Every Verb in its Right Place? A Roadmap for Operationalizing Developmental Stages in the Acquisition of L2 German

Josef Ruppenhofer♡, Matthias Schwendemann♠, Annette Portmann♠, Katrin Wisniewski♠, Torsten Zesch♡

♡ CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics
FernUniversität in Hagen
josef.ruppenhofer,torsten.zesch@fernuni-hagen.de
♠ Herder Institut, Universität Leipzig,
matthias.schwendemann,annette.portmann,katrin.wisniewski@uni-leipzig.de

Abstract

Developmental stages are a linguistic concept claiming that language learning, despite its large inter-individual variance, generally progresses in an ordered, step-like manner. At the core of research has been the acquisition of verb placement by learners, as conceptualized within Processability Theory (Pienemann, 1989). The computational implementation of a system detecting developmental stages is a prerequisite for an automated analysis of L2 language development. However, such an implementation faces two main challenges. The first is the lack of a fully fleshed out, coherent linguistic specification of the stages. The second concerns the translation of the linguistic specification into computational procedures that can extract clauses from learner-produced text and assign them to a developmental stage based on verb placement. Our contribution provides the necessary linguistic specification of the stages as well as detaiiled discussion and recommendations regarding computational implementation.

Keywords: Acquisition, Syntax, Profiling

1. Introduction

Second language acquisition (SLA) depends on many factors, such as learners' age or educational background, which cause considerable variation in learner language. While learners' *interlanguage* (IL) is highly variable, it is also known to be systematic (Selinker, 1972). Therefore, finding stable linguistic developmental features that are largely independent of such factors and of characteristics of the target language has been a key goal in researching learner language development. The concept of *developmental stages* (Hulstijn et al., 2015) is key in this regard. It refers to core grammar phenomena that are hypothesized to be acquired by learners in an ordered, step-like manner: learners may be faster or slower, but stages cannot be skipped or re-ordered (so-called *implicationality* (Rickford, 2002)). Even explicit instruction cannot enable this (Pienemann, 1989; Arntzen et al., 2019).

Early research on developmental stages focused on the morpheme acquisition order in English (e.g. (Dulay and Burt, 1973)), but subsequently, most attention has been paid to the acquisition of word order and, in particular, verb placement (Diehl et al. (2000); Gunnewiek (2000); Tschirner and Meerholz-Härle (2001); Pienemann (2005a); Jansen (2008); Jordens (1990); Baten and Håkansson (2015); *inter alia*). For L2 learners, knowing where to place the verb(s) is a crucial challenge. For example, in German main clauses, the canonical word order is subject - verb - object (SVO) as in: *Regina kauft Kuchen* (Regina buys cake). However, if something is placed before the subject, SVO is given up as the finite verb has to remain in the second position (V2), as in: *Heute kauft Regina Kuchen* (*Today buys Regina cake). In subordinate clauses, finally, the finite verb has to be placed at the end, as in: *Ich glaube, dass Regina Kuchen kauft* (*I believe that Regina cake buys). Developmental stage research assumes that these (and other) structures are acquired in a fixed sequence.

Research into the validity and variation-independent stability of developmental stages is important as a theoretical concern within SLA as the assumption of stages is all but uncontroversial. Competing approaches that conceptualize variation as a driver of L2 development (e.g. Complex Dynamic Systems Theory (Larsen-Freeman, 2020)) tend to discard the idea of developmental stages, casting doubt on the universality of stages once IL variation is fully taken into account. Even outside such frameworks, studies challenge the assumption of general developmental stages showing, for example, that order of acquisition might depend on the L1 (Murakami and Alexopoulou, 2016) or on the task (Ehl et al., 2018).

Research on developmental stages also intersects with learner corpus research as well as research on proficiency and language testing. In the case of German, the L2 we focus on, the stages play a considerable role in real-world contexts. A number of L2 assessments, such as Profilanalyse (Grießhaber, 2012; Heilmann and Griesshaber,

LREC-COLING 2024, pages 6655–6670

2012; Grießhaber, 2019) or LiSe-DaZ (Schulz and Tracy, 2011; Schulz and Grimm, 2012), and the Rapid Profile software (Keßler and Liebner, 2011) build on these stages. Decisions based on such test instruments can have direct consequences on test-takers and other stakeholders. For instance, placement decisions for newcomers to the German (or Austrian) school system with only limited knowledge of German may depend on such assessments as they transition from language preparation to regular classes (Wisniewski, 2023; Gamper, 2023). Furthermore, the stages are regularly used in teaching contexts, where they often guide the grammar curriculum (Winkler, 2014; Schroeder and Gamper, 2016; Baten, Kristof and Keßler, Jörg, 2019). Importantly, as pointed out by Wisniewski (2020), the relation between the grammar-focused developmental stages and the communicatively oriented CEFR (Common European Framework of Reference for Languages) levels requires further study, especially for languages other than English. Research into developmental stages so far has focused on data sets of limited size and heavily relied on manual annotation and classification based mostly on theoretical and example-based descriptions. In this contribution, we want to make progress towards large-scale automated analysis of multiple datasets. Accordingly, we first give an overview of work on developmental stages and specifically introduce the framework of Processability Theory (PT) (Pienemann, 1998; Pienemann et al., 2005) (§2). We use PT here as it is an important theory of developmental stages and has often been used to discuss German, our L2 of interest. We then discuss the challenges for deriving a technical specification from the theoretical constructs of developmental stage research (§3). Next, we discuss the challenges of implementing the specification computationally (§4) before offering a summary and outlook in §5.

2. Developmental Stages

Developmental stages for word order were first described for L2 German in the ZISA project (Clahsen et al., 1983). These ideas were then adopted and further developed mainly in Processability Theory (PT; (Pienemann, 1998, 2005b; Pienemann et al., 2022; Pienemann and Lenzing, 2020; Lenzing et al., 2019)). While stages are central in other approaches, too, for instance in the generatively oriented Organic Grammar (Vainikka and Young-Scholten, 2011), PT is the most wide-spread and broad theoretical approach that is based on both psycholinguistics (Levelt, 1989) and formal grammar (Lexical Functional Grammar, LFG (Kaplan and Bresnan, 1982; Dalrymple et al., 2019)).

2.1. Processability Theory

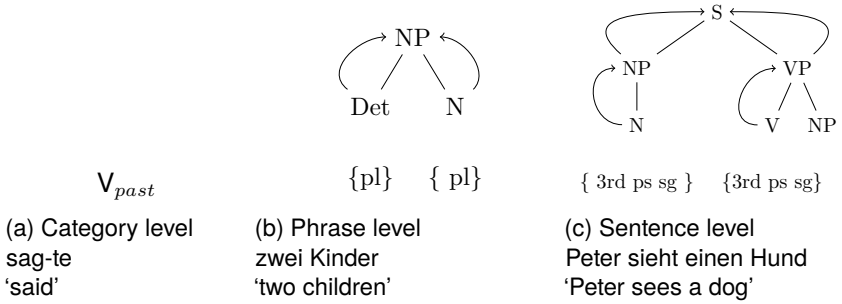
The core of PT is the *processability hierarchy*, shown in abbreviated form in Fig. 1. It encapsulates the idea that acquisitional order from simpler to more complex structures results from the fact that the capabilities of the human language processor (Levelt, 1989) expand in a specific sequence as it develops new processing procedures for handling ever more advanced grammar rules. At first, learners can only produce single words or chunks as they lack any processing procedures. In the first stage of acquisition, learners learn to associate lexical items with categorial information, enabling for instance inflection. Phrasal procedures (e.g. within NPs) require the exchange of information on features such as gender or number. The sentence procedure requires the exchange of information across phrases so as to ascertain subject-verb agreement.[1] Clearly, not all languages have gender, case, or subject-verb-agreement. Still, even if the linguistic features that are indicative of a learner's being at a specific stage (*core indicators*) are different across languages, the assumption is that the processing procedures underlying the stages are not. As all language use relies on processing procedures, so developmental stages are assumed to be relevant for all L2s. And, in fact, syntactic and/or morphological stages have been reported in a number of languages apart from English and German, namely Arabic, Chinese, Japanese, Swedish, Italian, and others (Pienemann, 2015; Bettoni and Biase, 2015; Ozeki and Shirai, 2007a,b).

2.2. Developmental stages for German

We now look at five of the PT stages for German: Svo, Adv, Sep, Inv, and V-End. Each of these is briefly illustrated and defined in Table 1. In what follows, we discuss the characteristics of the individual stages necessary for automatic recognition. Certain exceptions and special cases cannot be addressed here. For details on these, we refer the reader to (Pienemann, 1998; Bettoni and Biase, 2015). Instead, we focus on describing the respective core indicators and their operationalisation for the different stages as precisely as possible.

Svo (or Canonical Word Order) In a first step, learners gain access to the canonical word order of the target language. For German, PT takes this to be SVO. Learners are able to produce utterances with the subject in first position, followed by a verb in second position and a subsequent object, as in *Ich brauch eine neue Wohnung* [I need a new

[1] The information matching within and across phrases is usually seen as a mechanism mirroring LFG's *feature unification*.



(a) Category level
sag-te
'said'

(b) Phrase level
zwei Kinder
'two children'

(c) Sentence level
Peter sieht einen Hund
'Peter sees a dog'

Figure 1: Simplified PT hierarchy

Short Name	**Structure**	**Description**	**Example**
SVO	S→NP$_{subj}$ V (NP$_{obj1}$) (NP$_{obj2}$)	canonical word order	*Ich suche eine neue Wohnung.*
ADV	S→ (XP) NP$_{subj}$ V (NP$_{obj1}$) (NP$_{obj2}$)	adverb preposing	*Darum ich suche eine neue Wohnung.*
SEP	(XP) NP$_{subj}$ VP$_{Vaux/mod}$ {(NP$_{obj}$) / (X)} VP$_{Vinf}$	verb separation	*Darum ich muss eine neue Wohnung suchen.*
INV	(XP) V NP$_{subj}$ (NP$_{obj}$) / (X)	inversion	*Darum muss ich eine neue Wohnung suchen.*
V-END	(COMP)$_{ROOT}$ NP$_{subj}$ (NP$_{obj1}$) (NP$_{obj2}$) (ADJ) (V)$_{INF=-}$ (V)$_{INF=+}$	verb-end	*Weil ich eine neue Wohnung suche.*

Table 1: Developmental stages for verb placement in German

The structure descriptions are adapted from (Pienemann, 1998). Parentheses denote optionality. Subscripts represent feature annotations.

apartment]).[2] At the same time, this canonical word order seems to be more immediately accessible and comprehensible to learners since the various semantic roles correspond directly to the position in the underlying constituent structure (Jansen, 2008). Feature unification is not yet in place.

ADV At this stage, learners acquire the ability to place something before the canonical SVO structure, which is left unchanged. The sentence-initial saliency position (XP) can be filled with *Wh*-question words, adverbial complements, nominal or prepositional phrases. The resulting structures are ungrammatical in L1-German, as in: **Deshalb ich war immer krank.* ['Therefore I was always sick'] instead of the correct *Deshalb war ich immer krank*. Importantly, the salient sentence-initial position is filled with an additional structure that requires no feature unification. Therefore, it is assumed that this stage is acquired before the following stages, which all do require some feature unification.

SEP Next, learners acquire feature unification inside phrases. For German, verb separation is the key indicator for this stage. German builds verbal brackets with split verb phrases (VPs) where a finite (inflected) auxiliary or modal verb is separated from a non-finite (uninflected) verb (participle or infinitive), as in: *Und ich **soll** ein Schwimmbad **kaufen**.* [Lit. 'And I shall a swimming pool buy']. The finite verb has to be separated from the nonfinite verb by one or more objects and/or other element(s) such as adverbs. Canonical word order still persists, with S preceding the finite verb.

INV Later, learners acquire feature unification across phrasal boundaries. In German, the key indicator for this is taken to be subject-verb-inversion in sentences where the first position is not occupied by the subject, which is instead found after the finite verb. Compared to ADV, where two constituents precede the finite verb, learners now realize that the verb needs to come in second position (V2) and accordingly move the subject to the right of the finite verb as in: *Zuerst **habe ich** einige Fragen über die Sprache* ['First, I have some questions regarding the language'].INV involves the so-called sentence procedure and is the first point in the acquisitional hierarchy where learners produce structures not even partly featuring SVO.

[2]All examples in this section are taken from the MERLIN learner corpus (Wisniewski et al., 2013).They are presented in the form produced by the learners.

V-END At the highest stage in the processing hierarchy, learners differentiate main and subordinate clauses. In German, the key indicator for this is the final position of the finite verb in subordinate clauses as in: *wenn man sein Land verlässt*[Lit. 'when one one's country leaves']. As the arguments and adjuncts of German verb-final sentences can be variously ordered, V-END cannot simply be equated with OSV or SOV word order.

3. Challenge I: Specification

Having introduced processability theory and the five stages under consideration in this paper, we now discuss the challenges faced when trying to precisely define each stage. The specification is available at https://github.com/dakoda-project/devstages.

3.1. Lack of linguistic precision

Many formulae (rewrite rules) for structural description of PT's stages say nothing about optional material intervening between or co-occurring with the specified elements. For instance, in clauses with a preverbal subject and a postverbal object, adverbs may occur between verb and object as in (1).

(1) Er aß gestern Tofu.
He ate yesterday tofu.
'He ate tofu yesterday.'

Interpreting the SVO formula strictly, (1) would not match. But strict application is likely not intended.

A more critical case concerns the understanding of the named elements in the formulae. Consider the SEP stage where the finite verb in second position is separated from a non-finite verb or cluster of non-finite verbs at the end of the clause. In some prior research, certain types of intervening element such as simple negators like the adverb *nicht* ['not'] were not always accepted as evidence of separation (Jansen, 2008). Another similar issue concerns the precise understanding of what $O(BJ)$ means. While intuitively O covers objects, it is not clear which ones. German has three kinds of nominal objects: in addition to the prototypical accusative objects, there are dative and genitive objects. Further, the inventory of object types in LFG, PT's grammar formalism of reference, has changed over time and thus references to $O(BJ)$ in different PT papers may assume different types of objects. Additional cases, where the applicability of PT's formulae is unclear, are discussed in Appendix A.[3]

Conclusion In cases where abstract stage definitions were not precise, our own specification is based on surveying the prior literature to determine the most widely shared understanding.

3.2. Incomplete empirical coverage

ADV is presented as a structure used by learners that is not available in the L1. However, Müller (2003) has shown that L1 German does feature quite a few cases of so-called doubly filled prefields, that is cases, where two constituents precede the finite verb in second position. Many of them, including example (2), match the ADV-rule and thus look like IL productions of L2 speakers.

(2) [Vermutlich] [ein Defekt an der
Presumably a leak on the
Gashauptleitung] hat ... eine
gas main line has ... a
Gasexplosion ... verursacht .
gas explosion ... caused .
'Presumably a defect in the gas main line caused a gas explosion .. .' (=Müller 5b)

Other issues of coverage concern structures that are either rare in the written language or tied to specific genres/text types or modalities. For instance, the forms *weil* and *obwohl*, which in written L1-German are canonically used as subordinating conjunctions governing V-END, are often used in the spoken variety in ways similar to coordinating conjunctions in SVO clauses. Given PT's emphasis on the spoken form of the target language, such uses should ideally not be treated automatically as deviant L2 forms failing to exhibit V-END.

(3) Diese Reaktion finde ich gut , weil
This reaction find I good , because
das ist leider Realität.
that is unfortunately reality.
'I think this reaction is good, because unfortunately that is reality.'

More importantly, certain German clause types are out of scope altogether. Yes/No questions and imperatives (as typically verb-initial sentences) are outside of the purview of PT as are all non-finite clauses, even though both are subject to clear verb placement restrictions. Non-finite clauses, for instance, exhibit V-END (Wöllstein, 2010).

Conclusion In our specification, we point out cases where structures may be target-like or not depending on specific factors such as genre or modality so that either a manual or automatic post-analysis can be employed to resolve them. We also account for many additional verb placement constellations beyond the ones PT takes into view.

As we saw in the preceding discussion, the existing definitions of PT developmental stages are not precise enough to be technically implemented and they leave out of consideration many sentence

[3] For the submission phase, we leave in the references to any appendices we might add to a final version.

types that are found in L1 and also with advanced learners. To mitigate that fact, we will release a comprehensive collection of test sentences that tries to cover as many syntactic variants as possible. We refer the reader to appendix B.

3.3. Analyzing learner language

Interlanguage poses conceptual problems because it contains structures that are ungrammatical in L1 and whose analysis is ambiguous (Gass and Selinker, 2008; Ragheb and Dickinson, 2012). Consider (4), where we have two verbal forms *bleiben* 'remain' and *stehen* 'stehen', which in the context of the HaMaTac maptask corpus we interpret as parts of a complex verb *stehenbleiben* 'stop'.

(4) Wo bleiben ich stehen
where remain I stand
'Where do I stop ?' (HaMaTaC 008)

The issue is with the form *bleiben*: it can express (among others) the 1st or 3rd person plural present tense or the infinitive but not the 1st person singular present tense, which we expect. The specific error assumed here – use of plural instead of singular (i.e. incorrect number agreement) versus use of a non-finite form – has a big impact: in the former case, we have a finite clause and proceed with our verb placement analysis, whereas in the latter, we lack a finite clause and do not perform a verb placement analysis. Given the limited size of most learner texts, by excluding even just one instance we may lose key evidence that a learner has acquired a verb placement type. But especially for shorter L2 texts from cross-sectional studies we lack enough data to assess the state of a learner's grammar so any choice between error types would be entirely arbitrary. Prior research gives no indication how these cases were handled.

Conclusion The easiest approach would be to maximize compatibility with L1.[4] For (4), we would choose the number agreement error as the one that lets us proceed with the verb placement analysis. An alternative with the same result is what statistical parsers such as spaCy (Honnibal et al., 2020) do: in (4) they simply assign first person singular morphology to *bleiben* because that is most coherent with the available subject pronoun. We might interpret this as saying that, for the learner, *bleiben* is a first person form. But this is clearly a speculation about the learner's personal grammar that cannot be substantiated. A very strict alternative approach would be to exclude all verbal instances that are potentially non-finite. The strict approach likely removes a lot of learner data from verb placement analysis.

[4]This is similar to Dickinson and Ragheb (2013)'s idea of giving learners the benefit of the doubt.

The discussion above related to errors that are ungrammatical regardless of context. We refer the reader to Appendix C for discussion of context-bound ambiguities in learner language.

3.4. Chunks and borrowing of structures

Another challenge for the study of learner output consists in the use of multi-word chunks or formulaic sequences and the re-use of material given in task prompts. Some prior PT studies mention the exclusion of such material from the analysis of verb placement (e.g. Wisniewski (2020)) but do not detail the criteria for identifying it in the data.

Regarding formulaic sequences, one can use available inventories of chunks (or procedures for the data-driven discovery of chunks), multi-word expressions, and similar notions. However, in the end one cannot know what utterances a learner does produce compositionally and thus one has to simply assume that a formulaic L1 sequence also represents a chunk for an L2 learner. Similarly, while we can try to trace to what extent the lexical material and syntax in learner productions overlaps with that in the task prompt, it is not clear how such "borrowing" could be delineated from use of lexical material that is unavoidable given the topic, or from lexical and structural priming effects that native speakers and writers are also subject to.

Conclusion The handling of chunks and overlaps with task prompts is in need of further research.

3.5. Emergence criterion

For individual learners, PT employs a so-called emergence criterion: a stage counts as acquired if some N instances are produced in contexts where the relevant verb constellation is expected by L1 standards. 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. Note that high overall accuracy is not required for emergence (cf. Wisniewski (2020)).

Conclusion We recommend multiple analyses with different settings to see if results are stable.

4. Challenge II: Implementation

Having formally specified PT developmental stages and created a development and test sentence suite, we turn to issues of computational implementation.

4.1. Robust Preprocessing

Because the training data of our parsers comes from written L1 data, we have an imperfect fit to the L2 data that we want to analyze, which moreover contains spoken data as well. We lack high quality sentence-segmentation for spoken language and lower-level learners do not necessarily observe the punctuation conventions of German in their writing. Written learner data contains casing and other orthographic errors. Raw spoken language transcripts usually lack any casing and may employ some non-standard orthography to capture non-standard pronunciation, though several of our spoken learner corpora come with an orthographic normalization layer, including correct casing for intrinsically cased words (German nouns, for instance, are capitalized). Ideally, to improve parsing, our input would be sentence-split, true-cased and orthographically and lexically normalized.

Spoken language segmentation PT considers spoken L2 data to be the most relevant for analysis. However, a literature survey shows that no well-documented transcription and sentence segmentation standards are in common use in PT studies on spoken language. For German, two candidates are HIAT (Ehlich and Rehbein, 1979; Ehlich, 1993), which uses punctuation, and GAT 2 (Selting et al., 2009), which does not. Of the relatively few spoken German L2 corpora (Hirschmann and Schmidt, 2022), none seems to have fully applied an available standard. For instance, while the HaMaTaC corpus follows HIAT, it uses no punctuation (HZSK, 2010).

Thus, use or adaptation of segmentation approaches for L1 spoken German is unavoidable. However, prior work in this area is limited, most likely because much influential research on spoken language tends to find the sentence to be unsuitable as a basic unit of analysis (Deppermann and Proske, 2015). Still, there exists some recent work studying the manual (Westpfahl et al., 2018; Westpfahl and Gorisch, 2018) and automatic (Rehbein et al., 2020) segmentation of linguistic spoken language transcripts into sentence-equivalent units for L1 German. An alternative approach might be to first add punctuation (Tuggener and Aghaebrahimian, 2021), perform truecasing (Lita et al., 2003) and then perform segmentation, for instance using NLTK (Bird, 2006).

Case study: segmenting HaMaTaC data We explore the differences between the two available approaches using the spoken data on the HaMaTaC corpus. The punctuation-less one (punctless) uses only the transcribed tokens (Rehbein et al., 2020), while the other approach (punctful) performs punctuation insertion[5] (Guhr et al., 2021) and truecasing[6] before segmentation.

	punctless		punctful	
Sent.	5825		4378	
	spaCy	syntaxdot	spaCy	syntaxdot
V	2274	2295	2276	2291
V fin	1853	1889	1850	1886

Table 2: Comparison of verb analysis on HaMaTaC for two segmentation approaches and two different parsers (totals across all documents)

Table 2 shows that the two segmentation approaches differ significantly with respect to the number of sentences or sentence-like units: punctless (5825) has a third more sentences (4378) than punctful. The reason is that `punctless` produces a lot more one-word sentences for items like *ja* 'yes' , *genau* 'exactly' etc. whereas `punctful` includes them in larger segments. Importantly, we note that the number of verbs identified by our parsers, spaCy[7] and syntaxdot[8], is about the same across conditions . For both `punctless` and `punctful`, we have finite forms with about 80-82% of verbs, suggesting that recognition of this key category works about equally well in both settings. The syntaxdot parser recognizes slightly more finite verbs than spaCy. Detailed analysis (not included above) shows that it recognizes more imperative forms and more finite forms of auxiliaries than spaCy does.

To get a sense of the quality of the segmentations and verb analysis on HaMaTaC, we perform a manual analysis on the syntaxdot versions of a subset of transcripts selected for their varied lengths: with 6366 tokens, HaMaTaC documents 2, 6, 8,10, 18 and 25 make up 28% of the corpus total. In the `punctless` condition, the 6 documents feature 606 tokens tagged as verbal forms. Of these, 32 tokens have an at least partially incorrect analysis. In addition, 6 tokens tagged as non-verbs really are verbs. Among the forms incorrectly tagged as verbs are incomplete tokens, adverbs or interjections, notably the disfluency marker *äh* 'um'. In the other direction, isolated verbs in one-word utterances are mistaken for nouns. There are also morphological mis-analyses within the verbal category. For instance, imperatives and infinitives are mistaken for present tense forms and the other way around. These confusions between finite and non-finite forms are important as they negatively impact the verb placement analysis which should apply to

[5] https://github.com/oliverguhr/fullstop-deep-punctuation-prediction

[6] https://github.com/nreimers/truecaser

[7] https://spacy.io/

[8] https://github.com/tensordot/syntaxdot

all and only finite forms. Repeating the same analysis for the `punctful` setting, we find 44 tokens with the same kinds of errors as observed in the no-punct setting, with slightly more morphological mis-analyses.

Conclusion We recommend exploring both the punctuation-less and the punctuation-insertion approaches further for spoken language segmentation.

4.2. Topological + dependency parsing

While PT is tied to LFG, for German no broad-coverage LFG grammar is publically available. However, we can cast the task of identifying verb placement constellations as a combination of topological field and dependency analysis. For the topological aspect, consider Table 3. According to topological field theory, sentences consist of a sequence of slots (fields) which can contain certain elements (though not others) and which may not always be filled. For instance, in main clauses the finite verb (no matter if lexical, a tense auxiliary or a modal) is said to occupy the left sentence bracket (Linke Satzklammer (LSK)) (a-g), whereas any non-finite verb sits in the right sentence bracket (Rechte Satzklammer (RSK)) (d-g). In subordinate clauses, the finite verb is in the right bracket (h), potentially preceded there by non-finite forms (i). To differentiate among the non-VEND placement constellations, we need to consider the grammatical relations that the elements occupying the slots bear towards the verb. This information, which corresponds to LFG's f-structure, we can get from a dependency parser.[9] In SVO, for instance, the subject is in the Vorfeld ('pre-field', VF) and the object is in the Mittelfeld ('midfield', MF). In the case of INV, the subject is in MF and some other element occupies the VF slot. Note that topological field theory also applies to sentence and clause types that PT sets aside: (j) shows for instance, that in imperatives the verb occupies the left bracket, whereas VF is empty.

Available tools and data To our knowledge, no parsers trained specifically on L2 German are available. Prior work on parsing German learner data has used parsers trained on written German (Köhn et al., 2016; Rehbein et al., 2012; Ott and Ziai, 2010). Manual syntactic annotations are available for the Falko corpus (Reznicek et al., 2012), but they are applied to the target hypotheses (cf. 4.4) and thus cannot serve to train models for raw text. Treebank construction for learner language according to UD has recently seen growing interest, but the existing efforts are focused on written English as the L2 (Berzak et al., 2016; Lyashevskaya and Panteleeva, 2018; Morgado da Costa et al., 2022). Notable exceptions are a treebank for Chinese as the L2 (Li and Lee, 2018) and a treebank for spoken L2 English (Kyle et al., 2022). For L2 German data, however, no is treebank available.

Some efforts have been made to adapt tagsets for German POS annotation to spoken language (Rehbein and Schalowski, 2013; Westpfahl, 2014) but there exists little research on annotating spoken German syntax. Falenska et al. (2020) created the GRAIN-S corpus with manual syntactic annotations on six German radio interviews (626 sentences, 11274 tokens) for the purposes of language technology evaluation. The data was syntactically annotated according to the Tiger guidelines (Brants et al., 2004) with certain adaptions such as additional sentence node subtypes for parenthetical sentences and additional annotation conventions for spoken phenomena such as preposition stranding. While GRAIN-S represents a step away from written German, interviews do not well represent spontaneous spoken conversation. Also, the underlying transcripts are cleaned up for readability and do not contain unfinished words and lexical fillers. Other prior research on spoken language syntax exists within the Universal dependencies initiative. However, as Dobrovoljc (2022) points out, the spoken UD resources are currently still very heterogenous not only in how they apply the UD schema to their data but also likely with respect to transcription and segmentation. There also exist no common guidelines specific to spoken language syntax within UD.

Conclusion We can realistically only apply syntactic parsers for written L1 to written and spoken L2 material (cf. Köhn et al. (2016)). We suggest combining analyses from two different parsers, spaCy and syntaxdot, as shown in Fig. 2. While spaCy uses a German-specific inventory of dependencies from the Tiger schema (Brants et al., 2004) and is trained mostly on mainstream newspaper text, the syntaxdot parser uses the language-agnostic Universal dependencies scheme (Nivre et al., 2017) and is trained on the TüBa D/Z corpus (Telljohann et al., 2012) which notably contains data from the *taz* newspaper, which features a more colloquial and informal variety of German. Importantly, the syntaxdot parser also provides (flattened) topological field annotations.

4.3. Dependency on learner proficiency

For the study of developmental stages, early learners are most relevant as their grammar is less target language-like and evolving more. However, their productions are harder to analyze for L1-trained tools. Besides word order, all aspects of language from orthography, including punctuation, to mor-

[9] From a dependency parse, we could also infer constituents, similar to LFG's c-structure. However, the exact constituents are not important for our purposes.

		(VVF)	VF	LSK	MF	RSK	NF
a	SVO		Ich	suche	eine neue Wohnung		
b	SVO	Aber	ich	suche	eine neue Wohnung		
c	ADV		Darum ich	suche	eine neue Wohnung		
d	SEP		Ich	muss	eine neue Wohnung	suchen	
e	SEP	Und	ich	habe	eine neue Wohnung	gefunden	
f	INV		Darum	muss	ich eine neue Wohnung	suchen	
g	INV	Und	darum	muss	ich eine neue Wohnung	suchen	
h	V-END			weil	ich eine neue Wohnung	suche	
i	V-END			weil	ich eine neue Wohnung	gesucht habe	
j	V1-IMP			Suche	eine neue Wohnung		

Table 3: Verb placement and topological fields
Examples in part repeated from Table 1. VF =Vorfeld (prefield); LSK=Linke Satzklammar (left sentence bracket);
MF = Mittelfeld (midfield); RSK = Rechte Satzklammer (right sentence bracket); NF = Nachfeld (postfield); VVF = Vorvorfeld (pre-prefield)

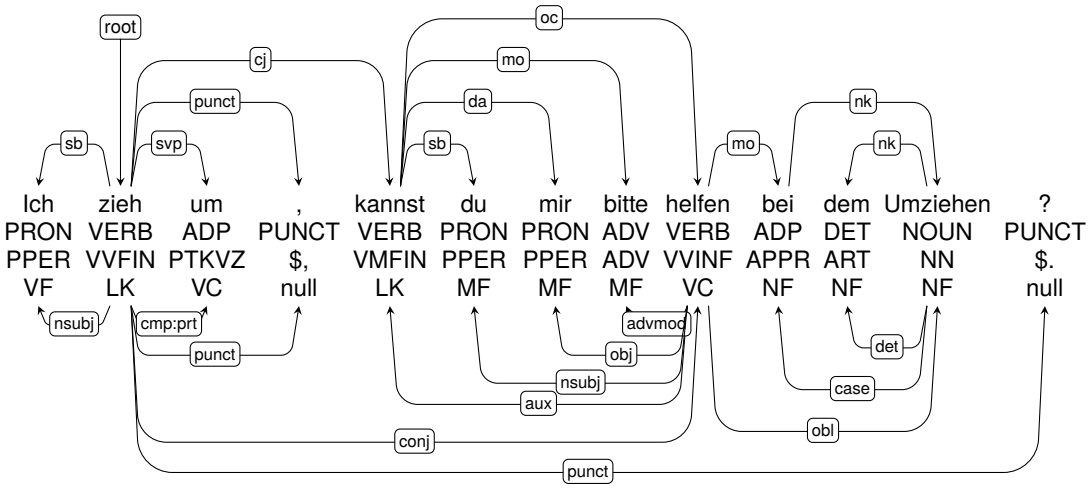


Figure 2: Token annotations: Forms, UD PoS, STTS PoS, Topological fields
Dependency arcs: spacy (Tiger scheme) top; syntaxdot (UD scheme) bottom

phology and syntax often diverge from L1 norms. Errors are not only more numerous, but also more severe. Fig. 3 illustrates this. The parser is robust against misspellings: though *umsiechen* is not a German word and *siech* occurs only as an adjective, both are treated as verbs, which is plausible in the context given the casing. But the lacking segmentation - a comma or period is missing after the third token - in combination with other problems - for instance, *Hilfe* 'help' should be changed to verbal *helfen* –leads to a poor parse where e.g. the token *umsiechen* is attached as a complement to the finite verb *zieh* rather than anywhere in the second clause. The combined errors prevent the recovery of verb placement.[10]

4.4. Use of Target Hypotheses

For some written L2 corpora, the creators provide so-called target hypotheses aligned to the learner productions. The target hypotheses (THs) explicate the assumed minimally different grammatical structures that the learners aimed for, and against which traceable error annotations can be made (Lüdeling, 2008; Reznicek et al., 2013). As noted in section §4.3, the sentence in Fig. 2 represents a target hypothesis for the sentence in Fig. 3. Given that THs are grammatical (if not always perfectly idiomatic), they are much easier to parse than the raw learner productions, especially at the lower CEFR levels. Given THs, one can parse both the learner output

[10] Identifying the specific contribution of an individual problem to an erroneous parse is not possible. However, studies for individual aspects have shown their impact on parsing. The notable influence of punctuation on parsing, for instance, has been confirmed repeatedly across different parsing approaches (Jones, 1994; Spitkovsky et al., 2011; Søgaard et al., 2018).

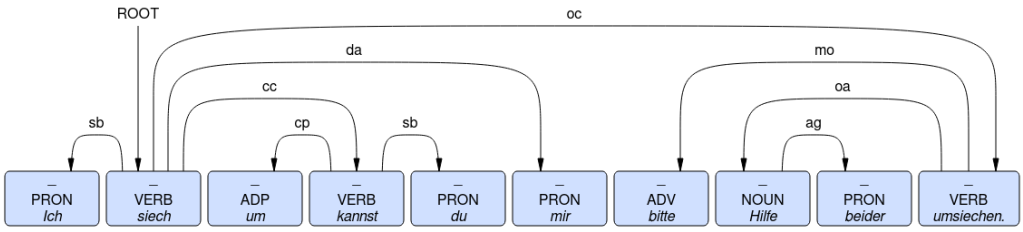


Figure 3: spaCy parse of a raw A1-Learner sentence from Merlin
(Lacking punctuation, orthographic and grammatical errors result in erroneous parse. Target hypothesis is given in Fig. 2)

and the TH in parallel and then project information from the TH back onto the learner output, correcting, for instance, erroneous POS or dependency annotations (cf. Rehbein et al. (2012)).

The challenge for this approach to analyzing verb placement lies in the inconsistent availability of THs across datasets. THs are not the same as what e.g. grammatical error correction or grammar checking systems provide. The outputs of these systems are usually more smoothed and idiomatic than what is required for our analysis, with more edit operations than are strictly necessary to create a minimally different but grammatically well-formed version.

Conclusion Research is needed to create a TH-generating system that does a more limited correction of learner productions.

4.5. Lack of grammatical information

Our parsing approach lacks information on valence or a(rgument)-structure. This prevents us from recognizing as correct certain spoken language structures that omit an argument (cf. 5a), where the same sequence of overt tokens would be judged ungrammatical if produced in the written modality (with a different interpretation) (cf. 5b).

(5) Kann ich lesen .
Can I read .

a. ✓ I can read that. (spoken, topic-drop)

b. ∗ I can read (stuff). (written, habitual)

To distinguish the above two cases, we would need to postprocess the parses with a semantic role labeler that tracks required arguments. Similarly, while subordinate clauses in German normally require VEND, there are cases of SVO and other structures in subordinate clauses that are acceptable in the spoken variety (see (3) in §3). To disambiguate acceptable from incorrect SVO, INV etc. in subordinate clauses, we would need to postprocess the output of the parser in the light of metadata on the modality from which the clauses originate.

Conclusion While we currently cannot disambigutate the above cases, we will inspect automatic analyses to ascertain if errors on these structures may compromise our overall analyses.

5. Summary

The notion of developmental stages focuses on features of language that are acquired in an unchanging order. Given that fixed stages are a controversial notion, especially in view of the observed variation in interlanguage and because of their real-world importance in language testing, we need to enable computationally backed large-scale studies. Here, we laid the groundwork for investigating the stages hypothesized by Processability Theory for the acquisition of verb placement in German. We addressed two challenges. First, we contributed a detailed linguistic specification of the stages. Second we discussed problems of computational implementation and provided recommendations for how to deal with them based on surveying the available data and tools for German.

6. Acknowledgments

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 16DKWN035A).

7. Ethics statement

We believe that, besides its relevance to scientific theory-building, research on developmental stages is necessary from an ethical point of view. Developmental stages form the basis for various test instruments and the results of such tests may significantly impact the school careers of newcomers to German-speaking countries, who are often migrants and / or refugees for whose integration into society a successful educational trajectory is key. Making sure that relevant tests are based on sound theory advances educational equity.

8. Bibliographical References

Ragnar Arntzen, Jörg-U Keßler, Arnstein Hjelde, and Gisela Håkansson, editors. 2019. *Teachability and learnability across languages*. John Benjamins Publishing Company.

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.

Baten, Kristof and Keßler, Jörg. 2019. Research timeline : the role of instruction : teachability and processability. In Arntzen, Ragnar and Håkansson, Gisela and Hjelde, Arnstein and Keßler, jörg, editor, *Teachability and learnability across languages*, volume 6 of *Processability Approaches to Language Acquisition Research & Teaching*, pages 9–26. John Benjamins.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal dependencies for learner english. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746. Association for Computational Linguistics.

Camilla Bettoni and Bruno Di Biase. 2015. *Grammatical development in second languages. Exploring the boundaries of Processability Theory*. Lulu.com.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Journal of Language and Computation*, 2:597–620.

Harald Clahsen, Jürgen M Meisel, and Manfred Pienemann. 1983. *Deutsch als Zweitsprache: Der Spracherwerb ausländischer Arbeiter*. Narr.

M. Dalrymple, J.J. Lowe, and L. Mycock. 2019. *The Oxford Reference Guide to Lexical Functional Grammar*. Oxford University Press.

Arnulf Deppermann and Nadine Proske. 2015. Grundeinheiten der Sprache und des Sprechens [Basic Units of Language and Speaking]. In Christa Dürscheid and Jan Georg Schneider, editors, *Handbuch Satz, Äußerung, Schema*, pages 17–47. De Gruyter, Berlin , München, Boston.

Markus Dickinson and Marwa Ragheb. 2013. Annotation for Learner English Guidelines, v. 0.1. Technical report, Indiana University, Bloomington, IN. June 9, 2013.

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.

Kaja Dobrovoljc. 2022. Spoken language treebanks in Universal Dependencies: an overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Dudenredaktion. 2009. *Die Grammatik - Unentbehrlich für richtiges Deutsch*, 8th edition, volume 4. Bibliographisches Institut GmbH, Mannheim.

Heidi C. Dulay and Marina K. Burt. 1973. Should we teach children syntax? *Language Learning*, 23(2):245–258.

Birgit Ehl, Michèle Paul, Gunnar Bruns, Elisabeth Fleischhauer, Miriam Vock, Anna Gronostaj, and Michael Grosche. 2018. Testgütekriterien der „profilanalyse nach grießhaber". evaluation eines verfahrens zur erfassung grammatischer fähigkeiten von ein- und mehrsprachigen grundschulkindern. *Zeitschrift für Erziehungswissenschaft*, 21(6):1261–1281.

K. Ehlich. 1993. Hiat - a transcription system for discourse data. In J.A. Edwards and M.D. Lampert, editors, *Talking Data: Transcription and Coding in Discourse Research*, pages 123–148. Psychology Press.

Konrad Ehlich and Jochen Rehbein. 1979. Erweiterte halbinterpretative Arbeitstranskriptionen (HIAT2): Intonation [Extended semi-interpretative working transcriptions (HIAT2): Intonation]. *Linguistische Berichte*, 59:51–75.

Agnieszka Falenska, Zoltán Czesznak, Kerstin Jung, Moritz Völkel, Wolfgang Seeker, and Jonas Kuhn. 2020. GRAIN-S: Manually annotated syntax for German interviews. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5169–5177, Marseille, France. European Language Resources Association.

Jana Gamper. 2023. Die Profilanalyse für Neuzugewanderte: Ein ambivalenter Blick aus der Schulpraxis [Profile Analysis for Recent Immigrants: An Ambivalent Look at the Practice in Schools]. *Deutsch als Fremdsprache*, pages 107–111.

S.M. Gass and L. Selinker. 2008. *Second Language Acquisition: An Introductory Course*. Routledge. 3rd edition.

Wilhelm Grießhaber. 2012. Die Profilanalyse. In Bernt Ahrenholz, editor, *Einblicke in die Zweitspracherwerbsforschung und ihre methodischen Verfahren.*, DaZ-Forschung, pages 173–194. De Gruyter.

Wilhelm Grießhaber. 2019. Profilanalysen. In Stefan Jeuk and Julia Settinieri, editors, *Sprachdiagnostik Deutsch als Zweitsprache*. De Gruyter Mouton.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans-Joachim Böhme. 2021. Fullstop: Multilingual deep models for punctuation prediction. In *Swiss Text Analytics Conference*.

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.

Hanna Hedeland and Thomas Schmidt. 2012. Technological and methodological challenges in creating, annotating and sharing a learner corpus of spoken German. In Thomas Schmidt and Kai Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, pages 25–46. John Benjamins.

B. Heilmann and W. Griesshaber. 2012. *Diagnostik & Förderung - leicht gemacht : [das Praxishandbuch]*. Deutsch als Zweitsprache in der Grundschule. Klett Sprachen.

Gerhard Helbig and Joachim Buscha. 2013. *Deutsche Grammatik [German Grammar]*. Klett-Langenscheidt, München.

Hagen Hirschmann and Thomas Schmidt. 2022. Gesprochene lernerkorpora: Methodisch-technische aspekte der erhebung, erschließung und nutzung. *Zeitschrift für germanistische Linguistik*, 50(1):36–81.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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.

Bernard E. M. Jones. 1994. Exploring the role of punctuation in parsing natural text. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*, Kyoto, Japan.

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

Ronald Kaplan and Joan Bresnan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press.

Jörg-U Keßler and Mathias Liebner. 2011. Diagnosing L2 development. In Manfred Pienemann and Jörg-U. Keßler, editors, *Studying Processability Theory: An Introductory Textbook*, pages 131–147. John Benjamins.

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*, pages 144–153, Duisburg-Essen, Germany. German Society for Computational Linguistics and Language Technology.

Christine Köhn, Tobias Staron, and Arne Köhn. 2016. Parsing free-form language learner data: Current state and error analysis. In *Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, Bochum, Germany, September 19-21, 2016*, volume 16 of *Bochumer Linguistische Arbeitsberichte*.

Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. A dependency treebank of spoken second language English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45, Seattle, Washington. Association for Computational Linguistics.

Diane Larsen-Freeman. 2020. Complex Dynamic Systems Theory. In Bill VanPatten, Gregory D. Keating, and Stefanie Wulff, editors, *Theories in*

Second Language Acquistion, pages 248–270. Taylor & Francis.

Anke Lenzing, Howard Nicholas, and Jana Roos, editors. 2019. *Widening contexts for Processability Theory: Theories and issues*, volume 7. John Benjamins Publishing Company.

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

Keying Li and John Lee. 2018. L1-L2 parallel treebank of learner Chinese: Overused and underused syntactic structures. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Lucian Vlad Lita, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla. 2003. tRuEcasIng. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 152–159, Sapporo, Japan. Association for Computational Linguistics.

Olga Lyashevskaya and Irina Panteleeva. 2018. REALEC learner treebank: annotation principles and evaluation of automatic parsing. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories, TLT 2018, Prague, Czech Republic, January 23-24, 2018*, pages 80–87.

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.

Luís Morgado da Costa, Francis Bond, and Roger V. P. Winder. 2022. The tembusu treebank: An English learner treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4817–4826, Marseille, France. European Language Resources Association.

Akira Murakami and Dora Alexopoulou. 2016. L1 influence on the acquisition order of english grammatical morphemes. *Studies in Second Language Acquisition*, 38(3):365–401.

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

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics.

Niels Ott and Ramon Ziai. 2010. Evaluating dependency parsing performance on German learner language. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, NEALT Proceeding Series.

H. Ozeki and Y. Shirai. 2007a. The consequences of variation in the acquisition of relative clauses: An analysis of longitudinal production data from five japanese children. In O. W. Robinson Y. Matsumoto, D. Y. Oshima and P. Sells, editors, *Diversity in language: Perspectives and implications*, pages 243–270. CSLI Publications.

Hiromi Ozeki and Yasuhiro Shirai. 2007b. Does the Noun Phrase Accessibility Hierarchy predict the difficulty order in the acquisition of Japanese relative clauses? *Studies in Second Language Acquisition*, 29:169–196.

Manfred Pienemann. 1989. Is Language Teachable? Psycholinguistic Experiments and Hypotheses. *Applied Linguistics*, 10(1):52–79.

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

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

Manfred Pienemann, editor. 2005b. *Cross-linguistic aspects of processability theory*. Benjamins.

Manfred Pienemann. 2015. An Outline of Processability Theory and Its Relationship to Other Approaches to SLA : An Outline of Processability Theory. *Language Learning*, 65(1):123–151.

Manfred Pienemann, Bruno Di Biase, Satomi Kawaguchi, and Gisela Håkansson. 2005. Processing constraints on L1 transfer. In J.F. Kroll and A.M.B. de Groot, editors, *Handbook of Bilingualism : Psycholinguistic Approaches*, pages 128–153. Oxford University Press.

Manfred Pienemann, Frank Lanze, Howard Nicholas, and Anke Lenzing. 2022. Stabilization. A dynamic account. In Alessandro G. Benati and John W. Schwieter, editors, *Second Language Acquisition Theory: The legacy of Professor Michael H. Long*, pages 29–76.

Manfred Pienemann and Anke Lenzing. 2020. Processability Theory. In Bill VanPatten, Gregory D.

Keating, and Stefanie Wulff, editors, *Theories in Second Language Acquisition*, chapter 8, pages 162–191. Routledge.

Marwa Ragheb and Markus Dickinson. 2012. Defining Syntax for Learner Language Annotation. In *Proceedings of COLING 2012: Posters*, pages 965–974, Mumbai, India.

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

Ines Rehbein, Josef Ruppenhofer, and Thomas Schmidt. 2020. Improving Sentence Boundary Detection for Spoken Language Transcripts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7102–7111. European Language Resources Association.

Ines Rehbein and Sören Schalowski. 2013. STTS goes Kiez - Experiments on Annotating and Tagging Urban Youth Language. *Journal for Language Technolology and Compututational Linguistics*, 28(1):199–227.

Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing target hypotheses in the Falko corpus. In Ana Díaz-Negrillo, Paul Thompson, and Nicolas Ballier, editors, *Automatic treatment and analysis of learner corpus data*, pages 101–123. John Benjamins Publishing Company.

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.

John Rickford. 2002. Implicational Scales. In Peter Trudgill James K. Chambers and Natalie Schilling-Estes, editors, *The Handbook of Language Variation and Change*, pages 142–167. Blackwell.

Christoph Schroeder and Jana Gamper. 2016. Sprachliche Bildung für Neuzugewanderte. Ein Plädoyer für einen erwerbssequentiellen Ansatz [Language instruction for new immigrants: A plea for an approach based on order of acquisition]. *Osnabrücker Beiträge zur Sprachtheorie*, 89:217–230.

P. Schulz and A. Grimm. 2012. Spracherwerb. In H. Drügh, S. Komfort-Hein, A. Kraß, C. Meier, G. Rohowski, R. Seidel, and Helmut Weiß, editors, *Germanistik. Sprachwissenschaft – Literaturwissenschaft – Schlüsselkompetenzen*, pages 155–172. Metzler, Stuttgart.

Petra Schulz and Rosemarie Tracy. 2011. *Linguistische Sprachstandserhebung - Deutsch als Zweitsprache (LiSe-DaZ): Language Test for Children with German as a Second Language*. Hogrefe, Göttingen.

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

Margret Selting, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Martin Günthner, Susanneand Hartung, and Friederike Kern. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2) [Conversation analytic transcription system 2]. *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, 10:353–402.

Anders Søgaard, Miryam de Lhoneux, and Isabelle Augenstein. 2018. Nightmare at test time: How punctuation prevents parsers from generalizing. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 25–29, Brussels, Belgium. Association for Computational Linguistics.

Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2011. Punctuation: Making a point in unsupervised dependency parsing. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 19–28, Portland, Oregon, USA. Association for Computational Linguistics.

Philipp Stoeckle. 2005. Doppelte Infinitivkonstruktionen im Alemannischen – Eine empirische Untersuchung anhand der spontansprachlichen Daten des SSA [Double infinitive constructions in Alemmanic - An empirical investigation based on the spontenous speech data of the SSA]. Master's thesis, Albert-Ludwigs-Universität, Freiburg im Breisgau. Wissenschaftliche Arbeit zur Zulassung zur Wissenschaftlichen Prüfung für das Lehramt an Gymnasien.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2012. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft.

E. Tschirner and B Meerholz-Härle. 2001. "Processability Theory: Eine empirische Untersuchung" [Processability Theory: An empiriical 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.

Don Tuggener and Ahmad Aghaebrahimian. 2021. The Sentence End and Punctuation Prediction in NLG Text (SEPP-NLG) Shared Task 2021. In *Swiss Text Analytics Conference*.

Anne Vainikka and Martha Young-Scholten. 2011. *The Acquisition of German*. De Gruyter Mouton.

Swantje Westpfahl. 2014. STTS 2.0? improving the tagset for the part-of-speech-tagging of German spoken data. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 1–10. Association for Computational Linguistics and Dublin City University.

Swantje Westpfahl and Jan Gorisch. 2018. A Syntax-Based Scheme for the Annotation and Segmentation of German Spoken Language Interactions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 109–120. Association for Computational Linguistics.

Swantje Westpfahl, Nadine Proske, Melanie Hobich, Anton Borlinghaus, and Hanna Strub. 2018. Guideline: Syntactic annotation and segmentation in the SegCor Project. Working paper, Leibniz-Institut für Deutsche Sprache.

Steffi Winkler. 2014. Für eine psycholinguistisch orientierte Fremdsprachendidaktik. Anmerkungen auf Grundlage einer exemplarischen Studie zum Negationserwerb im Deutschen [In favor of a pyscholinguistically oriented foreign language didactics. Remarks based on a sample study on the acqusistion of negation in German]. In U. Bredel, I. Ezhova-Heer, and S. Schlickau, editors, *Zur Sprache.kɔm*, page 51. Universitätsverlag.

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. 2023. Zum Einsatz der Profilanalyse für neu zugewanderte Menschen in Bildungskontexten: Ein Forschungsaufruf [On the Use of Profile Analysis for Recent Immigrants in Educational Contexts: A Call for Research]. *Deutsch als Fremdsprache*, pages 42–45.

Katrin Wisniewski, Karin Schöne, Lionel Nicolas, Chiara Vettori, Adriane Boyd, Detmar Meurers, Andrea Abel, and Jirka Hana. 2013. MERLIN: An online trilingual learner corpus empirically grounding the European Reference Levels in authentic learner data. In *ICT for Language Learning Conference (ICT4LL 2013)*.

A. Wöllstein. 2010. *Topologisches Satzmodell*. Kurze Einführungen in die germanistische Linguistik. Winter.

Gisela Zifonun, Ludger Hoffmann, and Bruno Strecker. 1997. *Grammatik der deutschen Sprache [Grammar of the German Language]*. De Gruyter, Berlin, New York.

9. Language Resource References

HZSK. 2010. *HAMATAC - the Hamburg MapTask Corpus*. Archived in Hamburger Zentrum für Sprachkorpora. PID http://hdl.handle.net/11022/0000-0000-6330-A. Version 0.3.

Appendix A. Additional challenges for theoretical specification

Further cases of lacking empirical coverage

While the INV stage is defined in relation to declaratives and wh-questions, the PT literature makes no mention of other structures that may occasionally feature a preposed constituent before a verb in second position. For instance, in (6), we have a yes/no question (asked to seek reassurance from the interlocutor) with a fronted topic NP, while (7) features an imperative with a fronted object NP.

(6) Das Buch kauft Hannelore heute?
The book buys Hannelore today?
'The book Hannelore buys today?'

(7) Den Namen sag mir an!
The name tell me
'Tell me the name!'

While both Yes/No-questions and imperatives are usually verb first, the question arises what should in principle be done about minor cases such as (6) and (7): is there a good reason to ignore them when investigating developmental stages of verb placement or is it an oversight that they have not been discussed as cases of INV?

Similarly, clause-final placement of finite verbs (VEND) is said to be conditioned by the value "-" of the ROOT-feature on complementizers (COMP) in subordinate clauses. One problem with this definition of VEND is that German also allows quite a few perfectly grammatical sentence-types that look structurally like subordinate clauses but are used by themselves without a governing predicate. For instance, (8) is a free-standing degree exclamative that formally has the structure of an embedded question, while (9) is a free-standing question (representing interior monologue) with the structure of an embedded polar question.

(8) Was das gekostet hat!
What that cost has!
'How much that cost!'

(9) Ob er es weiß?
Whether he it knows?
'If he knows it?'

On a strict interpretation of the VEND rule, sentences (8) and (9) would not be matches for the VEND rule if *was* and *ob* are treated as COMP with the feature ROOT=+. However, that would raise the question why there are cases where such clauses are governed by embedding predicates. Potentially, this could be solved by assuming polysemy of *was* and *ob* so that there exist parallel entries with the feature ROOT=-. However, that is an inelegant solution as the issue is fully general for all German question words and complementizers.

The VEND constellation is also a relevant reference point for presentational relative clauses typically found in spoken German narration such as 10. In these constructions, VEND fails to appear in the relative clause but instead SVO, INV or other orderings are found. PT's discussion of verb placement has nothing to say about these structures.

(10) Ich hatte mal einen Freund, der
I had once a friend who
hieß Sascha.
was-called Sascha.
'I once had a friend who was called Sascha.'

Regionally or socially defined varieties

A special subset of issues with empirical coverage concerns lacking awareness of non-standard varieties. Different varieties of a language admit or rule out different grammatical structures. This also applies to German verb placement. Example 11 from Alemannic German shows a constellation where the right sentence bracket does not contain all the expected verbal material or does not contain it in the right order, with an object intervening between two non-finite verb forms.

(11) dann **hat** man **können** einen traktor
then has one can a tractor
kaufen
buy
'then one was able to buy a tractor' (= (4), Stoeckle (2005))

Learners acquiring German in areas where such varieties are spoken could potentially pick up such forms. Given that PT is focused on spoken language, the use of such structures would then ideally be treated as a case of a regional variety (which may be inappropriate in some contexts such as the classroom) rather than as a garden variety error.

Appendix B. Comprehensive collection of test sentences

In order to evaluate any automatic approach for recognizing verb placement constellations, we need a test set. We constructed such a set where the 415 finite clauses of 249 distinct sentences are annotated for the verb placement constellation found in them. This material represents L1 German as it is meant to test in principle how well verb constellations can be recognized in the absence of orthographic, grammar and other errors found in L2 productions.

The verb constellations we take into account include not only the five major categories of Processability Theory (SVO, ADV, SEP, INV, VEND) but also additional ones that PT does not discuss and which may be used only by more advanced learners. For instance, we include verb-first subordinate conditional sentences. The reason for the larger inventory is to evaluate the full range of German sentence types. Being able to study the full range is useful in analyzing the language of proficient learners (C1, C2) and also for comparing learner language to native language text.

Note, too, that each constellation is represented multiple times, with different verbs and/or at different positions within the overall sentence. For instance, we want to make sure that VEND in subordinate clauses is recognized not only at the end of the overall sentence but also when the subordinate clause starts the overall sentence.

The inventory in the test set was put together based on discussion of German sentence types in Wöllstein (2010) as well as in various German grammars (Helbig and Buscha, 2013; Zifonun et al., 1997; Dudenredaktion, 2009).

Appendix C. Ambiguity of learner language

Apart from the issue of erroneous forms which may represent multiple distinct types of errors, we also have the issue that a (learner) sentence may be grammatical dependent on the context or the specific interpretation assumed. Example (12) from (Köhn and Menzel, 2015) illustrates this: if we take the grammatical form as is, the sentence must refer to the son sending his mother someplace. However, in the context of a given learner task such as the retelling of a picture story we may know that the opposite scenario is the one to be expressed, in which case the produced form is an error and the nominative case article *der* is incorrectly used where the accusative form *den* is required. Note that the interpretation of the case forms seen in example (12) also impacts the verb placement constellation that is being observed: we either have SVO or we have inversion, with OVS.

(12) Die$_{nom/acc}$ Mutter$_{nom/acc}$ schickt der$_{nom}$ the mother sends the Sohn$_{nom}$. son
'The son sends the mother' or 'The mother sends the son'

A manual approach to analyzing learner language can detect such cases but statistical parsers maximize grammaticality and lack awareness of context / ground truth interpretation and therefore would choose the context-inappropriate reading that's solely based on the forms. We can identify such potential errors through morphological analysis that points out the ambiguities of individual forms and assess through manual analysis of random samples if we have incurred errors of this kind.