# Where it's at: Annotating Verb Placement Types in Learner Language

**Josef Ruppenhofer**[1], **Annette Portmann**[2], **Matthias Schwendemann**[2], **Christine Renker**[3], **Katrin Wisniewski**[2], **Torsten Zesch**[1],
[1]FernUniversität in Hagen, [2]Universität Leipzig, [3]Universität Bamberg,
**Correspondence:** torsten.zesch@fernuni-hagen.de

## Abstract

The annotation of learner language is often an ambiguous and challenging task. It is therefore surprising that in Second Language Acquisition (SLA) research, information on annotation quality is hardly ever published. This is also true for verb placement, a linguistic feature that has received much attention within SLA. This paper presents annotations of verb placement in German learner texts at different proficiency levels. We argue that as part of the annotation process target hypotheses should be provided as ancillary annotations that make explicit each annotator's interpretation of a learner sentence. Our study demonstrates that verb placement can be annotated with high agreement between multiple annotators, for texts at all proficiency levels and across sentences of varying complexity. We release our corpus with annotations by four annotators on more than 600 finite clauses sampled across 5 CEFR levels.[1]

## 1 Introduction

Acquiring the different options for verb placement, and more generally constituent order, in finite clauses of German is a well-known challenge for learners and a frequent object of second language acquisition (SLA) studies on German (Jordens, 1990; Diehl et al., 2000; Gunnewiek, 2000; Tschirner and Meerholz-Härle, 2001; Jansen, 2008; Czinglar, 2013; Baten and Håkansson, 2015; Wisniewski, 2020; Schlauch, 2022; Schwendemann, 2023).

One key reason to study verb placement is its theoretical significance for theory building. While learners' *interlanguage* (IL) has been found to be highly variable, it is also known to be systematic (Selinker, 1972). Processability Theory (PT) (Pienemann, 1998, 2005) posits that German verb placement options are acquired in a fixed order by all learners regardless of other factors, such as learners' age or educational background. This systematicity is attributed to the fact that it depends on the *processability* of the grammatical structures producing the different orders. These grammatical mechanisms build on each other to the effect that there is no skipping or re-ordering possible among the five major placement options that PT focuses on. Unsurprisingly, such strong claims are contested within the field of SLA (De Bot et al., 2007; Hulstijn et al., 2015). A second important reason to verify claims about the acquisition of verb placement empirically is that some common instruments for proficiency testing that are used in educational settings rely on verb placement as a key diagnostic (e.g. MIKA-D in Austria (Glaboniat, 2020; Blaschitz, 2023)): a theory whose application affects educational trajectories in the real world had better be sound.

Recently, Ruppenhofer et al. (2024) published specifications for the computational implementation of a system detecting verb placement types as a prerequisite for an automated analysis of learner (L2) language development on a large scale. However, that paper did not show that a key prerequisite for automation holds, namely that verb placement analysis can be performed reliably by human annotators. Moreover, as far as we could ascertain, agreement on verb placement analysis also has never been evaluated within SLA, where most studies on the topic seem to be based on single coding by one of the authors.

While the above specifications suggest that this should be an eminently doable task on proficient native (L1) data, we think it needs to be tested empirically how well human coders agree on verb placement in **learner text**, which is orthographically, semantically, and/or morpho-syntactically non-canonical. As an illustration, consider example (1).

[1] https://github.com/dakoda-project/annotating_verb_placement_with_ths

(1) Wann möchtst wir Treffen ?
when would-like we meeting ?
'When would you like (for) (us) to meet ?'

The last two tokens in (1) cannot combine if taken at face value: *wir* 'we' is a nominative case personal pronoun but *Treffen* 'meeting' is a noun. In this and other similar cases, any labeling of the learner data rests on adopting a particular interpretation of what the learner was trying to say.[2]

In the remainder of this paper, we argue for using an annotation protocol where verb placement annotations are performed in conjunction with the annotation of target hypotheses that can explicate the understanding of difficult learner productions such as (1). To that end, we present the design and results of an annotation study on essay data of L2 German learners at different levels of proficiency. We focus on the following research questions. How good is agreement between human annotators on verb placement overall? Can we observe differences related to the texts' proficiency levels (given in terms of CEFR ratings)? Is there an effect of sentence complexity on agreement?

## 2 Theoretical Background

To motivate our study design, we first present the SLA theory whose verb placement inventory we use for annotation and then discuss the use of target hypotheses in the analyses of learner data.

### 2.1 Processability Theory

The core of PT is the idea of a *processability hierarchy*. It encapsulates the idea that at least for some phenomena an acquisitional order from simpler to more complex structures results from the fact that the capabilities of the human language processor (Levelt, 1989) expand in a specific sequence as it develops new processing procedures for handling ever more advanced grammar rules. While the specific linguistic phenomena that exhibit fixed acquisition may differ across languages, the assumption is that all languages have phenomena of this kind because all languages must rely on grammatical processing procedures. In the case of German, verb placement is taken to be a core grammatical feature whose fixed acquisitional order is owed to the processability hierarchy. Table 1 illustrates the major patterns that Processability Theory (Pienemann, 1998) has focused on. These concern only finite clauses. In non-finite clauses, German verbs are always placed in final position so there is no variation to acquire. There also exist further minor finite sentence types with additional placement options. For instance, German allows so-called narrative verb-initial sentences. Since these minor sentence types are not the focus of the SLA literature, we set them aside here, too.

In **SVO** order, the verb is in second position, preceded by the S(ubject) and followed by an O(bject). **ADV**(erbial) is an order said to be used transitorily by learners (but ungrammatical in L1 German)[3], where an adverbial is placed before an SVO sequence for information structural reasons. **SEP**(aration) is a constellation that is used with complex verb clusters consisting of a finite modal or auxiliary in second position and a non-finite participle or infinitive in final position. Usually the finite and non-finite verbs are separated from each other by intervening arguments and/or modifiers. **INV**(ersion) is the L1-appropriate way of achieving the discursive ends intended by learners using ADV. But different from ADV, in INV the subject moves to the right of the verb so that only the adverbial remains to its left, which fulfills the constraint that in L1-German only one item should fill the preverbal slot. Once learners master INV, they no longer use ADV. The last placement type, **VEND** is used in subordinate clauses that are marked as such by subordinators or complementizers.

Note that some of the above placement types can co-occur. For instance, example (2) below shows both SEP(aration) of the finite and non-finite verbs *muss* and *suchen* and INV(ersion) of the subject pronoun *ich*. We refer the reader to Ruppenhofer et al. (2024) and their specifications for more discussion of such cases.

(2) Darum muss ich eine neue Wohnung
therefore must I a new apartment
suchen .
look-for .
'That's why I need to look for a new apartment.'

### 2.2 Annotating Target Hypotheses

Target hypotheses (THs) are a type of ancillary annotation that is often used in learner corpus linguistics. In that context, the TH makes explicit the aimed-for production the analyst assumes as

[2] We illustrate the multiple possible normalizations for example (1) below in Figure 1.

[3] Müller (2003) shows that there are some limited cases where ADV-like structures do occur in L1 German.

| Short Name | Description | Example |
|---|---|---|
| SVO | canonical word order | *Ich **suche** eine neue Wohnung .*<br>I look-for a new flat.<br>'I am looking for a new flat.' |
| ADV | adverb preposing | *Darum ich **suche** eine neue Wohnung .*<br>therefore I look-for a new flat .<br>'Therefore, I am looking for a new flat.' |
| SEP | verb separation | *Ich **muss** darum eine neue Wohnung suchen .*<br>I must therefore a new flat look-for .<br>'I have to look for a new flat.' |
| INV | inversion | *Darum **suche** ich eine neue Wohnung .*<br>therefore look-for I a new flat .<br>'Therefore, I am looking for a new flat.' |
| V-END | verb-final | *Weil ich eine neue Wohnung **suche** .*<br>because I a new flat look-for .<br>'Because I am looking for a new flat.' |

Table 1: Verb placement types in German (Pienemann, 1998) (**bold** = finite verb; underline = non-finite verb)

a reference when performing error annotation on a learner production (Lüdeling, 2008). For German as an L2, MERLIN (Boyd et al., 2014) and the Falko corpora (Lüdeling et al., 2008) are well-known resources that feature THs. The guidelines of the Falko project (Reznicek et al., 2012) in fact distinguish several types of THs. So-called minimal target hypotheses (called TH1) are supposed to feature only the minimal edits to make a learner production morpho-syntactically grammatical (and automatically parsable), though not necessarily idiomatic and contextually appropriate. Extended target hypotheses (aka TH2), by contrast, are less constrained: they also aim to make the utterance semantically and pragmatically appropriate to the context. In addition to TH1s and TH2s, the Falko corpus also features TH0 hypotheses. These are like their TH1 counterparts except that word order changes necessary for TH1 are undone. This means that TH0 may contain ungrammatical word orders. Table 2 provides an illustration.

Inter-annotator agreement for TH-based annotation has not been reported or discussed very much, as most corpora with any type of TH are only singly annotated. A notable exception is the ComiGs corpus of picture story retellings (Köhn and Köhn, 2018). It includes a subset of learner texts for which two annotators produced both a TH1 and a TH2 following the Falko guidelines. The authors report a high level of agreement with a $\kappa$ of 0.765 for which tokens on the learner text need to be changed. The reasons for the absence of multiple THs in most corpora likely are the time and cost required: the Falko guidelines for THs, for instance, span more than 20 pages.

The field of Grammatical Error Correction (GEC) distinguishes between reference normalizations that involve "minimal edits" (similar to Falko's minimal THs (TH1)) and reference normalizations that include "fluency edits" (similar to Falko's extended THs (TH2)). Of the datasets used in the recent Multilingual GEC shared task, most datasets only feature minimal edits and none seems to have multiple references at the same level of correction (Masciolini et al., 2025). To make up for the lack of multiple reference normalizations, the evaluation of GEC systems often uses reference-free metrics which enable the evaluation of model output without relying on a single (or, at best, a few) gold-standard references (Bryant et al., 2023).

## 3 Annotating Verb Placement with Ancillary Target Hypotheses

Broadly speaking, we can distinguish two types of difficult cases for verb placement analysis: (a) productions whose meaning is understandable but which are not obvious to normalize and (b) productions whose meaning is difficult to understand. Our introductory example (1) exemplifies the former situation: while we can understand the semantic import of the learner's utterance (especially in view of the task context of this production), the learner's production is syntactically incoherent and its normalization is not obvious.

Figure 1 shows several possible THs for the learner production in (1). The different THs themselves have different verb placement annotations and they lead to different conclusions about verb

| | | |
|---|---|---|
| L | Erstens gibt es viele Frage **muss** man im voraus zu überlegen.<br>firstly gives it many questions must one in advance to consider<br>'First, there are many issues that one has to think about in advance.' | |
| TH0 | Erstens gibt es viele Fragen, die **muss** man sich im Voraus überlegen. | raw word order |
| TH1 | Erstens gibt es viele Fragen, die man sich im Voraus überlegen **muss** . | corrected order |
| TH2 | Erstens gibt es viele Fragen, *über* die man im Voraus *nachdenken* **muss** . | fluency edit (italics) |

Table 2: Example with three levels of target hypotheses from Falko L2 corpus (fu129_2006_10a)

placement on the learner layer. The first target hypothesis, TH-a, treats *wir* as an erroneous realization of the accusative form *uns* and interprets *Treffen* as an erroneously capitalized infinitive form rather than as a noun. In addition, the TH adds a subject pronoun *du* to make the sentence grammatical. Accordingly, the clause shows SEP(aration) between the finite verb 'möchtst' and the non-finite verb 'Treffen'. This also applies to the learner layer, which has counterparts for both verbal tokens as well as an intervening token. However, since the learner layer lacks a subject, it cannot be labeled as an instance of INV(ersion). By contrast, TH-b treats 'Treffen' as a noun and exhibits INV because the sole finite verb is followed by its subject and preceded by a non-subject. However, because the learner layer lacks a post-verbal subject, it cannot be labeled as INV. In fact, none of PT's labels applies.

An example of the second type of difficult case is found in (3). Here the verb *sagen* may or may not be taken to have a complement clause (cf. possible interpretations a-c). Depending on how the two finite verbs/clauses relate, we make different assumptions about the type of clause and verb placement we need to assign to the finite form *wurde*.

(3) und Sie sagen mir gut Konzert wurde 18
and she say me good concert became 18
märz.
March

(a) 'And she tells me there is a good concert on March 18th.'
(b) 'And she tells me okay. The concert was on March 18th.'
(c) 'And she tells me if the concert on March 18th turned out to be good.'

Given cases such as (1) and (3), it seems unavoidable to explicate coders' target hypotheses: simply comparing annotations on the learner layer without reference to THs risks making the annotations appear less valid and reliable than they might be. As a correlate, for instances where multiple THs are plausible, multiple gold standards for verb placement must be entertained.[4]

Beyond explicating the understanding attributed to the tokens on the learner layer, THs serve a second function that is important within the language acquisition context: they spell out the structure that was expected in context. For instance, in (4), the learner uses SVO (verb-second) in the complement clause. A possible TH for this clause would re-order it to final placement of the finite verb (*daß immer mehr Menschen lieber alleine als in einer Großfamilie leben*).

(4) ... so kann man Sagen, [ dass immer mehr
... so can one say, [ that always more
Menschen **leben** lieber alleine als in
people live preferably alone than in
einer Großfamilie ].
a big-family ].

' ..., then we can say that more and more people prefer living alone to living in an extended family.'

By the logic of Processability Theory, a data point such as (4) serves as a piece of negative evidence, suggesting that the learner has not mastered verb-final placement as they fail to use it in a context where it ought to be used. Without THs, no such evidence is available.

### 3.1 Source Data

The data on which we carried out our study comes from the MERLIN (Boyd et al., 2014) and DISKO (Wisniewski et al., 2022) corpora. Both of them include written texts, specifically essays, for which a manual CEFR rating is available. We used MERLIN data to represent the lower CEFR levels A1, A2, and B1, while we sample DISKO for more advanced B2 and C1 data.[5] We did not include texts rated as C2 since they are too few in number and of lesser interest as the acquisition of verb placement likely is completed prior to that level of proficiency.

[4]While we are concerned directly only with the analysis of verb placement, the idea of capturing multiple acceptable analyses of learner language should be relevant to learner language tree-banking in general.

[5]We consider the proficiency level TDN3 of the DISKO corpus to be equivalent to B2 for our purposes, whereas DISKO's level TDN5 serves as comparable to CEFR-level C1.

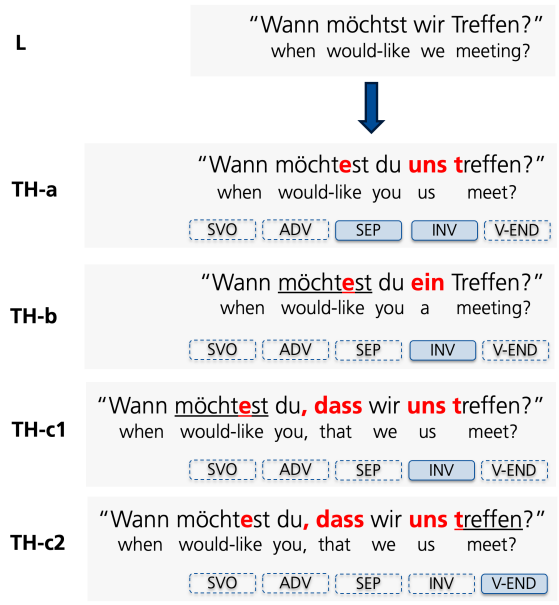



Figure 1: Different annotators might come up with different target hypotheses potentially leading to different analyses regarding verb placement. Note that TH-c produces two analyses because it assume two finite verbs/clauses.

The MERLIN corpus contains texts produced as part of standardized tests. The most common L1s in the German part of MERLIN[6] are Russian, Polish, Hungarian, French, and Spanish. The DISKO corpus contains language tests taken by L2 speakers studying at German universities. The most common L1s in DISKO are Russian, Arabic, and Spanish.

All annotations are performed on the learner data (abbreviated as **L**) as well as the annotated target hypothesis (**TH**). If the learner sentence was grammatical, the target hypothesis (and the resulting annotations) is usually a copy.

### 3.2 Type of target hypothesis to aim for

In the context of our annotation of verb placement types in the DAKODA project, annotators were instructed to produce target hypotheses that (i) reflect their interpretation of the learner text, (ii) are grammatical, and (iii) make minimal changes. They were, however, given no further criteria for which 'edit operations' they should consider more or less costly but instead use a holistic approach when weighing alternatives. Our instructions thus match neither the minimal (TH1s) nor the extended target hypotheses (TH2s) defined by Falko (Reznicek et al., 2012). While TH1s emphasize criteria (ii) and (iii), they may ultimately not reflect the contextually understood interpretation of a learner utterance in the interest of staying close to the lexico-syntactic material the learner provided. TH2s, by contrast, often don't observe desideratum (iii) and make more fluency edits than we would like to see from the annotators. For instance, for our purposes verbal constructions should not be replaced by nominal ones or vice versa. Nor should finite and non-finite constructions be switched, even at the cost of idiomaticity.

Our annotators were aware of the general 'downstream' analytic interest in verb placement, but they were not explicitly told to adhere to any additional desiderata such as the ones about preserving (finite) verbs. By refraining from imposing specific rules for which kinds of normalizations to prefer, we hoped to avoid suppressing alternative possible interpretations and alternative normalizations. Note that the TH guidance we used should not be seen as a poor man's approximation of TH1s: we purposely deviate from the Falko guidelines to enforce more faithfulness to interpretation than TH1s do, while allowing somewhat more formal variation than THs1 allow (but still less than TH2s do).[7]

The resulting data thus allows one to study how often annotators converge on the same or similar THs even without detailed guidance. This approach may be of interest for other research settings where the creation of highly controlled THs is not feasible.

### 3.3 Annotation Process

We split the annotation into 6 rounds. Per round, we asked for 100 finite clauses to be identified and annotated. For each round, we provided the annotators with a series of randomly sampled texts within which they were asked to perform a set of annotation steps (explained in the next paragraph) on the learner text until they had reached 10 finite clauses from the start in a given document. Limiting the annotation to at most 10 clauses from a given document/learner was done so as not to bias results to any particular learner. If a document contained fewer than 10 clauses, annotators were asked to annotate additional clauses in another document.

[6]The overall MERLIN corpus is trilingual with German, Italian, and Czech as targets of language acquisition.

[7]While we also hoped to see, as a welcome side effect, a speedup of TH construction relative to using the detailed Falko guidelines for TH1s, we did not perform an empirical comparison and thus do not know if any time savings materialized.

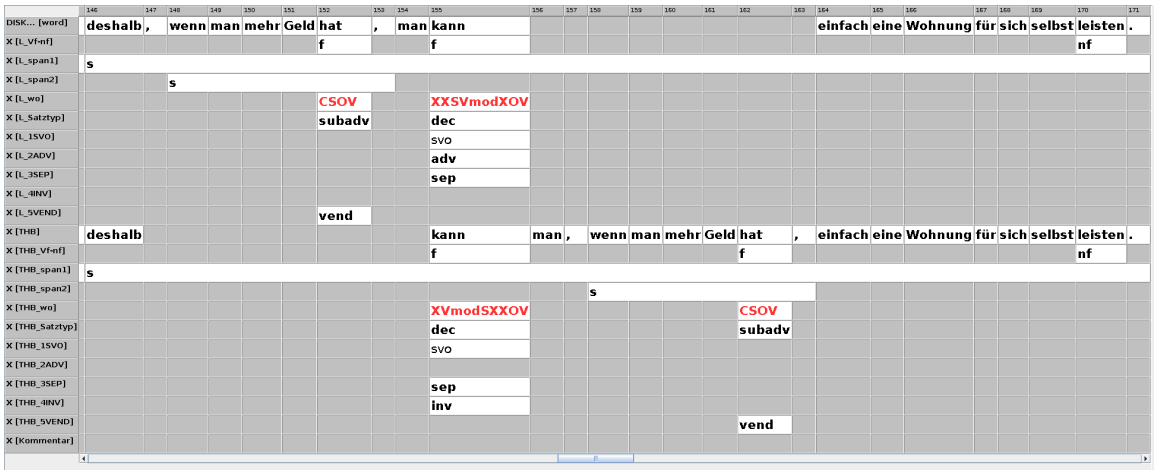



Figure 2: Annotation in Exmaralda
'If you have more money, you can readily afford a place of your own.'

Each round included documents from each of the five CEFR levels under consideration. Overall, data is drawn from 66 distinct documents.

**Annotation Steps**

- segment the text into sentences and clauses (as needed)
- identify any verbal forms and mark them as finite (f) or non-finite (nf)
- classify finite clauses into predefined sentence types (cf. Appendix A)
- record the ordering of the major constituents in each finite clause
- provide one or more labels characterizing the verb placement in a finite clause (cf. section 2.1)

Note that the annotators ran through the above annotation steps in one go. That is, we did *not* create an adjudicated set of finite verb instances before letting annotators proceed to the sentence type and verb placement analysis.[8] This choice was made with the expectation that agreement would be high for identifying finite verbs anyway.

**Tool** We used Exmaralda[9] (Schmidt and Wörner, 2014) because some of our annotators had prior familiarity with it and because our corpora are available in a format that Exmaralda can read. As we did not want to carry over any bias from automatic tools, the annotators worked on raw text, that is, they had no access to any manually or automatically assigned POS-tags or lemmas etc. For that reason, we explicitly asked for the annotations related to clause and verb identification in addition to verb placement labels.

Figure 2 shows a screenshot of annotations on a text from the DISKO corpus. In the example, the target hypothesis involves a reordering and the analysis of the matrix clause headed by the modal *kann* differs accordingly: for instance, while the learner clause exhibits ADV, the TH clause features INV.

**Annotators** We had 4 annotators ranging from master's students to post-docs with expertise in the area of German as a foreign or second language and familiarity with PT. They met to discuss questions after every round of annotation. A subgroup of two annotators finally produced an adjudicated gold standard. Importantly, this gold standard allows for multiple correct labels if they result from target hypotheses with different clausal orders.

## 4 Annotation Analysis

In the final dataset, we have 849 tokens annotated as verbs on the learner layer **L**. On the target hypothesis layer **TH**, we have 847 instances. Table 3 gives the breakdown per CEFR level. As we have complex sentences in our data even on the lower levels, we reached more than the 600 verb instances to be expected if we only had atomic finite clauses.

Figure 3 shows the combinations of sentence type and verb placement found on the learner layer. What we observe are mostly combinations that would be expected for German. For instance, INV(ersion) structures are commonly found in questions and declaratives, while verb-final (VEND) structures are found exclusively in

[8] In other words, unitizing was not completed before categorization in the sense of (Mathet et al., 2015).

[9] www.exmaralda.org

| | L | | TH | |
|---|---|---|---|---|
| Level | # verbs | % finite | # verbs | % finite |
| A1 | 159 | .74 | 158 | .73 |
| A2 | 152 | .73 | 151 | .74 |
| B1 | 161 | .70 | 162 | .70 |
| B2 | 173 | .68 | 173 | .68 |
| C1 | 204 | .73 | 203 | .73 |

Table 3: Total verb instances per CEFR level

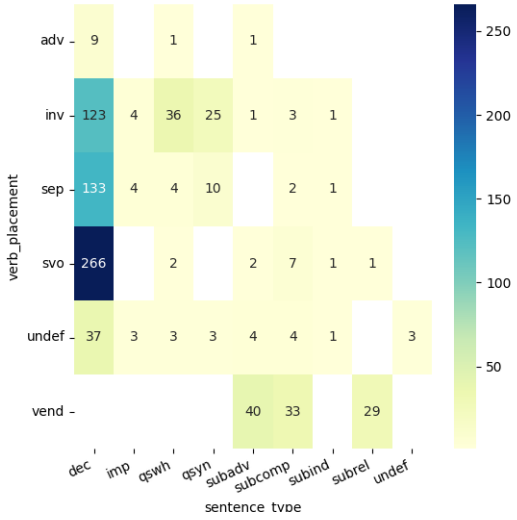



Figure 3: Combinations of sentence type (cf. Appendix A) and verb placement (cf. section 2) on the Learner layer

subordinate clause types. However, we can also observe some unexpected combinations involving SVO in various types of subordinate clauses.

### 4.1 Overall agreement

We first consider overall agreement per layer. Table 4 shows Fleiss $\kappa$ values for 4 annotators calculated using the python re-implementation of the IRR_CAC package.[10] Importantly, as we had expected, agreement is very high for identifying finiteness. And in fact, agreement is also high for sentence type and verb placement, with surprisingly small differences between the two layers. The high agreement on annotations based on THs suggests that ancillary THs formulated without detailed Falko-style guidelines are adequate for our task.

### 4.2 By CEFR level

To address our second research question, we analyze the level of agreement obtained for texts with

[10] https://github.com/afergadis/irrCAC

| | L | TH |
|---|---|---|
| finiteness | .97 | .98 |
| sentence type | .84 | .85 |
| verb placement | .83 | .83 |

Table 4: Overall agreement on learner text (**L**) and target hypothesis (**TH**) in terms of Fleiss' $\kappa$

different **proficiency** levels to see if there is evidence for either of two seemingly conflicting intuitions. On the one hand, agreement might get better, the higher the proficiency level gets because more proficient texts are more grammatical and understandable. On the other hand, the constructions found in lower-proficiency texts may exhibit less variance and may be simpler, making clauses easier to analyze.

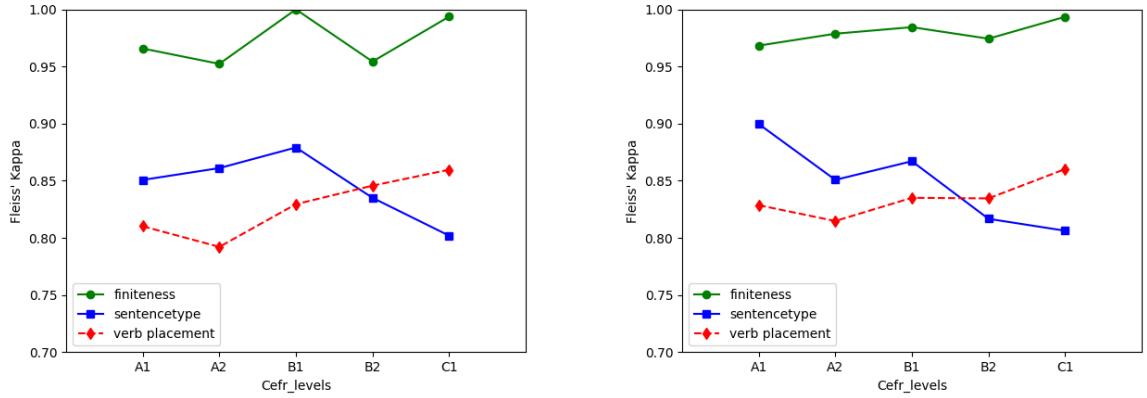
Figure 4 provides plots for agreement by CEFR level. For the learner data, agreement on finiteness is high throughout, with a peak for documents at level B1. On the target hypothesis layer, the results are similar but the peak at B1 is absent.

For the annotation of sentence type on the TH layer, the texts at level A1 yield higher agreement than those at level B1, whereas on the learner layer the peak is at level B1. This may be due to non-target language-like characteristics of early learners' L2 German, whereas on L1 German the annotation of sentence type becomes more difficult, the more sophisticated the texts become. The finding for early L2 German learners might seem counter-intuitive at first sight. Since beginning learners make more errors, one might expect that it would be more difficult to agree on a common interpretation. However, early learner's language is also characterized by a smaller repertoire with a large proportion of ready-made chunks. This might constrain the range of interpretational options for annotators and thus make it an easier task to agree on annotations.

For verb placement, agreement improves slightly across levels for both learner and TH layers. On the learner layer, there is a dip for the highest level. However, overall the differences between CEFR levels do not seem very pronounced, which potentially means that both intuitions apply at the same time: we get fairly steady high agreement, though for different reasons at different levels.

### 4.3 By complexity

Addressing our third research question, we want to see if sentence **complexity**, operationalized here




(a) Learner (**L**)

(b) Target Hypothesis (**TH**)

Figure 4: Agreement by CEFR level

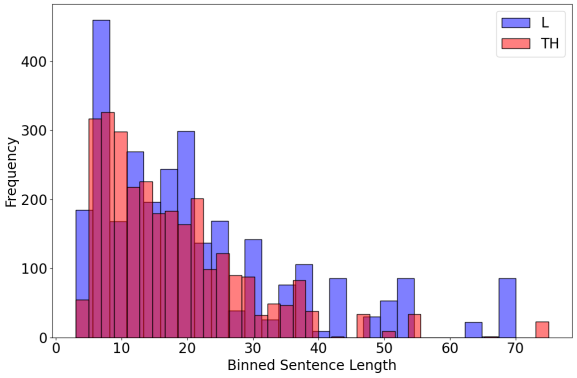



Figure 5: Distribution of sentence lengths

in rough terms as the number of tokens, influences agreement. Note that we use *complexity* here in the sense of (Bulté et al., 2024) as focused on formal features of linguistic items , in contrast to *difficulty*, which refers to items' cognitive load.

Figure 5 shows the right-skewed distribution of sentence lengths in both the learner and the TH layers. Most outliers at the end of the long tail are owed to the learner layer. Re-segmentation on the TH layer eliminates many of them.

We split the annotated instances into 10 bins of equal size. Figure 6 shows the agreement results for **L** and **TH**, respectively. Agreement on finiteness is a bit lower for the shorter sentences on the learner layer than on the TH layer. Agreement on sentence type trends downward as sentences get longer. For verb placement, agreement peaks for the 4th bin (median sent. length 12) on the leaner layer but for the 7th bin (median length 21) on the TH layer.

Notably, for both sentence type and verb placement, results are lower on the TH layer for the longest sentences than on the learner layer. This may be due to the fact that during the creation of target hypotheses the material could be re-segmented. This eliminated many long "sentences" that lack correct punctuation in the learner text. The long sentences that remain on the TH layer are complex ones that are harder to analyze.

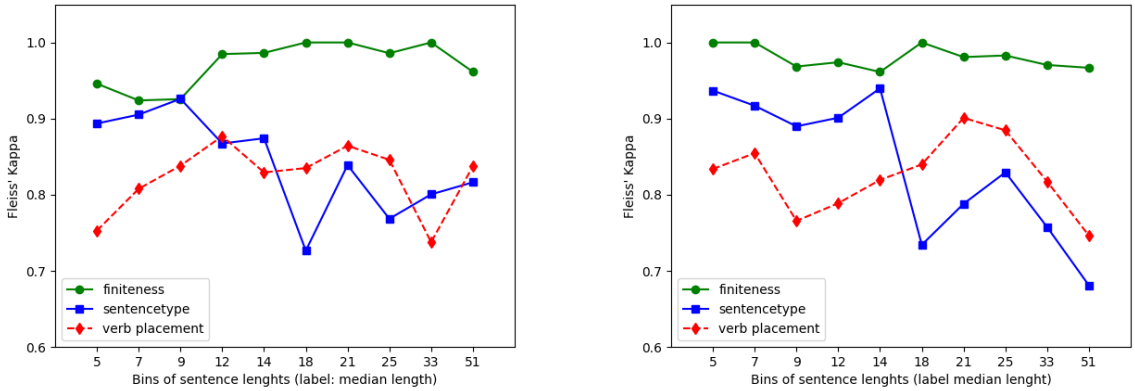
### 4.4 Illustration of disagreements regarding verb placement

Some disagreements result from unclear grammatical relations.[11] In example (5), the token *alle* is mismatched with the verb *geht*. On one analysis, the author aimed for *allen geht es sehr gut*, where *allen* is an indirect object; on another, the author aimed for *alles geht sehr gut*, with *alles* as a subject.

(5) ich Hoffe alle **geht** sehr gut .
i hope all goes very well
'I hope everybody is doing very well. / I hope everything is going well'.

Other disagreements regarding verb placement are downstream of disagreements about whether a token is verbal or not. Example (6) is, even in its full context, very hard to make sense of. Some annotators treated *sein* as a non-finite form of the verb *sein* 'to be' that is in construction with the finite form *ist* 'is', while others didn't treat it as a verb but rather as the homophonous and homographic possessive determiner 'his'. On the first analysis, we observe an instance of a verbal bracket (SEP) , on the second analysis we do not.

[11] For discussion of disagreements about finiteness and sentence type, we refer the reader to appendix D.




(a) Learner (**L**)

(b) Target Hypothesis (**TH**)

Figure 6: Agreement by sentence length

(6) wann ist deine Kinder **sein** .
when is your children {be/his}

Another group of disagreements includes cases such as (7) where one could either recognize a lexicalized separable prefix verb (e.g. *gutgehen*) that gives rise to a bracket when the parts are separated, or a compositional use where a simple verb (e.g. *gehen*) is modified or complemented by an adverb.

(7) Wie gehtt's dir, mir geht **gut** und meine
how goes you, me goes good and my
famile auch .
family also .

'How are you doing? I'm well and my family is, too.'

Finally, we find cases of ambiguity between two verb placement types, for instance, between INV and ADV. In (8) the issue is whether the first token, *so*, is a modifier for the date phrase ('circa in 1975') or a clausal adverb ('Thus/therefore, in 1975 . . . '). On the first analysis, there is only one preverbal constituent and the sentence exhibits INV. On the second analysis, there are two preverbal constituents and the sentence exhibits ADV.

(8) So im Jahr 1975 **bestanden** fast die Häfte
so in year 1975 consisted almost the half
von der Haushälte in Deutschland aus 3
of the households in Germany out-of 3
und mehr Personen .
and more persons .

'Thus/Circa in the year 1975 almost half the households consisted of three or more persons.'

## 5 Conclusion

Our corpus – the Multiply annotated verb placement corpus (MAVPC) – is the first dataset for SLA studies where verb placement is multiply coded and where target hypotheses are available as ancillary annotation rationales. We have shown that on essay data sampled from two corpora and stratified across CEFR levels, high levels of agreement could be achieved for the core annotation categories of finiteness, sentence type, and verb placement. This holds both on the raw learner text and on the THs. The corpus features not only the raw annotations of four annotators but also one or more gold standard labels that reflect contextually plausible interpretations of clausal structure and verb placement. The data can serve as a test set for automatic systems performing verb placement analysis.

While the high agreement on the Learner layer might suggest that THs are not needed at all, we would caution against that conclusion. The concomitant annotation of THs may improve agreement on the learner layer in a way that might be absent if no THs were constructed. Also, our data represents just one written text type and a limited set of L1s. Further studies on additional written text types and especially on spoken language are needed.

## Acknowledgements

The authors were supported by BMBF (German Federal Ministry of Education and Research) as part of the project DAKODA (Datenkompetenzen in DaF/DaZ: Exploration sprachtechnologischer Ansätze zur Analyse von L2-Erwerbsstufen

in Lernerkorpora des Deutschen [Data competencies in German as Foreign Language / German as a Second Language: Exploring language technology-based approaches to the analysis of L2 acquisition levels in learner corpora of German]), grant number 16DKWN035.

We would also like to thank Jamila Bläsing and Iulia Sucutardean for their support in annotating the data. Finally, we thank Nils Reiter for a fruitful exchange on the details of our agreement evaluation.

## Limitations

The annotation carried out as part of this study covers only two corpora of learner essays. While we suspect that agreement would also be quite high in other written task settings, it is unclear just how well the findings would generalize. More significantly, this study does not include any transcripts of spoken learner language. Spoken language data, unlike our essay data, usually comes without punctuation and is transcribed not in terms of sentences or clauses but in terms of utterances or turns. Accordingly, manual annotation of verb placement on such data would be liable to exhibit disagreements resulting from differences in segmentation. In addition, spoken language transcripts contain disfluencies such as hesitations and repetitions which would have to be consistently factored into or out of the annotations. Further, since L1 spoken language admits certain structures that would be ungrammatical in the written modality, annotators should then not correct such structures on L2 data in their target hypotheses.

Our approach to TH creation relied on very little detailed guidance. While we think that that approach could be suitable for other research contexts, too, we acknowledge that it may limit the usefulness of the resulting annotations for re-use in research that requires high internal consistency across the breadth of grammatical phenomena.

## References

Kristof Baten and Gisela Håkansson. 2015. The development of subordinate clauses in German and Swedish as L2s : a theoretical and methodological comparison. *Studies in Second Language Acquisition*, 37(3):517–547.

Verena Blaschitz. 2023. "Zeig mir bitte: Banane" – kritische (sprach-)wissenschaftliche Anmerkungen zum Deutschscreening "MIKA-D" [Please show me: banana – critical remarks from (language) science regarding the German language screening instrument MIKA-D ]. *ÖDaF-Mitteilungen*, 39(12):174–197.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

Bram Bulté, Alex Housen, and Gabriele Pallotti. 2024. Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*, n/a(n/a).

Christine Czinglar. 2013. *Grammatikerwerb vor und nach der Pubertät: Eine Fallstudie zur Verbstellung im Deutschen als Zweitsprache [Grammar acquisition before and after puberty: A case study on verb placement in German as a second language]*. De Gruyter Mouton.

Kees De Bot, Wander Lowie, and Marjolijn Verspoor. 2007. A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition*, 10(1):721.

Erika Diehl, Helen Christen, and Sandra Leuenberger. 2000. *Grammatikunterricht: Alles für der Katz? Untersuchungen zum Zweitsprachenerwerb Deutsch [Teaching Grammar: All For Naught? Investigations into the Second Language Acquisition of German]*, 1. edition. Niemeyer, Tübingen.

Manuela Glaboniat. 2020. MIKA-D. Eine Betrachtung aus testtheoretischer Perspektive [MIKA-D. An examination from a test-theoretic perspective]. *ide - informationen zur deutschdidaktik*, 4:61 73.

Lisanne Klein Gunnewiek. 2000. *Sequenzen und Konsequenzen: zur Entwicklung niederländischer Lerner im Deutschen als Fremdsprache [Sequences and consequences: On the Development of Dutch Learners of German as a Foreign Language]*. Rodopi.

Jan H. Hulstijn, Rod Ellis, and Søren W. Eskildsen. 2015. Orders and sequences in the acquisition of l2 morphosyntax, 40 years on: An introduction to the special issue. *Language Learning*, 65(1):1–5.

Louise Jansen. 2008. Acquisition of German Word Order in Tutored Learners: A Cross-Sectional Study in a Wider Theoretical Context. *Language Learning*, 58(1):185–231.

Peter Jordens. 1990. The acquisition of verb placement in Dutch and German. *Linguistics*, 28(6):1407–1448.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Willem J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. The MIT Press.

Anke Lüdeling. 2008. Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora [Ambigutities and Categorization: Problems in the Annotation of Learner Corpora]. In Maik Walter and Patrick Grommes, editors, *Fortgeschrittene Lernervarietäten*, pages 119–140. Max Niemeyer Verlag.

Anke Lüdeling, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt, and Maik Walter. 2008. Das lernerkorpus falko [the falko learner corpus]. *Deutsch als Fremdsprache*, 45(2):67–73.

Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, and Robert Östling. 2025. The MultiGEC-2025 shared task on multilingual grammatical error correction at NLP4CALL. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–33, Tallinn, Estonia. University of Tartu Library.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma ($\gamma$) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Stefan Müller. 2003. Mehrfache Vorfeldbesetzung. *Deutsche Sprache*, 31(1):29–62. `https://hpsg.hu-berlin.de/~stefan/Pub/mehr-vf-ds.html`.

Manfred Pienemann. 1998. *Language processing and second language development. Processability theory*. Benjamins,.

Manfred Pienemann. 2005. An introduction to Processability Theory. In Manfred Pienemann, editor, *Cross-Linguistic Aspects of Processability Theory*, pages 1–60. John Benjamins.

Marc Reznicek, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas. 2012. Das Falko-Handbuch. Korpusaufbau und Annotationen. Technischer Bericht, Humboldt-Universität zu Berlin. Version 2.01.

Josef Ruppenhofer, Matthias Schwendemann, Annette Portmann, Katrin Wisniewski, and Torsten Zesch. 2024. Every verb in its right place? a roadmap for operationalizing developmental stages in the acquisition of L2 German. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6655–6670, Torino, Italia. ELRA and ICCL.

Julia Schlauch. 2022. Erwerb der Verbstellung bei neu zugewanderten Seiteneinsteiger:innen in der Sekundarstufe. Eine Fallstudie aus dem DaZ-Lerner:innenkorpus SeiKo [Aquisition of verb placement among newly arrived immigrants: Lateral entrants in secondary school. A case study based on the German-as-a-second-language learner corpus SeiKo]. *Korpora Deutsch als Fremdsprache*, 2(2):4362.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *The Oxford Handbook of Corpus Phonology*. Oxford University Press.

Matthias Schwendemann. 2023. *Die Entwicklung syntaktischer Strukturen [The development of syntactic structures]*. Studien Deutsch als Fremd- und Zweitsprache. Erich Schmidt Verlag GmbH & Co. KG Berlin.

Larry Selinker. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1-4):209–232.

E. Tschirner and B Meerholz-Härle. 2001. Processability Theory: Eine empirische Untersuchung [Processability Theory: An empirical investigation]. In K. Aguado and C. Riemer, editors, *Wege und Ziele: Zur Theorie, Empirie und Praxis des Deutschen als Fremdsprache (und anderer Fremdsprachen). Festschrift für Gert Henrici*, pages 155–175. Schneider, Hohengehren.

Katrin Wisniewski. 2020. SLA developmental stages in the CEFR-related learner corpus MERLIN: Inversion and verb-end structures in German A2 and B1 learner texts. *International Journal of Learner Corpus Research*, 6(1):1–37.

Katrin Wisniewski, Elisabeth Muntschick, and Annette Portmann. 2022. Schreiben in der Studiersprache Deutsch: Das Lernerkorpus DISKO [Writing in German as a Language of Instruction: The DISKO learner corpus]. In K. Wisniewski, W. Lenhard, J. Möhring, and L. Spiegel, editors, *Sprache und Studienerfolg bei Bildungsausländer/-innen*. Waxmann, Münster.

## A Annotation of sentence types

The sentence type definitions in Table 5 are meant to apply only to *finite* clauses because PT's theorizing about verb placement does not include non-finite clauses. Thus, though German allows e.g. the use of infinitives and participles as imperatives, such constructions are not part of our annotation. Finally, note that while PT makes no explicit reference to sentence types in defining the verb placement types, previous findings point to a potential effect of sentence type on acquisition order (Diehl et al., 2000).

| | |
|---|---|
| imp | imperative |
| dec | declarative main clause |
| qswh | matrix wh-questions |
| qsyn | matrix yes/no-question |
| subadv | adverbial clauses |
| subcomp | complement/object clauses |
| subind | embedded interrogative clauses |
| subrel | relative clauses |
| undef | other |

Table 5: Sentence types

## B Verb placement and developmental stages within Processability Theory

Within Processability Theory, categorizing the placement of verb tokens in learner text is done in service of determining the learners' so-called developmental stage. For instance, as noted in the body of the text, a learner who has mastered INV is more advanced than one who uses ADV. One important question is how mastery is assessed. Here, PT employs a so-called emergence criterion: a stage counts as acquired by an individual learner if some $N$ instances are produced in contexts where the relevant verb constellation is expected by L1 standards, so-called obligatory contexts.

To exclude formulaic language and repetition from counting towards emergence, often a lexical diversity criterion for verbs is employed.

For instance, if INV placement is observed with only one verb that is less clear evidence that INV has been acquired than if instances were found for $M$ verbs, where $M$ usually is $\geq$3. The exact values of $N$ and $M$ vary somewhat in the PT literature.

Two considerations are important here. First, high overall accuracy is not required for emergence (cf. Wisniewski (2020)). Second, given how few learners figure in some corpora and how short their texts are, conclusions on individual learners or a cohort may be quite significantly influenced by a few verb tokens being categorized one way or another. For that reason we argue that at least the data should be public , if at all possible, and target hypotheses should be created to explicate the understanding of the learner layer.

## C Additional agreement results

### C.1 By round of annotation

We look at the development of agreement across rounds of annotation to see if we can observe a **training effect**. Our baseline assumption is that agreement will rise across successive rounds. Figure 7 shows the results for the learner layer and the target hypothesis. The level of agreement overall is high and the trends are broadly similar for both layers. The annotation of finiteness is always easiest. The annotation of sentence type tends to have higher agreement than that for verb placement. For verb placement on the learner layer, we find continual improvement through round 5 after an initial dip, and a slight drop-off for the last round. On the target hypothesis layer, the climb close to peak performance happens earlier.

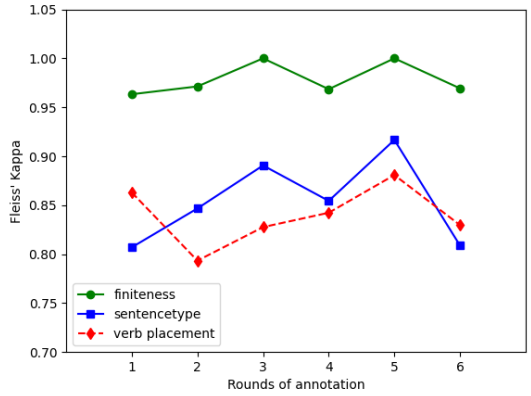
### C.2 Agreement by number of ratings

Some verbal instances in the dataset were not completely labeled on all layers by all annotators. We therefore wanted to see if the lacking annotations might reflect a greater difficulty of the relevant items. Figure 8 plots agreement depending on how many ratings the items minimally received. The figure suggests that agreement on the full dataset, where items were annotated by as few as 2 persons is, in fact, slightly better than on the subset where each item was annotated by everybody. We therefore think that the lacking annotations mostly result from the fact that we had no consistency enforcement in our annotation tool to make sure that items that were labeled as finite also received labeling on other layers. The setup thus allowed oversights to go unnoticed.

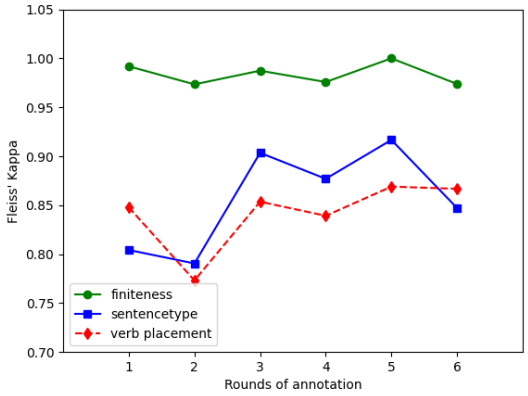
On the target hypothesis layer, we find the same trend as on the learner layer (cf. Fig 9).

## D Further illustrations of annotator disagreements

**Finiteness** Disagreements with regard to **finiteness** are very rare overall. One subset of these cases represents instances where some annotators do not treat a token as verbal at all, while others do recognize a verb. For example (9), one subset




(a) Learner (**L**)




(b) Target Hypothesis (**TH**)

Figure 7: Agreement by round of annotation

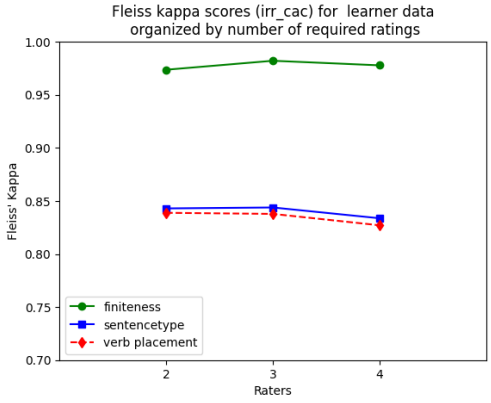



Figure 8: Agreement on learner layer for different numbers of required ratings

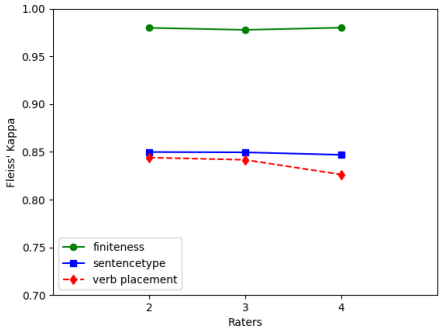



Figure 9: Agreement on target hypothesis layer for different numbers of required ratings

of annotators treated the token *besoche* 'visit' as a finite verbal form, whereas the second group of annotators treated it as a nominal form governed by the verb *nehme* 'take'.

(9) Ich nehme **besoche** meine Tochter .
I take visit my daughter .
'I visit my daughter .'

Example (10) is a case where all annotators perceive the token in question, *kommen* 'come', as verbal but differ as to finiteness. The disagreement is plausible since the *when*-clause lacks a subject, which normally suggests a non-finite construction. On the other hand, temporal adverbial clauses marked by *wann* 'when' ought to be finite according to the grammar of L1 German.

(10) Bringst du mir mit wann du hier in
bring you me with when you here in
Deutschland **kommen** . .
Germany come.
'You'll bringt it to me when you come here to Germany .'

**Sentence type** Disagreements with respect to sentence type may result from the tension between a sentence's form and its illocution. In (11), the sentence employs INV(ersion) as is appropriate for a yes/no question but the utterance is clearly a request.

(11) **Küsst** du für mich deine Kinder . .
kiss you for me your children.
'Kiss your children for me .'

The annotators were supposed to annotate based on form type (i.e. they should all have preferred

the yes/no question analysis for 11) but they did not always manage to overrule conflicting signals from illocution.

A significant group of disagreements involve subordinate clauses with unexpected word order. In example (12), the token *leben* 'live' occurs in an object clause marked by the complementizer *dass* 'that' and governed by the verb *sagen* 'say'. The expected word order for that constellation is verb-final (VEND) but in fact *leben* seems to occupy the second position as would be appropriate for either a matrix clause or a complement clause without a complementizer. Matching the overall structure, one subgroup of annotators (correctly) recognized an object clause whereas another group annotated a matrix declarative, following the signal given by the word order.

(12) Betrachtet man die Entwicklung der letzten
considers one the development the last
Jahren so kann man Sagen, dass immer mehr
years so can one say, that always more
Menschen **leben** lieber alleine als in
people live preferably alone than in
einer Großfamilie .
a big-family.

'If we consider the developments of recent years, then we can say that more and more people prefer living alone to living in an extended family. .'

Another example is shown in (13), where a sentential relative clause exhibits main clause word order rather than verb-final order. Some annotators chose the relative clause analysis that fits the overall context while others chose an analysis as a declarative sentence that is consonant with the clause-internal word order.

(13) In mein Heimatland LandX , wohnen
In my home-country countryX , live
immer viele Menschen in einem Haushalt
always many people in one household
manchmal sogar eine ganze Familie was
sometimes even a whole family which
**führ** zu eine Hilfsbereite und relativ
lead to a helpful and relatively
Tolerante Gesellschaft .
tolerant society .

'In my home country countryX, many people live together in a single household, sometimes even a whole family, which makes for a helpful and tolerant society. '