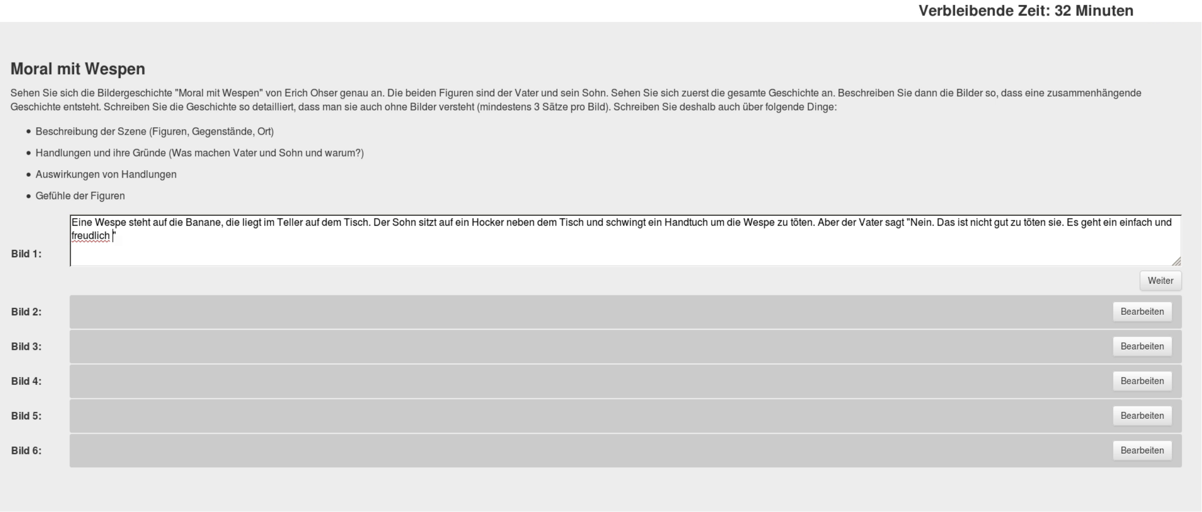
We designed a picture story retelling task in which language learners narrate the story shown in a comic strip, which we describe in more detail in Section 2. We name the result **Comi**c **S**trips Retold by Learners of **G**erman: the ComiGS corpus. The circumstances of the corpus collection and the type and amount of data collected are reported in Section 3 and 4.

Manual annotation is time-consuming and expensive. Therefore, an annotation serving more than one purpose is desirable and we annotated the corpus in a way that it can be utilized for different tasks. Section 4 explicates the annotations we added to the learners' texts and what they can be used for.

2 Picture Story Retelling Task

In the task, learners retell what they see in a comic strip. In order to meet the criteria mentioned above, the picture story as well as the instructions must be selected carefully. The story must

- not contain text
- provide enough material to write about, especially have more than one actor



Verbleibende Zeit: 32 Minuten

Moral mit Wespen

Sehen Sie sich die Bildergeschichte "Moral mit Wespen" von Erich Ohser genau an. Die beiden Figuren sind der Vater und sein Sohn. Sehen Sie sich zuerst die gesamte Geschichte an. Beschreiben Sie dann die Bilder so, dass eine zusammenhängende Geschichte entsteht. Schreiben Sie die Geschichte so detailliert, dass man sie auch ohne Bilder versteht (mindestens 3 Sätze pro Bild). Schreiben Sie deshalb auch über folgende Dinge:

- Beschreibung der Szene (Figuren, Gegenstände, Ort)
- Handlungen und ihre Gründe (Was machen Vater und Sohn und warum?)
- Auswirkungen von Handlungen
- Gefühle der Figuren

Bild 1: Eine Wespe steht auf die Banane, die liegt im Teller auf dem Tisch. Der Sohn sitzt auf ein Hocker neben dem Tisch und schwingt ein Handtuch um die Wespe zu töten. Aber der Vater sagt "Nein. Das ist nicht gut zu töten sie. Es geht ein einfach und freudlich "

Bild 2:

Bild 3:

Bild 4:

Bild 5:

Bild 6:

Figure 2: User interface for participants. Top to bottom: Title of the picture story, general task description, input boxes for each picture in the story. Top right: remaining time.

- be easy to understand for learners
- be pleasant enough in order to make the learner comfortable writing about it

Since we wanted to make the results of our corpus collection freely available, we also required that the picture stories can be distributed together with the corpus. We found the numerous *Vater und Sohn* (“father and son”) stories by Erich Ohser to be fitting, of which some had already been used for teaching German as a foreign language (da Luz Videira Murta, 1991; Eppert, 2001), and the copyright expired at the end of 2014[1]. We selected five stories, which we regarded as most suitable[2].

The task instructions should prevent the learners from writing short and superficial stories but we did not want to point the learner directly to the events in the story to avoid imposing our interpretation and our choice of words on the learner. Instead, the task instructs the learner to first look carefully at the whole picture story and then write a detailed and coherent story. The learners were required to write at least three sentences per image but we did not enforce this requirement. Then we added a list of things that the learner should include: description of the scene (characters, items, locations), actions and their causes (what do the characters do and why?), consequences of actions, the characters’ feelings.

We made one exception to our “no textual influence” policy: We stated the context of the story, i.e. the story’s name, the author’s name and the family’s characters: the father, the mother and the son. One of the titles contains a word which references an important item in the story but it is unlikely that the learners know the word, so we explained it in a few words.

We tested the complete task (instructions and story) on native speakers. Afterwards, we ranked the stories by perceived fit for the task and selected the two best stories for our corpus collection task (see Figure 1a). Then we estimated the difficulty of the task to determine the CEFR[3] level the learners should have in order to do the exercise. We estimated the taskto be suitable for levels A2 (elementary, i.e. upper beginner) and B1 (intermediate). We consulted a German as a foreign language instructor and

[1] Erich Ohser was prohibited from working under the Nazi regime and published the father and son comics under a pseudonym. He and Erich Knauf were arrested for making political jokes in 1944. Ohser commited suicide the day before his trial, Knauf was executed. (`https://en.wikipedia.org/wiki/E._O._Plauen`)

[2] None of the stories contains text, but not all of them are completely devoid of symbols: One of the stories (Der Schmöker “The page-turner”) contains two question marks to indicate a vacant chairs.

[3] The **C**ommon **E**uropean **F**ramework of **R**eference for Languages: Learning, Teaching, Assessment defines three levels of language proficiency for learners, which are each subdivided into two levels: A1/A2 (basic user), B1/B2 (independent user) and C1/C2 (proficient user).

he confirmed our estimation of level and time needed to complete the exercise (90 min for two stories). Therefore, all learners had 90 min for the task but the time limit was not strictly enforced. However, the actual time the participants needed for each picture story is documented. Learners with levels A2 and B1 were given two stories. Learners who had a higher level or who were very fast were given one additional story (the third best story). Learner levels and mother languages are reported in Table 1.

3 Data Collection

The language learners typed their text into a web interface (see Figure 2). Since the task did not provide a vocabulary list, they were asked to bring a dictionary or a dictionary app with them but were not allowed to translate whole sentences using tools such as Google Translate.

The interface showed the exercise text and a text area for each image. This way, each story part was assigned to one image without manual annotation. Naturally, this mapping is rough: A part may reference other pictures (e.g. relating the current situation to a previous one) or it may contain content which is not inferable from the story (e.g. the learners own experience) or is not inferable from one picture alone (e.g. if something changes from one picture to the other and the learner describes the actions to cause the change). Despite these shortcomings, we found the mapping to be very useful when annotating the target hypotheses.

An important question when designing the interface was whether or not to include some kind of spellchecking. Without a spellchecker, the texts would be as if the learner would write them offline by hand but this would not be a realistic scenario. Most learners would probably use a spellchecker when writing on a computer – at least to avoid typing errors. Furthermore, in the subsequent error annotation, the annotator interprets misspelled words when formulating the target hypothesis. In our opinion, the learners should at least get the chance to correct their error. Therefore, we decided to make the learners aware that they might have typed something wrong: We implemented a compromise between using no spellchecking at all and a full-fledged spellchecker. We used the browser-internal spellchecking which underlines possibly misspelled words but we decided not to provide suggestions on how to correct the error[4]: The main reason for this is that we feared that learners confide too much in the spellchecker and distort their texts, e. g. choose a wrong suggestion over their correctly spelled word.

CEFR level	#
A2	6
A2/B1	3
B1	11
B1/B2	1
B2	5
B2/C1	4

Mother language	#
Italian	5
Chinese	5
Spanish	4
Russian	4
Portuguese	3
English	3
Persian	2
Turkish	1
Romanian	1
Polish	1
Korean	1
Igbo	1
Armenian	1

Table 1: Learner levels and mother languages represented in the corpus. For each learner, we report the maximum of their certified and self-reported level, as the certification was sometimes outdated. Some learners reported two mother languages (English and another language), in which case we count both.

In the beginning, the learners were instructed by the supervisor. They started with an input test page where they got used to the interface and were made aware of the spellchecker: They were told to regard the spellchecker as an aid to make them aware of possible mistakes but that the spellchecker is imperfect. They were given examples to show that not every misspelled word is underlined and that an underline does not necessarily mean that the word is incorrect. Note that the texts were not part of an exam and the learners were aware that their texts would not be graded. The learners were instructed that they should rather cover the stories' content entirely than to write perfect sentences.

Overall, we collected texts from thirty learners. Twenty wrote texts for two stories and ten wrote texts for three stories. We collected a variety of meta data about each learner and each learner's text, including the background of the language learner and apart from the common data, we also asked for their previous

[4] Due to technical problems, some learners used suggestions but we marked these texts.

experience with this kind of task. For most of the learners, we also stored the text and timestamp every time they pressed a key as well as a video capture of their screen. This makes it possible for future work to draw conclusions about the development of the texts. For every learner, we got a certificate, e. g. the completion of a course, about their language level. However, this might not be identical to the actual level on the day of the performance. We therefore noted level, type of certificate and time span between issuing the certificate and the day of the task.

4 Annotations

We wanted to create annotations suitable for a variety of tasks, either directly or as a basis. The tasks which we wanted to cover in particular are: analysis by means of corpus linguistics, grammatical error diagnosis, grammatical error correction (GEC), and the development of NLP tools such as parsers and taggers.

All tasks except GEC[5] either need or profit from syntax annotation and it is therefore essential. Next in importance are Part-of-Speech (PoS) tags. The first three tasks benefit from annotation with error tags, i. e. tags that mark and categorize errors according to an error classification scheme. Lüdeling (2008) argues that error tags should only be used with respect to a reconstructed utterance. Therefore, a reconstruction known as target hypothesis is needed to annotate error tags. There can be a wide range of acceptable target hypotheses for the same sentence and even trained teachers do not agree on the same correction.

Learner utterances can diverge from the target language on several linguistic levels: orthography, morphology, syntax, semantics and pragmatics. Existing corpora annotate corrections on different levels: The EAGLE annotation scheme (Boyd, 2010) considers sentences in isolation and considers them ungrammatical “if there is no context in which the sentence could be uttered”. With this definition, a sentences or a corrected sentence which is not semantically or pragmatically appropriate in the actual context may be regarded as correct. Many existing corpora address errors only on the grammatical level. Sakaguchi et al. (2016) argue that corpora for GEC should not aim at grammaticality but at what they call *native-language fluency*: They “consider a text to be *fluent* when it looks and sounds natural to a native-speaking population”. Reznicek et al. (2013) propose two target hypotheses with different scopes to cover these differences – a minimal target hypothesis (TH1) and an extended target hypothesis (TH2) – and show that depending on the TH, different phenomena can be studied. Both THs should change the original text as little as possible. TH1 addresses only lower linguistic levels. It changes the original text to make it grammatical without regarding semantics, pragmatics and style. In contrast, TH2 does not ignore these levels. It takes the given context into account and aims at creating a text that is as similar as possible to that of a native speaker (Reznicek et al., 2012). By design, the TH2 may deviate substantially from the original text (Reznicek et al., 2013).

For TH1, there are detailed annotation rules, whereas there are only rough guidelines and annotation examples for the TH2 (Reznicek et al., 2013; Reznicek et al., 2012). An interesting property of TH1 is that it is explicitly designed to serve as a normalization layer suitable for automatic processing, which has proven to be useful in the past (e. g. Rehbein et al. (2012)). In particular, syntax parses of TH1 can be mapped back to the original utterance, obtaining a syntax annotation of the learner utterance (Hirschmann et al., 2013).

Ragheb and Dickinson (2011) argue that annotation schemes for learner language should be specifically tailored for this purpose because comparing it to the target language (or to the L1 of the learner) might obscure properties of the interlanguage. Because of this, Ragheb and Dickinson (2012) developed an annotation scheme for syntax and PoS tags for learner English. Each word is annotated with two PoS tags since a single PoS tag is often not adequate (see Díaz-Negrillo et al. (2010) for a discussion): One captures morphological evidence and one distributional evidence. The syntax annotation includes a subcategorization layer and a dependency layer. The morphosyntactic dependencies are based on the surface forms and morphological evidence and the subcategorization frames represent arguments that are

[5] Some approaches for GEC have been shown to benefit from syntactic information, e. g. adding syntactic features to a classifier for agreement errors (Rozovskaya and Roth, 2014). Approaches that treat correction as a machine translation task and are now close to human-level performance (Grundkiewicz and Junczys-Dowmunt, 2018) do not need syntactic information.

required in the target language. This annotation scheme has been shown to achieve good inter-annotator agreement (Ragheb and Dickinson, 2013).

Another approach to annotating learner syntax is to use an existing annotation scheme for the target language and use it on learner language. Ott and Ziai (2010) annotated 109 sentences written by learners of German with the dependency scheme by Foth (2006) because using an annotation scheme specifically for learner language is not suitable for their purpose. They measured an inter-annotator agreement between the three annotators of 88.1% in terms of labeled attachment accuracy. However, for 42 sentences all three annotations (dependencies and labels) differed. For 6 of them this was due to differences on the underlying target hypothesis (though they were not annotated). Therefore, Ott and Ziai (2010) recommend to explicitly annotate target hypotheses as proposed by Lüdeling (2008).

Rosén and De Smedt (2010) and Hirschmann et al. (2013) argue that target hypotheses are necessary for automatic analysis of learner language. Rosén and De Smedt (2010) discuss problems with an early version of Ragheb and Dickinson's annotation scheme (Dickinson and Ragheb, 2009) which we think are still valid for the latest version.

Overall, the best approach for syntactic annotation of our corpus is to annotate target hypotheses and map their syntactic structures back to the original utterances. This way, we obtain corrections of the sentences and syntax annotations.

We annotated our corpus manually with two target hypotheses, a minimal TH and an extended TH, following the guidelines for the Falko corpus (Reznicek et al., 2012; Reznicek et al., 2013). We made a few adaptations which we will explain later. We manually annotated dependencies for the target hypotheses using the well-documented scheme by (Foth, 2006) which has the advantages that it is a genuine dependency scheme and was used for annotating the largest German dependency treebank HDT (Foth et al., 2014). We annotated the target hypotheses manually with lemmas and PoS tags using the STTS tag set (Schiller et al., 1999).

The ComiGS corpus contains 18k tokens. The sentence lengths of the original texts have a relatively even distribution with an average of 12.2 tokens, which is close to the median of 11 tokens (see Table 3). This shows that the learners on average produce more complex sentences than just canonical subject verb object sentences (see example sentences in Figure 1b). For comparison, the answers to the reading comprehension questions in the CREG-109 corpus only have an average sentence length of 8.26 tokens.

Tag	Description
`INS`	inserted token in TH
`DEL`	deleted token in TH
`CHA`	changed token in TH
`MOVS`	source location of moved token in TH
`MOVT`	target location of moved token in TH
`MERGE`	tokens merged in TH
`SPLIT`	tokens split in TH

Table 2: Automatic error tags from Reznicek et al. (2013), which are used for this corpus

	#orig	#TH1	#TH2
25%	8	8	8
Median	11	12	12
75%	15	16	16
Mean	12.2	12.5	12.6

Table 3: Distribution of sentence lengths (in tokens) in our corpus

4.1 Minimal and Extended Target Hypotheses

The minimal target hypothesis (TH1) and the extended target hypothesis (TH2) reconstruct the original utterance while attempting to minimize the changes to the original text. The TH1 only corrects errors on lower linguistic levels (orthography, morphology and syntax). In contrast, the TH2 aims at creating a reconstructed text which is as similar as possible to a native speaker utterance and, therefore, also considers semantics, pragmatics and information structure (Reznicek et al., 2012).

The rules for the TH1 cover many cases for disambiguating conflicting evidence, e.g. if the verb and the arguments do not match, the verb should be preserved and the arguments adjusted. Consider the examples

ctok	Die	Kind	ist	liegend	[...]
TH2	Das	Kind	liegt		[...]
	The	*child*	*is lying/lies*		[...]

(a) Example where two tokens are corrected into one token. Without explicit alignment it would be impossible to differentiate between deleting "ist" and correcting "liegend" and correcting both into a single word.

ctok	Die	Kind	ist	[...]	liegend
TH2	Das	Kind	liegt	[...]	
tmid			1	[...]	1

(b) Example where two tokens are corrected into one token, with other tokens in between. Same as 3a, but the correspondence has to be established by an additional linking layer.

ctok	Der	Mann	geht	weiter	[...]
TH2	Der	Mann	fährt	fort	[...]
	The	*man*	*walks/goes*	*on*	[...]
tmid			1	1	[...]

(c) Example where two tokens are corrected into two other tokens. The linking layer is used to denote that this is a single correction. Without this additional annotation, independent but neighboring corections could not be distinguished from a single correction.

Figure 3: Examples for alignments in the error annotation. ctok: manually corrected tokenization, TH2: correction layer, tmid: token move id.
Sources: Example 3a from text YRGN_2, Example 3b adapted from 3a, Example 3c from text bsPg_3.

in Figure 1b: The TH1 corrects the verb form error, adds a separable verb prefix to the first sentence and corrects the pronoun (*haltet er* → *hält ihn auf* "stops him") whereas the TH2 uses a more suitable verb which requires the addition of an adverb (*hält ihn davon ab* "prevents him from doing it"). The TH2 of the third sentence corrects the pronoun *es* "it" to *sie*, a female pronoun, because it references a previously mentioned noun with female gender. The TH1, in contrast, does not change this pronoun as the sentence is grammatically correct but not adequate.

An important feature of the target hypotheses is that they are manually aligned to the original text token by token. This makes it possible to compute automatic error tags in the form of edit tags as has been done for the Falko corpus (Reznicek et al., 2013). We annotated our corpus with the same tags (see Table 2). When searching for patterns in the corpus, these error tags are useful for constraining the search results. Note that an automatically created alignment would not allow to derive all of the error tags: Merged and split tokens could not be identified except for simple cases such as deleting or inserting spaces and an inserted token followed by a deleted token (or a deleted token followed by an inserted token) could not be distinguished from a changed token. If more detailed error tags are needed, e. g. for evaluating GEC systems with respect to error type, error types could be automatically annotated as proposed by Bryant et al. (2017). This approach has the advantage that the same error set can be automatically annotated to different corpora (as long as target hypotheses are available) yielding a unified error annotation.

4.2 Adaptations and Extensions to the Falko scheme

The Falko manual requires that split or merged tokens are annotated as overlapping spans if possible. Tokens from the original text are merged like shown in Figure 3a. While this is a useful feature as it shows when a token with the same function is changed, this information gets lost when the token is moved. Therefore, we added a layer called tokmovid for every TH which assigns a unique identifier (we used numbers) to any number of tokens which would have been annotated as an overlapping span (tmid= tokmovid for TH2), see Figure 3b. Note that this also applies to tokens which are contiguous but cannot be annotated by an overlapping span. This is the case when one set of tokens is changed into another, see Figure 3c. In that example, the learner used *geht weiter* as in "walks on" but this is changed into a separable verb *fährt fort* ("goes on") in TH2. The tokmovid indicates that the words are not replaced in isolation but as a unit. Note that the original learner sentence does not reflect the intention by the learner: "geht weiter" could be literally translated as "goes on", but would describe an actual movement in this context. From the picture story it is evident that the learner wanted to describe that the man continues an action, therefore the text needs to be corrected.

The annotation of the tokmovid layers means an extra annotation effort, but we consider it rather low compared to the result: This information is mostly obvious to human annotators, but it can neither be easily recovered automatically later nor to the same extent. We also use this layer to explicitly annotate every movement of tokens. This way, we ensure that token movements can reliably be identified even if they are changed, e. g. due to spelling correction. For the Falko corpus, automatic identification of token movements was used. This has two disadvantages: A movement where the token was altered in the TH cannot be identified and a deletion of a token t in the original text and an insertion of token t in the TH would automatically be considered as a movement although they might be unrelated, e. g. consider deletion of an article in one noun phrase and insertion of the same article in another noun phrase.

The tokmovid layers do not influence the assignment of automatic error tags, e. g. a token which is moved and changed is tagged as `DEL` at the position where it was moved from and as `INS` at the position where it was moved to. However, the tokmovid can be queried in addition to find out whether the token was actually moved and changed.

We tried to adhere to the Falko annotation manual as far as possible since it already provides a reasonable and detailed rule set. Moreover, we wanted our annotations to be compatible with those of the Falko corpus to make comparison between learners productions in the ComiGS corpus to those of the Falko corpus possible. However, due to the differences between the tasks and language levels between both corpora, some minor changes or extensions to the guidelines were necessary which we documented in our annotation guide. For example, we do not discourage colloquial language for the TH2 in general. All in all, most of our changes can be regarded as an extension or clarification of the original guidelines so that the annotation is mainly compatible to that of the Falko corpus.

4.3 Annotation Process

The annotations were created by two annotators, annotator A and annotator B, using EXMARaLDA (Schmidt, 2004). We split the stories based on the learners into two sets, Set 1 and Set 2. Set 1 was jointly annotated by annotator A and B. Set 2 contains two independent annotations of THs, one by each annotator. The inter-annotator agreement was measured on Set 2 and it can be used as an evaluation set for systems trained on Set 1 as it contains two valid error corrections for each text.

We selected learners for Set 2 randomly but ensured that this set covers different mother languages and proficiency levels and also contains texts for the third story. Set 1 consists of 51 stories from 22 learners, totaling 12.5k tokens. Set 2 consists of 19 stories from 8 learners, totaling 5.4k tokens. The texts were automatically tokenized and segmented into sentences and the tokenization was manually corrected. Annotator A annotated TH1 and TH2 on Set 1 and met regularly with annotator B for discussions. Annotator B checked the annotations of annotator A and in case of disagreement discussed it with annotator A. Both annotators were forced to agree on one annotation. If necessary, the annotation was changed after the discussion. As a result, the annotations of the THs for Set 1 have been agreed upon by both annotators. While annotating Set 1, we developed the adaptation of the annotation guidelines for the THs. After annotating Set 1, both annotators annotated Set 2 independently with THs. The TH1 was annotated twice for measuring inter-annotator agreement and the TH2 was annotated twice for having two independent interpretations of the original learner utterance in Set 2.

Afterwards, one annotator annotated dependency trees including PoS tags and the other checked the annotation. The annotation was performed using the manual annotation interface of jwcdg (Beuck et al., 2013), a parser based on defeasible grammaticality rules. During the manual annotation, rule violations are displayed to highlight potential annotation errors. We did not annotate the same data twice with dependency structures and therefore did not perform an inter-annotator agreement evaluation on the dependency structures. Language learner sentences can be syntactically annotated with high inter-annotator agreement (91.5%) using the HDT annotation scheme even on the uncorrected learner sentences (Köhn et al., 2016). We assume that annotation of the already grammatically corrected sentences with explicit visual context can be annotated even more consistently.

4.4 Inter-annotator Agreement

The rules for TH1 are designed to make the TH1 structurally close to the original utterance. This and a good inter-annotator agreement is important to draw reliable conclusions about the original texts based on the TH1. Therefore, we measured the inter-annotator agreement on the minimal target hypotheses from Set 2 using Cohen's Kappa coefficient (Cohen, 1960). For the computation, token changes are always considered with respect to a token in the original text. A token from the original text is considered changed if there are any tokens on the TH1 layer between the end of the previous (original) token and the end of the current (original) token that differ from the current token[6]. A token is considered unchanged if the token is the same on the original and on the TH1 layer. Note that the κ coefficient does not measure agreement on the type of token change (in terms of Table 2) and that we use the manual alignment for the computation and not an automatic alignment based on extracted texts. The latter means that an insertion of token i followed by a deletion of token d by one annotator and a deletion of token d followed by an insertion of token i are considered as a disagreement although the resulting text is identical. Therefore, the κ values are likely to be an underestimation of agreement with respect to the resulting text.

Of the 5424 tokens in the original texts from Set 2, annotator A changed 1215 tokens and B 1248 while annotating TH1. Much more changes were made for TH2 than for TH1 (A: 2k changes, B: 1.9k changes). For TH1, 273 tokens were changed by one annotator but not by the other. If considering identical changes (agreement on which token to change and the correction) the annotators disagreed on 448 tokens. Overall, the κ coefficient for agreement on which tokens to change is 0.856 and for identical token change 0.765. We consider the agreement exceptionally high given the nature of language learner texts. Obviously, we cannot infer from this how precise the annotators worked because there are no comparable annotations available but the result shows that the clear annotations guidelines of TH1 in conjunction with a task-based corpus result in a reliable annotation.

Even the TH2 level, for which the annotation guidelines are less restrictive, is annotated with high agreement. The κ coefficient for identical token change on TH2 is 0.507 and agreement on which tokens to change is 0.728. For comparison, Dahlmeier et al. (2013) measured an average agreement of 0.3877 for the agreement of token change on the NUCLE corpus.

5 Conclusions and Outlook

We presented a task with a strong visual context which constrains the texts written by language learners without severely restricting their input and without linguistic priming. The task encourages learners to produce complex utterances. Despite this, the intentions of the utterances can be recovered quite well by annotators due to the visual context. This is reflected in the exceptionally high inter-annotator agreement for both the minimal and extended target hypothesis which shows that the annotation is reliable.

The corpus contains several layers of annotations for the texts, including a minimal and an extended target hypotheses and dependency annotations to facilitate different directions of research. The visual context is not only helpful for consistent annotation but could also be used as an additional input for automatic grammatical error detection and correction, e. g. by encoding the information in the picture stories into a knowledge base (cmp. Köhn and Menzel (2015)).

We make this corpus freely available and provide the software and instructions necessary to extend the corpus under `https://nats.gitlab.io/comigs`. We hope that this corpus will be useful for a variety of tasks which need learner data.

Acknowledgements

We are grateful to Inga Kempfert for annotating the corpus together with the first author and for implementing a part of the software for automatic processing and format conversion. We thank Maike Paetzel

[6] Additions after the last token are also added to the last token.

for the initial implementation of the data collection software and Robert Schmidt for extending it and for conducting the majority of the data collection. We thank Sven Zimmer for his work on the syntax annotation interface. We thank Wolfgang Menzel for helpful discussions on our data collection effort. We thank Nils Bernstein for his feedback on our task design and the other teachers for German as a Foreign Language at the Sprachenzentrum for forwarding our request to language learners. We thank Sebastian Beschke and the anonymous reviewers for their feedback. We thank the participants for their work.

References

Stacey Bailey and Detmar Meurers. 2008. Diagnosing Meaning Errors in Short Answers to Reading Comprehension Questions. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 107–115, Columbus, Ohio, USA, June. Association for Computational Linguistics.

Niels Beuck, Arne Köhn, and Wolfgang Menzel. 2013. Predictive incremental parsing and its evaluation. In Kim Gerdes, Eva Hajičová, and Leo Wanner, editors, *Computational Dependency Theory*, volume 258 of *Frontiers in Artificial Intelligence and Applications*, pages 186 – 206. IOS press.

Adriane Boyd. 2010. EAGLE: an Error-Annotated Corpus of Beginning Learner German. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the International Conference on Language Resources and Evaluation*, Valletta, Malta, May. European Language Resources Association (ELRA).

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada, July. Association for Computational Linguistics.

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

Maria da Luz Videira Murta. 1991. "Vater und Sohn" im Anfängerunterricht: Eine Hörverstehensübung und ein Schreibauftrag. *Fremdsprache Deutsch: Zeitschrift für die Praxis des Deutschunterrichts*, pages 46–47. Issue 5: Das Bild im Unterricht, Klett Edition Deutsch, München, Germany.

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, Atlanta, Georgia, June. Association for Computational Linguistics.

Markus Dickinson and Marwa Ragheb. 2009. Dependency Annotation for Learner Corpora. In Marco Passarotti, Adam Przepiórkowski, Savina Raynaud, and Frank Van Eynde, editors, *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*, pages 59–70, Milan, Italy, December.

Ana Díaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2):139–154. Special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair.

Franz Eppert. 2001. *Deutsch mit Vater und Sohn: 10 Bildgeschichten von E. O. Plauen für den Unterricht Deutsch als Fremdsprache*. Max Hueber Verlag, Ismaning, Germany, 1st edition.

Eileen Fitzpatrick and M.S. Seegmiller. 2004. The Montclair Electronic Language Database Project. In Ulla Connor and Thomas A. Upton, editors, *Language and Computers, Applied Corpus Linguistics. A Multidimensional Perspective.* Rodopi.

Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because Size Does Matter: The Hamburg Dependency Treebank. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Language Resources and Evaluation Conference 2014*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Kilian A. Foth, 2006. *Eine umfassende Constraint-Dependenz-Grammatik des Deutschen*. Fachbereich Informatik, Universität Hamburg. URN: urn:nbn:de:gbv:18-228-7-2048.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot, editors. 2009. *International Corpus of Learner English. Version 2. Handbook and CD-Rom.* Presses universitaires de Louvain, Louvain-la-Neuve.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2018. Near Human-Level Performance in Grammatical Error Correction with Hybrid Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 284–290. Association for Computational Linguistics.

Michael Hahn and Detmar Meurers. 2012. Evaluating the Meaning of Answers to Reading Comprehension Questions: A Semantics-Based Approach. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 326–336, Montréal, Canada, June. Association for Computational Linguistics.

Hagen Hirschmann, Anke Lüdeling, Ines Rehbein, Marc Reznicek, and Amir Zeldes. 2013. Underuse of Syntactic Categories in Falko – A Case Study on Modification. In S. Granger, G. Gilquin, and F. Meunier, editors, *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*, Corpora and Language in Use – Proceedings – 1. Presses Universitaires de Louvain.

Levi King and Markus Dickinson. 2013. Shallow semantic analysis of interactive learner sentences. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–21, Atlanta, Georgia, June. Association for Computational Linguistics.

Christine Köhn and Wolfgang Menzel. 2015. Towards parsing language learner utterances in context. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2015)*, pages 144–153, University of Duisburg-Essen, Germany.

Christine Köhn, Tobias Staron, and Arne Köhn. 2016. Parsing free-form language learner data: Current state and error analysis. In Stefanie Dipper, Friedrich Neubarth, and Heike Zinsmeister, editors, *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, number 16 in Bochumer Linguistische Arbeitsberichte, pages 135–145. Stefanie Dipper, Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, Universitätsstr. 150, 44801 Bochum, Bochum, Germany, September.

Anke Lüdeling. 2008. Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In Maik Walter Patrick Grommes, editor, *Fortgeschrittene Lernervarietäten*, pages 119–140. Niemeyer.

Niels Ott and Ramon Ziai. 2010. Evaluating Dependency Parsing Performance on German Learner Language. In Markus Dickinson, Kaili Müürisep, and Marco Passarotti, editors, *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*, volume 9 of *NEALT Proceeding Series*, pages 175–186.

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.

Marwa Ragheb and Markus Dickinson. 2011. Avoiding the comparative fallacy in the annotation of learner corpora. In *Selected Proceedings of the 2010 Second Language Research Forum: Reconsidering SLA Research, Dimensions, and Directions*, pages 114–124, Somerville, MA. Cascadilla Proceedings Project.

Marwa Ragheb and Markus Dickinson. 2012. Defining syntax for learner language annotation. In *Proceedings of COLING 2012: Posters*, pages 965–974, Mumbai, India, December.

Marwa Ragheb and Markus Dickinson. 2013. Inter-annotator agreement for dependency annotation of learner language. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179, Atlanta, Georgia, June. Association for Computational Linguistics.

Ines Rehbein, Hagen Hirschmann, Anke Lüdeling, and Marc Reznicek. 2012. Better tags give better trees – or do they? *Linguistic Issues in Language Technology (LiLT)*, 7(10), 1.

Marc Reznicek, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas, 2012. *Das Falko-Handbuch.*

Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing target hypotheses in the falko corpus. In Ana Ballier Díaz-Negrillo and Paul Nicolas Thompson, editors, *Automatic Treatment and Analysis of Learner Corpus Data*, pages 101–123. John Benjamins Publishing Company, Amsterdam, NLD.

Victoria Rosén and Koenraad De Smedt. 2010. Syntactic annotation of learner corpora. In K. Johansen, H.; Tenfjord, editor, *Systematisk, variert, men ikke tilfeldig : antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag*, pages 120–132. Novus.

Alla Rozovskaya and Dan Roth. 2014. Building a state-of-the-art grammatical error correction system. *Transactions of the Association for Computational Linguistics*, 2:419–434.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4:169–182.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universität Stuttgart / Universität Tübingen.

Thomas Schmidt. 2004. Transcribing and annotating spoken language with EXMARaLDA. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*. ELRA.