Do large language models and humans follow similar learning stages? – Assessing GPT-2's order of Swedish grammar acquisition within the Processability Theory framework

Stella Lundqvist

Uppsala University
Department of Linguistics and Philology
Master Programme in Language Technology
Master's Thesis in Language Technology, 30 ECTS credits

June 23, 2025

Supervisors:
Johan Sjons, Uppsala University
Murathan Kurfalı, RISE Research Institutes of Sweden

Abstract

Human second language (L2) learners have been found to learn grammatical structures in a predictive, implicational order, hypothesized by Processability Theory (PT) to be restricted by the processing procedures that are accessible at the learner's stage of language development. This study investigates whether large language models (LLMs) acquire Swedish grammatical structures in the same implicational order, and whether this order is robust to the sequencing of the input data. We model Swedish L2 learners by fine-tuning four GPT-2 models pretrained on English on Swedish unlabeled data with different input orders in regards to the grammatical structures identified in the data. We present SwePT – a Swedish minimal pair Processability Theory dataset containing nine separate sets of minimal pairs targeting Swedish syntactic and morphological structures that are acquired by human L2 learners on four separate stages of language development – and evaluate the GPT-2 models on SwePT through an acceptability classification task at each 100th time step during fine-tuning. We find that the observed acquisition orders correlate across the fine-tuned models, while violating the implicational order sequence as hypothesized by PT. The observed relation between performance on the classification task and frequency distributions of the contrasting features in the minimal pairs suggests that the acquisition order can be explained by unigram and n-gram heuristics. While the adaptation of NLP methodologies into the PT framework requires further conceptual and methodological refinement, our findings suggest that the grammatical development predicted by PT does not naturally emerge from next-word prediction objectives.

Contents

Preface

This thesis would not have been possible without the continuous support and encouragement from my supervisors Johan Sjons and Murathan Kurfalı. I am very grateful to you both. I am also grateful to Joakim Nivre for reviewing the logic of my parsing scripts and lending me invaluable insights on Swedish grammar, as well as to Anna Flyman Mattsson at Lund University for introducing me to PT five years ago, and for answering my persistent emails about PT. I would also like to thank the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) for providing access to the computational resources that enabled this project.

Special thanks go to the Lofi Girl YouTube channel and Jeremy Sole's Skyrim soundtrack, without which I would not have found the concentration to write a single sentence of this paper. Lastly, I want to thank Irini Chamali for her unwavering emotional support.

1. Introduction

Human language has probably existed for hundreds of thousands of years, and its study can be traced back as far as to the 6th century BCE. Yet, many questions related to how humans process and acquire language remain. How do we process language? Is syntactic knowledge hardwired in our brains, or is the input and interaction we receive from our environment sufficient? And is syntactic knowledge required at all to learn a language?

In recent years, the arrival of artificial neural networks (ANNs) and large language models (LLMs) in the likes of ChatGPT have given rise to even further questions about the nature of language development and the similarities between human and artificial language processing and acquisition. Such questions have brought about research studies where concepts from human language acquisition theory are applied in the study of artificial cognition, and vice versa. The possibilities to model language acquisition by controlling input and observing the calculations of the layers of neurons in artificial "brains" may bring us closer to understanding which language-related features are required to carry out certain linguistic tasks, and may teach us not only about artificial language acquisition, but also about human language acquisition in the process.

One aspect of human language acquisition that has been extensively studied is the concept of developmental stages, in particular in regards to children's acquisition of their first language (e.g. Brown, 1973; Kuhl, 2004; Tomasello, 2005; Yang, 2016). Such research has in turn inspired studies on whether artificial learners display similar patterns (Choshen et al., 2021; Evanson et al., 2023; McCoy et al., 2018; Warstadt et al., 2023; Yedetore et al., 2023).

Although first language acquisition is known to differ from second language acquisition (L2A) in many aspects, the latter remain underresearched. The Processability Theory (PT; Pienemann, 1998b, 2005) remains one of the most influential theories on L2 grammatical development, and in the L2A research field overall. However, despite substantial cross-linguistic support from empirical studies on human speech production (e.g. Kawaguchi, 2008; Mansouri, 2008; Norrby and Håkansson, 2007; Wang, 2011) and a handful of studies testing grammatical perception within the same framework (Buyl and Housen, 2015; R. Ellis, 2008; Keatinge and Keßler, 2009; Spinner, 2013), no study has, to the best of our knowledge, investigated the hypotheses of PT on artificial learners.

In the present study, we aim to explore this research gap by fine-tuning a Generative Pre-trained Transformer 2 (GPT-2) model pretrained on English on Swedish, thereby modeling a Swedish L2 learner, and testing whether it exhibits a developmental trajectory in acquiring Swedish grammatical structures that aligns with the trajectory that has been observed within human L2 learners. We will also examine whether this trajectory is affected by the sequencing of the input during training.

Our research questions are as follows:

1. To what extent does GPT-2's acquisition order of Swedish grammatical structures follow the order sequence as stipulated by PT?
2. To what extent does input order of the training data (curriculum learning) affect the learning trajectory?

We will attempt to find answers to these questions by evaluating GPT-2 models through a curriculum learning approach, where the order of the training data varies between increasing complexity and decreasing complexity (as determined by the PT framework) of the grammatical structures present in the input. We will test the models' grammatical knowledge development using acceptability judgment tests (AJT) on a curated minimal pair dataset at regular intervals throughout training, where the model's ability to assign higher probability scores to the grammatical sentences compared to their ungrammatical equivalents with regards to the target morphosyntactic structures is measured. We will evaluate the AJT using adapted implementations of the emergence criterion and implicational scaling, that are traditionally used to test human learner output within the PT framework, as well as analyzing the learning trajectories.

Our main contributions are: (i) introducing a novel approach and bridging the fields of L2A and AI research in using the PT framework to the evaluation of LLMs; (ii) creating and publicly releasing the *Swedish Processability Theory Minimal Pair Dataset* (SwePT) including canonical word order SVO, plural, tense, attributive agreement, predicative agreement (with and without attractor nouns), inversion after topicalization, preverbal negation and non-inversion in indirect questions – along with our codebase; and (iii) evaluating GPT-2 on its performance and acquisition order of the structures in SwePT using acceptability judgment tests.

By evaluating whether human developmental patterns are shared by artificial learners such as GPT-2, the present project may contribute to the emerging research field of learning trajectories, to the expansion of the PT framework, and to our understanding of universal principles of language acquisition and how they differ between humans and artificial learners. Additionally, testing curriculum learning effects on grammatical development in artificial learners may contribute to the development of more efficient methods for training LLMs with input sequencing.

2. Background

2.1. Grammaticality and acceptability judgments

One of the most widely used methods of testing grammatical competence is acceptability judgment tests (AJT). Although initially applied within theoretical linguistics to test theories of grammar, they have since been widely used within L1 and L2 acquisition research, and more recently in NLP to test models.

At the core of AJT is the classification task, where the participant or model (henceforth: learner) evaluates whether a sentence is grammatically acceptable or not (sometimes under a specific reading; see e.g., Adger, 2003, p. 5). Aside from the classification task at the core of AJT, the implementation thereof follows no standard practice. A common approach is to use *minimal pairs*, where the learner is presented with one grammatical and one ungrammatical sentence that differ from each other on a single linguistic aspect (e.g., the position of a word in the sentence), and is tasked to determine which one of them is grammatical. The classification task may rely either on the Boolean paradigm (where a sentence assumed to be either grammatical or ungrammatical, e.g., Warstadt, 2019) or the gradient paradigm (where the sentence is assigned a score on a scale, e.g., Lau et al., 2017). The AJT implementation in the present study follows the boolean paradigm, where likelihood scores assigned to either the grammatical or the ungrammatical sequence of a minimal pair indicates which of the two sentences is more acceptable to the model.

It should be noted that although the terms "acceptability judgments" and "grammaticality judgments" are often used interchangeably in the field, the concepts of grammaticality and acceptability are not equivalent. Ponder the sentence "The woman the boy who I love loves loves the dog", which is grammatical but not necessarily acceptable. Similarly, the sentence "Last year more people visited Rome than I did" is ungrammatical, but arguably acceptable. While grammaticality is a theoretical concept referring to whether a structure conforms to the formal rules of grammar, acceptability is an empirical measure based on speaker judgments of whether a sentence sits right with their linguistic intuition or not. Moreover, Lau et al. (2017) refer to grammaticality as "the theoretical competence that underlies the performance phenomenon of speaker acceptability judgements". As such, linguistic competence can only be *inferred* from acceptability judgments, since factors beyond grammaticality, such as semantic plausibility and individual processing issues, also play into the grammaticality classification task.

2.1.1. Acquiring linguistic knowledge

There is no unified view among scholars on whether the linguistic knowledge used for the AJT classification task is categorical or probabilistic in nature – the former calling on an internal grammar that generates a set of structures (e.g., Chomsky, 1957), and the latter relying on a gradient probability distribution across some set of constituents in the language (e.g., Chater and Manning, 2006). However, it is well-supported that human language acquisition is guided by, and our linguistic knowledge based on, our *linguistic experience*. N. C. Ellis (2002, p. 162) writes that our linguistic knowledge can be understood "as a statistical ensemble of language experiences that changes slightly every time a new utterance is processed." This claim is supported by various observations. High frequency words are known to be recognized by humans more quickly than low frequency words (Grosjean, 1980), and syntactically ambiguous words (i.e. words that may have more than one part of speech in a certain linguistic context) are initially interpreted with their most likely part-of-speech (Crocker and Corley, 2008). Human speakers have been found to rate a sentence as more grammatical if they have read the same sentence or a sentence with the same syntactic structure earlier (Luka and Barsalou, 2005). And recent exposure of a syntactic structure can also make that structure easier to comprehend (e.g., Fine et al., 2013) – a concept known as *syntactic priming*.

Still, probability alone cannot predict acceptability. Lau et al. (2017) use the sentences "I saw a cat" and "I saw a yak" to exemplify the fact that while the former is a much more probable sequence of words, both sentences are equally as acceptable. Moreover, aside from lexical frequency, sequence length also plays into the probability score of a sequence, while not necessarily affecting its acceptability.

Lau et al. (2017) took these confounding factors into account when examining the relationship between acceptability and probability. They found a surprisingly strong correlation between probability values and crowd-sourced gradient acceptability judgment ratings, after normalizing probabilities through a function that discounts the non-predictive factors of lexical frequency and sequence length. They concluded that linguistic competence may be comprised, in part, by a probabilistic classifier. It is upon this assumption that AJT are used to test the development of grammar acquisition in GPT-2 in the current study.

2.1.2. Minimal pair datasets for AJT

One of the earliest and most significant contributions to applications of AJT is the work by Warstadt (2019), who introduced the Corpus of Linguistic Acceptability (CoLA). CoLA contains grammatical and ungrammatical sentences covering a broad range of theoretical syntax and was collected from linguistic literature and annotated by linguists. Shortly after CoLA, The Benchmark of Linguistic Minimal Pairs for English (BLiMP; Warstadt et al., 2020) was released, containing an even distribution of 12 linguistic categories constructed systematically using synthetic examples designed to isolate specific linguistic phenomena. A notable contribution in the context of our study is the Russian Benchmark of Linguistic Minimal Pairs (RuBLiMP; Taktasheva et al., 2024). RuBLiMP was constructed using an automated approach where sentences extracted from publicly available corpora were parsed, annotated and then searched to identify target linguistic structures. The minimal pairs were then curated through rule-based alteration processing – an approach that has been adopted in the present study.

The most significant contribution to Swedish AJT are the Dataset for Linguistic Acceptability Judgments (DaLAJ and DaLAJ-GED; Volodina et al., 2023, 2021), that were constructed from the error-annotated learner corpus SweLL (Volodina et al., 2019) which contains original erroneous sentences and their corrections.

In the present study, we present a Swedish minimal pair dataset targeting specific linguistic phenomena that have been shown to be acquired by human L2 learners in a specific developmental sequence. The theory behind this observed developmental sequence is accounted for in the next section.

2.2. Processability Theory

2.2.1. The theoretical foundation of PT

Processability Theory (PT) was first introduced by Pienemann (1998b). This psycholinguistic theory stipulates that a learner can acquire only those linguistic forms and functions that they can process, restricted by the processing procedures that are accessible at the learner's stage of language development. The procedures are accessed in a predictive order sequence, which has been supported through observing the acquisition order of linguistic structures in numerous cross-linguistic empirical studies (e.g. Kawaguchi, 2008; Mansouri, 2008; Norrby and Håkansson, 2007; Wang, 2011).

Derived from PT is the Teachability Hypothesis, which proposes that teaching should focus on grammar from the learner's subsequent developmental stage, since learners cannot "skip" developmental stages regardless of the content of formal instruction. While PT is known as an L2A theory, its predicted acquisition trajectory has been shown to generalize to first language acquisition as well, albeit with an observed faster progression between stages and even simultaneous acquisition of structures due to different initial hypotheses (Pienemann, 1998b, 2005).

PT builds upon Levelt's (1989) model of speech production, where the process of producing a spoken utterance is hypothesized to be modular. In the initial steps, after the speaker's intended message has been formed in the Conceptualizer, the message is sent to the Formulator where the lemmas associated with the message are first selected from the Lexicon and then grammatically encoded and annotated. It is the activation of lemmas that triggers syntactic procedures. For example, a V lemma triggers the building of a VP, and an N lemma triggers the building of an NP. This process is then formalized through the concept of *grammatical information exchange*, which is derived from the grammatical framework of Lexical Functional Grammar (LFG; Kaplan, Bresnan, et al., 1981). To illustrate, the sentence "Jag ser stora hundar" (*I see large dogs*) requires exchange or transfer of the number feature between the adjective "stora" and the noun "hundar" within the noun phrase (NP), of which the processing is made available to the learner through the phrasal procedure.[1] Once this procedure is acquired, the learner can then begin to process information exchange *across the phrase*, e.g. between an NP and a VP, as in the sentence "Hundarna är stora" (*The dogs are large*). The distance between the constituents between which information exchange takes place in the hierarchical syntax tree thus increases for each stage in the hierarchy, as illustrated in Figure 2.1 below.

The processes are described in more detail in the list below, retrieved verbatim from Buyl and Housen (2015, p.526):

1. The lemma procedure activates the lexical items.
2. The category procedure accesses the categorical information associated with the activated lemmas.
3. The phrasal procedure builds phrases by unifying information between constituents of the same phrase.
4. The S-procedure exchanges information between phrases in a sentence and accesses the target word order rules.
5. The subordinate clause procedure or S'-clause procedure operates on subordinate clauses, allowing learners to produce target-like word orders which are specific to such clauses.

[1] An example showing that the learner has not yet acquired this procedure could be the equivalent "Jag ser *stor hundar" (*I see large dogs*), where the grammatical number feature has not been exchanged between the adjective and the noun.

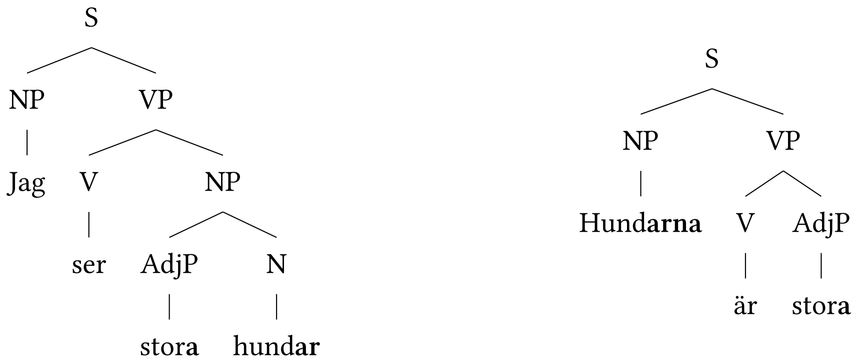


Figure 2.1.: Illustration of the structural complexity difference between attributive agreement (left), where grammatical information is exchanged within the NP, and predicative agreement (right), where information is exchanged between the NP and VP.

2.2.2. Swedish-specific linguistic structures tested within PT

While the procedures of the PT hierarchy are language universal and independent of the language learner's first language (i.e. the developmental trajectory is robust to transfer effects, while transfer may increase the speed of acquisition), the syntactical and morphological structures that are processable at each stage are language specific. Table 2.1 presents the morphological and syntactic structures in Swedish of which the acquisition trajectory as predicted by PT has been empirically tested and widely supported (e.g. Eklund Heinonen, 2009; Glahn et al., 2001; Håkansson and Norrby, 2006; Pienemann and Håkansson, 1999). Each structure is described in more detail in the subsections below and in Appendix A. Ungrammatical examples are marked with an asterix.

Stage	Processing procedures	Syntax	Morphology
5	Subordinate clause procedure	Preverbal negation in subclauses (negV) Non-inversion in indirect questions	
4	Interphrasal procedure	Inversion with topicalization (XVS)	Predicative agreement
3	Phrasal procedure	Non-inversion with topicalization (*XSV)	Attributive agreement
2	Category procedure	Canonical word order (SVO)	Plural, definite form, tense
1	Word access	Chunks	Word/lemma

Table 2.1.: Developmental stages in the L2 acquisition of Swedish morphology and syntax according to PT (Dyson and Håkansson, 2017).

Items from the first developmental stage require no procedure, and are represented in the L2 learner's language by single words, formulae or single constituents such as "Thank you" or "How are you?" that are learned as chunks and have yet not been grammatically analyzed. Since the lexicon is not yet annotated and there is no syntax or morphology to be acquired on this initial stage, it is generally excluded in PT studies, as in this study.

Canonical word order SVO

The category procedure at stage 2 does not contain any unification or information exchange between constituents, but enables mapping syntactic categories and functional roles to the lexicon such as subject and predicate, allowing the learner to structure sentences in canonical SVO order. Swedish is a verb-second language, which means that the finite verb is always placed directly after the topicalized constituent in main clauses (and occasionally after selected clause adverbials). Thus, observed *SOV word order is ungrammatical and indicates that the learner has not yet accessed the category procedure. The contrast between SVO and *SOV is illustrated in example 1 below.

(1)

a. Jag läser boken
I read.FIN book
'I read the book'

b.* Jag boken läser
I book.FIN read

Plural

Access to the category procedure also enables encoding lexical information, which is necessary for marking plural on nouns and tense on verbs. Plural is normally marked morphologically as a suffix on the noun with -or, -ar, -r or -n in indefinite form and -na or -en in definite form, depending on the declension of the noun. While the access to the category procedure in humans' language output can be assessed from all occurrences of plural and non-occurrences of plural in their obligatory contexts, such obligatory contexts must be explicit when using AJT. As an example, the sentence "Elefanterna är här" (*The elephants are here*) and "Elefanten är här" (*The elephant is here*) are equally as grammatical in Swedish, and while the context may be inferred in a speech sample, it cannot be inferred in an AJT where the sentences are isolated. Moreover, while the plural inflection on an attributive adjective modifying a plural noun such as -a on "stor" in "De stora elefanterna är här" (*The big elephants are here*) does serve as obligatory context for plural, it cannot be discerned whether recognizing such a context is due to proper processing of plural or of attributive agreement. Thus, plural numerals were selected to form the obligatory context for the plural feature on the noun used in this study, as in example 2 below.

(2)

a. En grå elefant
one gray.SG elephant.SG
'One gray elephant'

b. Två grå-a elefant-er
two gray-PL elephant-PL
'Two gray elephants'

c.* Två grå-a elefant
two gray-PL elephant.SG

Tense

In Swedish, tense is marked morphologically with inflection or suffixes. While the difference in acceptability when using one tense above another (e.g. "Jag ser[PRS] ett träd" (*I see a tree*) and "Jag såg[PST] ett träd" (*I saw a tree*)) might be inferrable from the context of the utterance, such context is not available in AJT. Thus, in this study the obligatory context of tense is simply identified in all occurrences of tensed verbs, with the verb converted into infinitive in its ungrammatical equivalent. This is illustrated in example 3 below.

(3)
a. Jag såg ett träd
I see.PST a tree
'I saw a tree'

b.* Jag se ett träd
I see.INF a tree

Definite form

The processing required for the morphological marking of definiteness on nouns is also enabled on the second developmental stage of the PT hierarchy. Since three other grammatical structures are already included to represent the 2nd developmental stage in the current study however, definite form as a category is excluded from analysis in this study due to time limitations.

Non-inversion with topicalization

The phrasal procedure at stage 3 enables unification of features within the phrase, and thus gives access to the processing of attributive agreement and topicalizing constituents. However, the process necessary for obligatory inversion of the verb after a topicalized constituent is not accessible until stage 4. This implicates that the language output containing topicalization indicating that a learner has reached stage 3 is ungrammatical by default. Thus, the models cannot be tested on the *XSV structure of this stage, which is why this structure is excluded in the present study.

Attributive agreement

Attributive agreement entails that features of gender (common and neuter), number (singular and plural) and definiteness (indefinite and definite) are unified within the noun phrase and marked on both article, noun and adjective, and thus requires access to the phrasal procedure. The example below showcases this exchange of features. Observe that the plural form of the adjective is identical to its definite singular form. The present study uses the common form of the adjective in the obligatory context of neuter or plural agreement as evidence for ungrammaticality. This entails that attributive agreement with common singular nouns are not identified as examples of this structure.

(4) a. Den stor-a elefant-en
DET.COM.SG.DEF big-SG.DEF elephant.SG-DEF.COM
'The big elephant'

b. Många stor-a elefant-er
many big-PL.IND elephant-PL.COM.IND
'Many big elephants'

c. * Många stor elefant-er
many big.SG.COM.IND elephant-PL.COM.IND

Inversion after topicalization

Through the procedure on stage 4, interphrasal information can be exchanged, allowing the learner to properly use inversion of the finite verb when topicalizing constituents (example 5b below). Swedish is a verb-second language, which means that the finite verb must always be placed directly after the topicalized constituent (XVS). Before the interphrasal procedure stage is acquired, learners will use topicalization without the obligatory inversion of the verb (*XSV), as illustrated in the ungrammatical sentence in example 5 below.

(5) a. Jag ska läsa boken imorgon
I will.FIN read book tomorrow
'I will read the book tomorrow'

b. Imorgon ska jag läsa boken
tomorrow will.FIN I read book
'Tomorrow I will read the book'

c. * Imorgon jag ska läsa boken
tomorrow I will.FIN read book

Predicative agreement

Grammatical information of number and gender is also exchanged interphrasally between the noun in the NP and the adjective in the VP. Mismatching the form of the predicative adjective with the form of the noun it governs is sufficient as counter-evidence for acquisition of the predicative agreement structure. Example 6 illustrates a grammatical and ungrammatical sentence with regards to predicative agreement, including an attractor phrase within brackets. Including an attractor noun may reveal patterns of the model's generalization strategies based on whether or not it marks the adjective with the grammatical information of the attractor in favor of the subject due to their linear proximity.

(6) a. Elefant-er-na [på savann-en] är stor-a
elephant-PL-DEF.COM on savanna.SG-DEF.COM be big-PL
'The elephants on the savanna are big'

b. * Elefant-er-na [på savann-en] är stor
elephant-PL-DEF.COM on savanna.SG-DEF.COM be big.SG.COM

Preverbal negation

The subordinate clause procedure of stage 5 does not contain any unification of features, rather it has as its prerequisite the acquisition of all word order constraints of the main clause. In Swedish, the access to the procedure of this developmental stage is revealed by the L2 learner placing the negation *inte* and other clausal adverbs in front of the finite verb in a subclause. The ungrammatical equivalent would be placing the negation after the finite verb, which is grammatical only in main clause syntax. The subject may precede or follow the negation (see "det" in examples 7a and 7b below). Before this process is accessed, learners generalize the SVneg word order from main clauses to subclauses, resulting in ungrammatical sequences such as example 7c.

(7) a. Jag går inte om det inte är kul
I go.FIN not if it not is.FIN fun
'I am not going if it is not fun'

b. Jag går inte om inte det är kul
I go.FIN not if not it is.FIN fun
'I am not going if it is not fun'

c. * Jag går inte om det är inte kul
I go.FIN not if it is.FIN not fun

Non-inversion in indirect questions

Canceling of inversion after question words in interrogative clauses is also accessed on the 5th and final developmental stage. As in English, in Swedish, direct and indirect questions have different word orders. While the verb precedes the subject in direct questions,[2] in indirect questions the subject precedes the verb in the interrogative subclause.[3] Before entering the 5th developmental stage of the PT hierarchy, L2 learners tend to overgeneralize the word order of direct questions to subordinate interrogative clauses in indirect questions, as illustrated in the examples 8b and 9b below.

(8) a. Jag undrar vad hon inte har gjort
I wonder what she not have.FIN do
'I wonder what she hasn't done'

b. * Jag undrar vad har hon inte gjort
I wonder what have.FIN she not do

(9) a. Jag undrar om hon kommer
I wonder if she come.FIN
'I wonder if she's coming'

b. * Jag undrar om kommer hon
I wonder if come.FIN she

Interrogative subclauses are not grammatically but semantically distinguished from regular relative subclauses. The lemma of the matrix verb is an indicator of the nature of the subclause, where verbs describing inquisitive and cognitive processes such as "undra", "fråga", "fundera", "undersöka", "veta", "gissa", "förklara", "diskutera" and "beskriva" (*wonder, ask, ponder, examine, know, guess, explain, discuss, describe*) are commonly found in the matrix clause (Josephson, 2020). Commonly, the embedded subclause verb (often the head of the question word) functions as a clausal complement.

2.2.3. Measuring receptive grammar acquisition within PT

The present study aims to test *receptive* skills within the PT framework, which presupposes a certain link not only between receptive and productive processes, but also between competence and performance, as well as between implicit an explicit knowledge. This section aims to illustrate the distinction between these concepts as well as to present related work, in order to motivate the aim and methodology of the present study.

PT is built upon Levelt's (1989) model of speech *production*, which inherently views production and reception as separate cognitive processes. Dyson and Håkansson (2017, p. 11) write that "[a]lthough the generativist distinction between performance and competence is [...] not theorized in PT, the concern with skill means that PT focuses on language performance not competence." Pienemann (1998b, p. 52) himself writes that "it is clear that comprehension and production are not mirror images of each other". It is thus not surprising that the vast majority of PT studies concern speech production, with a few of them (e.g. Norrby and Håkansson, 2007) including written data. However, as Spinner (2013, p. 710) points out, "there is reason to believe that the procedures of PT might also operate in the decoding of grammatical structures in the input". Four studies have investigated this claim by analyzing grammatical *comprehension* within the PT framework, namely R. Ellis (2008), Keatinge and Keßler (2009), Spinner (2013) and Buyl and Housen (2015).

R. Ellis (2008) made an important – and for this study relevant – distinction between implicit and explicit L2 knowledge by examining participant results from four tests on 17 structures from the English PT hierarchy: an oral imitation test, a timed

[2] e.g. "Vad äter du?" (*lit. What eat you (What are you eating)?)*) or "Äter du?" (*lit. Eat you (do you eat)?*)
[3] e.g. "Jag undrar vad/om du äter" (*lit. I wonder what/if you eat*)

written AJT, an untimed written AJT and a metalinguistic knowledge test. A Principal Components Factor Analysis[4] across the results from all four tasks revealed that the oral imitation test and the timed AJT loaded on one factor while the untimed AJT and the metalinguistic knowledge test loaded on another factor. Ellis thus concluded that the distinction between the measures should be drawn between implicit and explicit knowledge rather than between production and "decision". The claim that the timed AJT, in opposition to the untimed AJT, should test implicit knowledge is supported by the notion that "people adopt some preferred interpretation immediately, rather then delaying interpretation" (Crocker and Keller, 2006, p. 2). As such, the timed AJT – reliant on immediate responses from the participants – reflects the learner's interlanguage in a way that untimed AJT do not.

In order to test the PT hypotheses, Ellis used *implicational scaling*[5] on the results from the oral imitation test and the untimed AJT. He found that the rank order of difficulty aligned with the order predicted by PT for the oral imitation test but not necessarily for the untimed AJT. R. Ellis (2008) concluded that PT cannot predict the learning difficulty of explicit knowledge since it does not tap into the learners' interlanguages, as opposed to implicit knowledge.

From Ellis's study, Dyson and Håkansson (2017, p. 11) drew the conclusion that "PT does not predict the difficulties with explicit knowledge, such as the knowledge required to make grammaticality judgements". They did not address, however, the distinction that R. Ellis (2008) made between the timed and the untimed AJT, of which the former was not included in the PT analysis on basis of the fact that the PT framework concerns learner production rather than reception.

Timed AJT as a measure for implicit knowledge has been supported in other studies (R. Ellis, 2005; R. Ellis and Loewen, 2007; Han and R. Ellis, 1998) and was further explored within the PT framework by Spinner (2013). Spinner (2013) performed timed audio AJT using fifteen different structures from stage 2 through stage 6 in the English PT hierarchy, with five grammatical and five ungrammatical items per structure. Using accuracy and implicational scaling for analysis, Spinner (2013) found that the results did *not* align with the predicted hierarchy, however, indicating that PT does not predict the order of acquisition of processing procedures for reception as it does for production. Spinner (2013) proposed that these results could be attributed to either "(a) processing routines for production and reception are completely separate" or "(b) processing routines are the same in production and reception, but factors unique to reception mask or disrupt their normal operation in reception", among other theories (Spinner, 2013, p. 731).

Instead of using AJT, Keatinge and Keßler (2009) examined English L2 learners' current state of language development by observing their perception of the single feature of passive voice through a picture task. A sentence, either in passive or active voice, was read to the participants, who were then asked to select the appropriate picture out of several options (e.g., "The red fish is eaten by the blue fish"). One point per correct match was given. They found that learners below the 4th developmental stage had 1 or less points, and concluded that comprehension of the passive voice was one or two stages further developed (emerged earlier) than the production of the same feature.

[4] Principal Components Factor Analysis is a statistical method aiming to find underlying correlations between variables by grouping them into common factors.

[5] Implicational scaling is used widely in L2A research and in PT studies in particular to determine if there's a hierarchical relationship between the acquisition of different language structures. The analysis detects the case where acquisition of one structure implies acquisition of other "simpler" structures, while the reverse is not necessarily true if such a relationship is found (Rickford, 2004). Implicational scaling is described in more detail under the methodology section.

While Buyl and Housen (2015) focused on only one structure and used a limited number of 10 participants, Buyl and Housen (2015) did a similar study with a larger participant pool of child L1 French learners of L2 English, testing 6 linguistic structures from stages 2 and 5 in the English PT hierarchy. Buyl and Housen (2015) also used a picture selection task, requiring the participants to match a picture with one out of three options (For example, a picture of a cat and the options "cat" (incorrect), "cats" (correct) and "dog" (distractor)). In contrast to Spinner's results, Buyl's results indicated alignment with the development sequence predicted by PT. Buyl proposed that this discrepancy may be due to the fact that stages 3 and 4 were missing in their study, and that they counted a procedure as acquired if at least one of the structures of the respective stage was acquired, while Spinner required 2 out of 3 structures to be acquired.

The findings from R. Ellis (2008), Keatinge and Keßler (2009), Spinner (2013), and Buyl and Housen (2015) suggest that the application of PT to receptive skills requires further exploration. The lack of unity in the methodological approaches, as well as the different results across these studies, highlight the need for additional research to determine whether PT can predict receptive processing sequences with the same reliability as productive sequences. In the present study We aim to increase interpretability and transparency of our results by using a broad range of linguistic structures covering all stages of the PT hierarchy, while evaluating the results using both visual plots of the learning trajectories, implicational scaling and rank correlation metrics with different thresholds of acquisition.

Without full knowledge of how the cognitive processes of humans differ from those of artificial learners such as GPT-2 when performing AJT, it cannot be ascertained whether the AJT applied in this study resemble the "timed" or "untimed" tests applied by R. Ellis (2008), nor if the distinction between implicit and explicit knowledge is even applicable to language models. Nevertheless, seeing as the scores from the AJT as implemented in the present study reflect the same probability distribution that is used by the model to generate language output, We assume the "cognitive processes" behind production and reception to be equivalent in GPT-2.

2.3. Human vs. artificial language acquisition

2.3.1. Generalization and hierarchical inductive bias

In this section We aim to shed light on the differences and similarities between human and artificial language learning relevant to the analysis of linguistic knowledge, in order to warrant the approach of studying artificial language development within a framework of human language acquisition theory.

A key feature of human language acquisition is believed to be our access to an internal grammar that yields an unbounded number of hierarchical expressions (e.g. Chomsky, 1957). Consider the sentence "The keys to the cabinet are on the table", in which correctly predicting "are" as the correct token in favor of "is" requires identifying that it is the keys that are on the table and not the cabinet, more specifically that the verb is agreeing with the subject "the keys". Even though "the cabinet" is linearly closer to the verb than the subject is, a speaker of English would normally interpret the sentence correctly despite having never heard that exact sequence of words before. In other words, we would *generalize* from previous linguistic examples, and favor the hierarchical interpretation over the linear one in doing so.

The question of whether artificial learners can generalize syntactic structures in the same way as children do when acquiring their first language, as opposed to relying only on surface-level statistics (heuristics), has been the object of great interest in recent years. A common approach is to test models on subject-verb agreement using *attractors*, such as "the cabinet" in the example described above. If the model fails to recognize the true dependent of the verb in the presence of the attractor, the conclusion is drawn that the model lacks hierarchical constraints on its generalization abilities.

According to the influential Universal Grammar theory, the hierarchical bias of humans is due to us possessing an innate language faculty which constraints our linguistic interpretations (Chomsky, 1980). This claim is in turn supported by Chomsky's "poverty of the stimulus" argument, which assumes that children can acquire and generalize from complex linguistic rules that encompass hierarchical syntax despite being exposed to insufficient, incomplete or ambiguous data. In comparison, this does not seem to be the case for artificial language models. While children can acquire language on less than 100 million words of input, language models typically require data of the billions (Warstadt et al., 2023) – GPT-2 for instance was trained on 8 million web pages of data.

Some research on Long-Short-Term-Memory (LSTM) and Gated Recurrent Unit (GRU) architecture indicate that sequence-to-sequence (seq2seq) models that lack hierarchical constraints have a linear bias and tend to struggle with syntactic tasks (e.g., Linzen et al., 2016; McCoy et al., 2018). Similar results were found when testing LSTMs and Transformer models trained on multimodal input similar to that provided to children (Yedetore et al., 2023). Other findings suggest that only tree-based models explicitly built to represent syntax can generalize hierarchically consistently (McCoy et al., 2020).

In contrast, other recent studies demonstrate that certain types of models, particularly large, pretrained seq2seq models, *can* exhibit hierarchical generalization. Mueller et al. (2022) fine-tuned mono- and multilingual BART to perform syntactic transformations, and found that hierarchical generalizations were possible even across long-span dependencies, *as long as the model was pretrained on a much larger amount of structured language data than humans receive*. While supporting the "poverty of the stimulus" argument, such results indicate that hierarchical generalization is possible without explicitly inducing a hierarchical bias in the model's structure.

One explanation for *how* models learn to generalize hierarchically was suggested by Hewitt and Manning (2019) and Manning et al. (2020), who found that models trained with self-supervision learned syntactic dependencies and hierarchical structures through hierarchical representations, without being specifically trained for such a task. They investigated the internal vector representations of linguistic structures of BERT and ELMo by measuring distances between them, and found that approximations of entire syntax trees were embedded in the models' vector spaces.

In the present study, we include two evaluation sets of the predicative agreement structure, where one includes an attractor noun that differs from the subject in number or gender, and that is linearly closer to the predicative adjective than the subject. By including an attractor noun, we hope to reveal patterns of the model's generalization strategies and connect such patterns with the developmental trajectory hypothesized by PT.

2.3.2. The probabilistic account

Challenging the importance of hierarchical bias are studies that suggest that models *without* explicit hierarchical structure are superior in modeling human processing difficulty, and that probability scores can better predict such difficulties than inductive bias (Fossum and Levy, 2012; Frank and Bod, 2011).

The idea that humans process language input incrementally, word by word, and in real-time, is a well-established finding in psycholinguistics and cognitive science (e.g., Altmann and Kamide, 1999; Hale, 2001; Levy, 2008; Tanenhaus et al., 1995). Comprehension of a string of words encompasses anticipating upcoming structures and words and revising our interpretation of a sequence when we misparse it, e.g. in the case of so-called garden-path sentences; The sentences "The old man the boat" and "The horse raced past the barn fell" are two well-known examples.

One idea that has gained much support from linguistics research is Hale's (2001) surprisal theory, which argues that the processing difficulty of a word is proportional to its probability in the context within which it appears (Frank et al., 2013; Levy, 2008; Smith and Levy, 2013). In these aspects, human language processing is similar to that of causal language models such as GPT-2, that are designed to predict the next token in a text sequence based on its previous context. When generating text, the model relies on predictions that are based on a probability distribution derived from the training data. During "comprehension", when the model is processing input through its decoder, processing difficulty can thus be quantified with *surprisal* – the negative logarithm of the probability of the word in its context. The assumption is then that there is a direct correlation between a model's next-word prediction ability and its ability to predict human estimates of processing difficulty – its "psychometric predictive power" (Goodkind and Bicknell, 2018; Wilcox et al., 2020).

Wilcox et al. (2020) investigated the correlation between LLMs' predictive power and their performance on syntactic tasks. While GPT-2 was superior in performing such tasks in comparison to other models, they found no significant correlation between its Syntactic Generalization scores and its predictive power, after controlling for perplexity. Oh et al. (2022) further investigated the predictive power of GPT-2 models of different sizes and made the surprising finding that larger GPT-2 models' perplexity scores had a *positive* correlation with predictive power, in contrast to the negative correlation found in previous studies. They suggest that "when the training data is held constant, high-capacity LMs may be able to accurately predict the upcoming word while relying less on humanlike generalizations, unlike their lower-capacity counterparts" (Oh et al., 2022, p. 14).

2.3.3. Learning trajectories

While learning trajectories in language models is a fairly new area of research, several recent studies have examined how and when language models acquire different linguistic phenomena during pretraining (Blevins et al., 2022; Chiang et al., 2020; Choshen et al., 2021; Liu et al., 2021; Saphra and Lopez, 2018). Choshen et al. (2021) used a similar approach to the present study, training multiple models including GPT-2_small on English, and evaluating them on AJT with BLiMP. They found a systematic learning trajectory across LLMs with different initializations, architecture and training data, albeit at different speeds. The study showed that morphological phenomena was found to emerge at similar stages. In initial stages, the LLMs were found to rely on local cues such as the frequency of the preceding words, similarly to bag-of-words (BOW) models. Performance is high for tasks such as Part-of-speech (POS) tagging during this stage (Saphra and Lopez, 2018). This correlation subsides as training progresses, as the models seem to apply different strategies. For some simple linguistic structures, this change in strategies can cause accuracies that start high to drop (Choshen et al., 2021).

In later stages of training, the LLMs' accuracy scores correlate with those of n-gram models, suggesting that the models are relying less on simple frequencies and more on structural cues and global features. Simultaneously, syntactic depth becomes a greater predictor to performance than sentence length. As training progresses, the

LLMs' performances become more similar to humans', eventually reaching a plateau (Choshen et al., 2021).

While the linguistic phenomena and their acquisition trajectories in these studies are not categorized in accordance with their hypothesized processability, the observed progression from local to global cues aligns with the progression across the developmental stages as described within the PT framework. This further motivates the aim of the present study.

3. Methodology

This chapter describes the adopted methodology in the present study, aiming to investigate the alignment of GPT-2's acquisition order of grammatical structures with the trajectory hypothesized by PT. We fine-tune four GPT-2 models pretrained on English on unlabeled Swedish data using three different curricula, and evaluate their developmental stages in acquiring grammatical structures through SwePT – a minimal pair dataset containing Swedish sentences that include target grammatical structures from the PT hierarchy.

3.1. SwePT – a Swedish PT minimal pairs dataset

I present the Swedish Processability Theory Minimal Pair Dataset (SwePT), consisting of nine subsets of minimal pairs representing four stages of the Swedish PT developmental hierarchy, namely SVO (canonical word order SVO, 2nd stage), PLUR (plural, 2nd stage), TENSE (tense, 2nd stage), ATTR (attributive agreement, 3rd stage), PRED_a (predicative agreement, 4th stage), PRED_b (predicative agreement with attractors, 4th stage), INV (inversion after topicalization, 4th stage), NEGV (preverbal negation, 5th stage) and INQ (Non-inversion in indirect questions, 5th stage). Examples of minimal pairs of each subset are presented in Table 3.1.

The grammatical sentences of each minimal pair were extracted from the Swedish Talbanken and LinES treebanks from UD (De Marneffe et al., 2021). Although Talbanken and LinES are limited in size compared to other Swedish open-source datasets, they were selected mainly due to their high-quality annotations. LinES contains 4,564 annotated trees and 79,812 tokens worth of data translated from English, including literary works, online help manuals and the Swedish part of the Europarl corpus. Talbanken consists of roughly 6,000 annotated sentences and 95,000 tokens from a variety of informative text sources including newspaper articles and textbooks. Both treebanks were merged into a single CoNLL-U data file before processing.

3.1.1. Processing pipeline

SwePT was constructed with an automated approach similar to that of RuBLiMP (Taktasheva et al., 2024). The target linguistic structures were identified through a processing pipeline consisting of nine rule-based Python scripts targeting each PT structure, respectively.[1] The scripts were written by performing several manual iterations of systematically relaxing the heuristics and reviewing the output. The criteria for identification and perturbation of the structures are found in section A in the appendix.

The pipeline performs three main consecutive steps: 1) identifying and extracting sentences containing the PT structures from the source CoNLL-U files through a dependency tree search, 2) duplicating the sentences to form the minimal pairs, and 3) altering the duplicates into ungrammatical sentences with respect to their target structures. The first step of this process was also used for labeling the training data, which is elaborated on in section 3.2.1.

[1] The scripts and datasets are available here: https://github.com/stellson/SwePT

Structure	Pairs	Example
5 NEGV	303	Men det är viktigt, att förlusterna [inte] [blir] onödigt stora. *Men det är viktigt, att förlusterna [blir] [inte] onödigt stora. (*But it is important that the losses are not unnecessarily large.*)
5 INQ	94	Jag har lust att fråga honom varför [den] inte [trycktes]. *Jag har lust att fråga honom varför [trycktes] [den] inte. (*I want to ask him why it wasn't printed.*)
4 INV	2581	Ovanpå ett skåp i hörnet [satt] [Dobby] hopkrupen. *Ovanpå ett skåp i hörnet [Dobby] [satt] hopkrupen. (*On top of a cupboard in the corner crouched Dobby.*)
4 PRED_a	226	De flesta u-länder har varit [koloniserade] *De flesta u-länder har varit [koloniserad] *(Most developing countries have been colonized)*
4 PRED_b	27	Resultaten av uppväxten i denna miljö är rätt så [uppenbara]. *Resultaten av uppväxten i denna miljö är rätt så [uppenbar]. (*The results of growing up in this environment are quite obvious.*)
3 ATTR	213	Han har inget [civiliserat] ansikte. *Han har inget [civiliserad] ansikte. (*He does not have a civilized face.*)
2 TENSE	2000	Jag [är] min fars dotter. *Jag [vara] min fars dotter. (*I am my father's daughter.*)
2 PLUR	479	Måste du försöka göra åtta [saker] samtidigt? *Måste du försöka göra åtta [sak] samtidigt? (*Must you try and do eight things at once?*)
2 SVO	2519	Hon [hade] [en dämpad, tonlös röst] och bröt inte så kraftigt som mannen. *Hon [en dämpad, tonlös röst] [hade] och bröt inte så kraftigt som mannen. (*She had a soft, dry voice and her accent was slighter than her husband's.*)

Table 3.1.: Selected examples of minimal pairs (a grammatical sentence and its ungrammatical equivalent) from SwePT, including their translations. The target structures are displayed within square brackets.

Double citation marks were removed from the sentences due to spacing issues during the conversion into the ungrammatical sentences. All tokens without integer indices (floats representing implicit, omitted words in elliptical structures) were skipped, since these tokens are not explicit in the original sentences.

To form the minimal pairs of the syntactic structures (SVO, INV, INQ and NEGV), relevant grammatical constituents and arguments were identified and had their positions switched with respect to the target structure. The alteration of the morphological structures (PLUR, TENSE, ATTR and PRED_a) was performed by converting the conjugated target structures into their neutral form (lemma). The alteration process for the PRED_b minimal pairs was performed manually in order to minimize errors, due to the small amount of extracted sentences and the complexity of the alteration task. The details of the processes involved in creating each subset of minimal pairs in SwePT are described in Appendix A.

In cases where more than one instance of the target structure was found in the same sentence, only one instance was modified in the ungrammatical sentence.

The pipeline handles whitespace before and after punctuation to ensure that the grammatical and ungrammatical sentences do not differ in more aspects than the target structure. The script allows no duplicates and only returns the minimal pair if the grammatical and ungrammatical sentence are not identical.

In order to evaluate the minimal pair generation, a manual analysis was performed with the aim to identify the positive predictive value (precision) of the minimal pair generation process. 50 random minimal pairs from each subset of SwePT were examined, of which 25 pairs originated from the LinES corpus and 25 pairs from the Talbanken corpus. The false positives, in terms of the number of pairs containing incorrectly identified grammatical structures or incorrect generation of the ungrammatical sentence, were counted. The error rate per minimal pair subset was then calculated with a 95% Wilson Score Confidence Interval in order to account for the small sample size. The minimal pairs were found to have an average error rate of 2.89%, which corresponds to a precision score of 97.11%. The false positives and the corresponding error rates per minimal pair subset are presented in Table 3.2 below. Observe that the PRED_b subset does not contain any false positives by default, since the ungrammatical sentences were manually generated.

Subset	SVO	PLUR	TENSE	ATTR	PRED_a	PRED_b	INV	NEGV	INQ	Avg.
FP	2/50	2/50	0/50	2/50	3/50	0/50	0/50	2/50	2/50	
Error rate	4%	4%	0%	4%	6%	0%	0%	4%	4%	2.89%

Table 3.2.: Error rates per subset in SwePT. 50 randomly extracted minimal pairs from each subset were examined manually. The false positives (FP) were counted and the error rate was calculated with a 95% Wilson Score Confidence Interval. The average FP score is 2.89%, which corresponds to a precision score of 97.11%.

3.2. Experimental Setup

I fine-tuned six GPT-2 models (pretrained on English) on Swedish unlabeled data and tested the models on their linguistic knowledge through acceptability judgment tests at different time steps during fine-tuning. We evaluated the performance in terms of the models' adherence to the acquisition trajectory as predicted by Processability Theory (PT).

3.2.1. Fine-tuning data

For fine-tuning, we used the Swedish partition of the Common Crawl corpus Open Super-large Crawled Aggregated coRpus (OSCAR)[2]. Due to limited computational resources only 9% of the dataset (approx. 680k examples and 1B tokens) was extracted for the training set after shuffling the data (seed=42). The data was processed in multiple steps, with the objective to separate the data into four subsets representing stages 2-5 in the PT hierarchy. The parsing, labeling and grouping processes are described in the sections below.

Parsing

The dataset was first partitioned into ten separate files for parallel processing. Each sentence in the data was first parsed, annotated and converted into CoNLL-U format using Stanza (Qi et al., 2020), a Python package intended for linguistic analysis. Stanza was chosen over other popular parsers due to its reported high performance on Swedish dependency parsing[3] and its intuitive Python API. We used the `tokenize`, `pos`, `lemma` and `depparse` tools in our pipeline, with batches of 50 on one GPU. Certain sequences in the data (e.g. calendar tabulars) caused an `ArrayMemoryError` and were thus skipped, while the remaining sentences of the batch were processed individually in a separate function to minimize the loss of data. Web addresses were also removed using regex matching. Due to memory allocation limits during parsing, the dataset was separated into chunks after time-outs and processed separately based on indexing, after which

[2] https://huggingface.co/datasets/oscar-corpus/OSCAR-2201
[3] https://stanfordnlp.github.io/stanza/performance.html

the files were concatenated into a single CoNLL-U file.[4] Through this process, the dataset was reduced to 25,758,263 sentences/examples and 1,021,799,273 tokens (after truncating with max_length=512). As described in Section 3.2.1,

Since a manually inspected sample would be too small to yield meaningful insights, we evaluated parsing performance by comparing the distribution of linguistic categories in the parsed OSCAR subset with those in the gold-standard Talbanken and LinES corpora. The Chi-square test shows that the parsed OSCAR data differs significantly from the two gold-standard corpora. A visualization of the distribution[5] shows that the syntactic categories (SVO, INV, NEGV, INQ) are consistently more frequent in the evaluation data, compared to in the training data, while the opposite relationship is observed among the morphological structures (ATTR, TENSE, PLUR, PRED). Stanza has a reported 87.85 labeled attachment score (LAS)[6] evaluated on the Talbanken treebank. While this is considered a high score for Swedish dependency parsing, in comparison with the performance of other publicly available parsers for Swedish, it does leave room for improvement. It is likely that the observed difference in populations is attributed to error propagation from earlier stages of the processing pipeline, or to natural domain differences between the corpora. In comparison to the manually annotated and corrected Talbanken and LinES, OSCAR as a common crawl corpus is expected to contain noisy data and strings that are independent from grammatical structure such as headers and descriptions, thus increasing the dominance of morphological structures. Although a more thorough evaluation of parsing quality would have been desirable, we will assume that the output is good enough for the present purposes.

Dependency Tree-Search and Labeling

After parsing, the CoNLL-U sentences were processed through the same functions used to identify the structures for the SwePT dataset, from which each sentence is returned labeled with the structures identified within it.[7] 500,000 sentences were then randomly sampled from the file with labeled sentences in order to ensure that one epoch of training across the entire dataset would fit within 72 h of training (as calculated during a test run).[8]

Grouping of Fine-tuning Data

After parsing the raw text from OSCAR and labeling the training data, the sentences were grouped into four subsets representing each of the developmental stages (2–5) in the PT hierarchy, and converted into a `DatasetDict`. The subsets were populated in decreasing order, and the sentences in each subset thus only contains 1) structures from its respective stage, and 2) structures from lower stages, if occurring within the same sentences. The sentences that remained unlabeled after labeling (i.e., no PT structures were identified within them) were distributed into the four subsets in proportion to the original size of each subset. The distribution between stages is shown in Table 3.3. Observe that the subsets are different in size, since each developmental stage

[4] Due to unforeseen issues during this process, some of the data was skipped unintentionally. This should be taken into account if replicating this study.

[5] See Table B.1 and Figure B.1 in Appendix B.

[6] https://stanfordnlp.github.io/stanza/performance.html

[7] See section A in the appendix for details on how each grammatical structure was identified.

[8] While it would have been more time effective to only parse 500,000 sentences of OSCAR instead of parsing the entire dataset and reducing it greatly during this step, the test run that determined the amount of training data necessary for the fine-tuning task was performed after the parsing process.

Structure	Stage-2	Stage-3	Stage-4	Stage-5	Total
INQ				1 362	1 362
NEGV				10 645	10 645
PRED_b			51	1	52
PRED_a			7 855	217	8 072
INV			111 195	3 156	114 351
ATTR		64 569	26 289	2 848	93 706
TENSE	194 702	49 118	110 559	11 833	366 212
PLUR	13 329	4 001	6 992	397	24 719
SVO	47 764	15 241	15 604	3 004	81 613
Unlabeled	55 268	17 878	31 982	3 295	108 423
Sentences/stage	254 875	82 447	147 490	15 188	500 000

Table 3.3.: Distribution of the linguistic structure labels of sentences in all four training subsets (stages 2-5). The subsets are populated in decreasing order, and the sentences in each subset thus only contains 1) structures from its respective stage, and 2) structures from lower stages, if occurring within the same sentences. Unlabeled sentences are also included in each subset (amount proportionate to the size of the subset).

are represented by different numbers of linguistic structures that occur in varying frequencies in the training data.

3.2.2. Fine-tuning

Language Model Architecture

Four small (124 M parameters) GPT-2 models[9] pretrained on English were fine-tuned and evaluated in the present study. As a language model, GPT-2 learns to predict and generate text based on a probability distribution over sequences of words. It estimates $P(w_n|w_1, ..., w_{n-1})$, e.g. the probability of the next word given its previous context (Radford et al., 2019). GPT-2 is a causal (unidirectional) transformer, meaning that it takes only the previous and not the future context into account when generating predictions. This aspect is similar to the incremental processing of humans (e.g., Altmann and Kamide, 1999; Kuribayashi et al., 2025), which is one the reasons that this model was chosen for this project. Another reason is the relatively small size, allowing for effective fine-tuning on a smaller dataset, which was crucial due to limited computing resources.

Curriculum learning

I employ the method of curriculum learning (Bengio et al., 2009) during fine-tuning, where models are initially fine-tuned on simpler concepts and gradually move on to more complex concepts. Similar applications of curriculum learning, such as the work of Tsvetkov et al. (2016), supports the assumption that structured input sequencing can be beneficial to model performance. Tsvetkov et al. (2016) used Bayesian optimization to find the optimal curriculum and found that sorting the data in terms of increasing *or* decreasing syntactic complexity (determined crudely in terms of sentence length) as well as ordering the data in terms of increasing NPs and decreasing VPs, PPs and parse tree depth could lead to improvements over the baseline. While the curriculum in the present study is determined through a single, pre-hypothesized parameter of

[9] https://huggingface.co/openai-community/gpt2

complexity, the positive results of Tsvetkov et al. (2016) provide further motivation for using curriculum learning to test PT on a computational model such as GPT-2.

In the present study, we used three different curricula including one randomized input order, in order to test the robustness of the implicational acquisition order as stipulated by PT. We fine-tuned one model instance on input data ordered from simpler to more complex (`GPT-order`), one in reverse order (`GPT-mixed`) and two on all four subsets concatenated into one dataset (`GPT-mixed` and `GPT-mixed_2`), thus exposing the models to a randomized curriculum. The models were trained for 72 hours for one epoch. If `GPT-reverse` displays a similar acquisition order as the other models, it is implicated that the implicational acquisition order as stipulated by PT holds.

Table 3.4 presents an overview of all four models.

Models	Fine-tuning Data	Curriculum Order	Seed
`GPT-mixed`	OSCAR sampled 500k	Shuffled	42
`GPT-mixed_2`	OSCAR sampled 500k	Shuffled	123
`GPT-order`	OSCAR sampled 500k	Increasing	42
`GPT-reverse`	OSCAR sampled 500k	Decreasing	42

Table 3.4.: Overview of all GPT-2 models evaluated with AJT on SwePT. Increasing curriculum order signifies that the model was trained on data in increasing complexity as determined by PT, starting with the Stage-2 subset. Decreasing curriculum order signifies that the model was presented with the subsets in decreasing complexity order (Stages 5-4-3-2).

Training Arguments

The GPT-2 models were fine-tuned using the `Transformers` library from Hugging Face. We used the pretrained `AutoTokenizer` with the padding token set to the end-of-sequence (EOS) token, with padding and truncation at a max length of 512[10]. We trained for 1 epoch using the `Trainer` API with an effective batch size of 32 (16 batches per device with gradient accumulation steps of 2), and the `AdamW` optimizer with a learning rate of 2e-5, a weight decay of 0.01 and half-precision (fp16) to speed up training. Checkpoints were saved at each 100th time step and named according to their indices.

3.3. Evaluation

3.3.1. Acceptability Judgment Test

I follow the approach of Evanson et al. (2023) in conducting the AJT. At each checkpoint (every 100 training steps), we measure how acceptable the model finds each grammatical and ungrammatical sentence (i.e., a minimal pair), by summing the log-likelihoods of their tokens, as derived from the model's loss outputs. More specifically, the score is calculated as follows, $-\mathcal{L}(M, X) \times N = \sum_{t=1}^{N} \log P(x_t \mid x_{<t}, M)$, where the total log-likelihood of a sentence S equals the cross-entropy loss $\mathcal{L}(M, X)$ (negative average log-likelihood) of N tokens in the sentence.

The performance of each checkpoint is assessed by comparing the scores assigned to the grammatical or the ungrammatical sentence of each minimal pair at each

[10] Each example in the data subset corresponds to one sentence, meaning that truncation is applied on the sentence level. The sentences have an average length of 40.08 tokens (with the caveat that many examples may consist of single titles or headers, which contribute to lowering this average), with 23,106 sentences (0.09%) exceeding the 512 tokens limit.

checkpoint. The accuracy is calculated as the percentage of the pairs where the grammatical sentence was correctly identified, that is, was given a higher score than its ungrammatical counterpart.

In the present study, we use *acceptability* in favor of *grammaticality* when referring to the classification task, in light of the discussion under Section 2.1. It should be emphasized, however, that the rule-based scripts used for the generation of the minimal pairs rely on a rule-based grammar, and that inconsistencies where a sentence's grammaticality does not correspond to its acceptability may occur. Examples of such inconsistencies are discussed in Appendix A.

3.3.2. Acquisition time and the emergence criterion

When assessing L2A within the PT framework, the human L2 learner's speech output is usually observed and analyzed during a communicative task designed to elicit obligatory contexts for target grammatical features. While most L2A theories use native-like performance or accuracy as its metric for assessing grammatical knowledge, in PT studies the current level of the learner's language development is determined using the *emergence criterion* (Pienemann, 1998b). Emergence of a certain grammatical rule is represented by a learner's first production of a token of that rule, and marks the onset of the procedure that enables its acquisition. More specifically, the emergence criterion relies on consideration of four possible cases, namely (1) a lack of evidence (i.e. no present obligatory context for the target rule), (2) insufficient evidence (i.e. not enough number of examples), (3) counter-evidence (i.e. non-application of the rule in the presence of its obligatory contexts) and (4) evidence of rule application (i.e., sufficient examples of applications of the rule in the presence of its obligatory context; see Pienemann, 1998b.[11] Using the emergence criterion, PT assessment can account for the fact that L2 performance rarely is linear, meaning that L2 learners often produce grammatically inaccurate structures even after having acquired them.

Since acceptability judgments and not language output are used for measuring the models' acquisition in the present study, the emergence criterion must be adapted and reduced to case (3) (interpreted as higher average score assigned to the grammatical sentence) and (4) (interpreted as a higher average score assigned to the ungrammatical sentence). Moreover, before reaching a certain threshold during training, the model is expected to distribute the probabilities over the grammatical and ungrammatical sentences somewhat randomly, and a single correctly identified grammatical sentence would say very little about the model's acquisition of the structure without the context of the cumulative probabilities of the entire subset. Thus, a threshold at which the model can be considered to have acquired a structure must be defined.

Evanson et al. (2023) measured the acquisition time of a specific structure as the time step where the model has reached 90% of the final accuracy on that structure. However, they focused more on *mastering* of skills, which is of less interest within PT. Thus, in the present study, acquisition time or emergence is approximated as the models' first systematically correct predictions of the grammatical sentences, that is, when the cumulative accuracy has reached above chance level. To account for some noise around the chance level mark, we set this acquisition threshold at 60% accuracy. Buyl and Housen (2015) evaluated at both 50% and 80%, upon similar discussion. In favor of transparency and interpretability, we will evaluate on thresholds of both 50%, 60% and 80% accuracy.

[11] What number of examples that constitutes sufficient evidence varies across languages and studies. For example, while Pienemann (1998b) has initially suggested minimally one occurrence per sample for the syntactic structures as evidence for emergence, Håkansson and Norrby (2010) required two occurrences in their study.

Structure	n	k if π=50%	k if π=60%	k if π=80%
2_SVO	2519	1302	1553	2049
2_PLUR	479	258	306	398
2_TENSE	2000	1038	1237	1630
3_ATTR	2268	1174	1400	1847
4_PRED_a	213	119	140	181
4_PRED_b	27	19	21	26
4_INV	2581	1333	1590	2099
5_NEGV	303	167	197	255
5_INQ	94	56	65	82

Table 3.5.: Acquisition thresholds k per subset where n = number of minimal pairs, k = the number of pairs needed to be correct to ensure acquisition above the threshold π and $\alpha = 0.05$.

Following the approach of Buyl and Housen (2015), we will calculate the k number of sentences per subset that must be correct in order to ensure acquisition at each threshold. The acquisition threshold is determined per subset in relation to the number of minimal pairs (n) by finding the smallest number of correct guesses (k) possible to exceed chance level ($\pi = 50$) at the determined significance level ($\alpha = 0.05$). We used the reliability function representation, rewritten as $P(X \geq k) = \text{BinomSF}(k-1, n, \pi)$. The acquisition thresholds for 50%, 60% and 80% accuracy are displayed in Table 3.5.

3.3.3. Implicational scaling

In order to test whether the acquisition of grammatical structures follows a hierarchical, implicational pattern across learners, as predicted by PT, implicational scaling (Rickford, 2004) is used. Implicational scales are binary matrices that visualize what structures are acquired by each learner at the time of evaluation. PT predicts only the order of acquisition and thus allows for variation in terms of the speed in which learners acquire the processing procedures as well as the order among the structures that belong to the same developmental stage. In PT studies, using implicational scaling as a metric to measure consistency across individual learners' rank orders is standard practice, as it can account for learner variation within the theorized constraints of PT (Pienemann, 1998a). In our study, each checkpoint across all fine-tuned models are treated as individual learners in the implicational scales, thereby approximating a longitudinal PT study where the same learner is tested at multiple points throughout learning. Although treating each checkpoint independently may seem counter-productive considering that the purpose of this study is to examine developmental stages, implicational scaling aims to determine whether the observed acquisition order is implicational, that is, whether structures from lower developmental stages are *required* for the learner to acquire more complex structures. As such, the scales are agnostic to the specific point in time at which a learner is evaluated. For the purpose of examining learning trajectories, other methods are used.

A coefficient or index of reproducibility (IR) is often used to give an indication of the scalability of the implicational scales, with a smaller IR expressing that a high number of entries deviate from the pattern, and a larger IR indicating a more consistent scale (1.0 reflecting a perfect scale). However, since multiple structures can be acquired during the same stage (for example, PT does not predict whether SVO is acquired before plural or tense since they are all processable at the same developmental stage), IR should be interpreted with caution.

3.3.4. Learning trajectories and rank correlation

While implicational scaling is the main evaluation method applied in traditional PT studies, studies on NLM's learning stages generally visualize learning trajectories and perform statistical methods to determine whether the trajectories are consistent across models. In the present study, we perform a rank correlation permutation test[12], inspired by Evanson et al. (2023) and Liu et al. (2021). We rank the PT structures in terms of their acquisition time (the number of steps taken to reach an accuracy above the respective acquisition threshold) and then compute the rank correlation between each pair of the five models and average it. A null distribution is then created by randomly shuffling the ranks in one model per pair and recomputing the average correlation. If the true average correlation is higher than the correlation from the null distribution, the acquisition trajectory is consistent.

[12] This is also known as the Spearman's coefficient of rank correlation, or Spearman's ρ (Gibbons and Chakraborti, 2014).

4. Results and analysis

This chapter presents the empirical results and analysis, aiming to evaluate the adherence of the learning trajectory of GPT-2 to the developmental stages as hypothesized by PT. Several different approaches are used to investigate different aspects of acquisition order. In section 4.1, the results from the AJT at the final checkpoints of the fine-tuned models are presented and compared to those of the GPT-2 baseline and the Swedish GPT-SW3 model, serving as a comparison between the AJT performance from Swedish L2 learners, a non-Swedish speaker and a native Swedish speaker, while also functioning as a sanity check of the SwePT evaluation dataset and the fine-tuning.

In section 4.2 we present the acquisition times of all structures in SwePT across all four fine-tuned models, which serves to determine whether the acquisition times are consistent across models. In section 4.3 the results from the implicational scaling are presented, where each checkpoint across all four fine-tuned models is treated as a separate learner, the purpose of which is to determine whether there is a consistent implicational relationship between the acquired structures from different developmental stages.

Lastly, in section 2.3.3 we go beyond PT and present and analyze the models' learning trajectories in relation to previous research. The subsequent section 4.5 provides an analysis of the impact from the training data distribution on the observed learning trajectories.

4.1. Model evaluation

Table 4.1 displays the final accuracies from the AJT on SwePT of all fine-tuned models. In addition to the fine-tuned models, we evaluated the pretrained English GPT-2 model (without fine-tuning) as well as the 126M parameter GPT-SW3[1] model. GPT-SW3 was pretrained on 320B tokens of text in Scandinavian languages, mainly Swedish, and thus functions as a skyline. GPT-SW3's average accuracy score 95.38% roughly aligns with the manually calculated precision score of 97.11% (see section 3.1.1).

Model	SVO	PLUR	TENSE	ATTR	PRED_a	PRED_b	INV	NEGV	INQ	Avg.
SW3	**98.57%**	**99.58%**	**95.10%**	**94.62%**	**90.61%**	**88.89%**	**93.06%**	**99.01%**	**98.94%**	**95.38%**
GPT-2	57.53%	10.86%	54.75%	26.85%	24.88%	40.74%	51.65%	38.94%	51.06%	39.70%
mixed	91.50%	63.26%	86.90%	74.96%	45.07%	44.44%	67.61%	70.63%	86.17%	70.06%
mixed_2	90.75%	64.93%	86.50%	75.00%	45.54%	48.15%	66.80%	71.95%	85.11%	70.52%
order	90.51%	63.26%	84.65%	74.07%	44.13%	44.44%	72.30%	95.38%	88.30%	73.00%
reverse	91.07%	62.21%	87.95%	50.13%	32.39%	44.44%	56.95%	28.71%	85.11%	59.89%

Table 4.1.: Results from evaluating the last checkpoints of all fine-tuned models, the Swedish GPT-SW3 model and the base English GPT-2 model on SwePT.

The accuracies across all structures and fine-tuned models, with the exception of NEGV in `GPT-reverse`, are higher than the English pretrained GPT-2 but lower than the GPT-SW3. This indicates that while the fine-tuning was successful, the 20M Swedish tokens in the finetuning data cannot compare in size to the 320B tokens that GPT-SW3 was trained on and is likely insufficient to reach maximum performance. Moreover, GPT-SW3 was trained from scratch on Swedish, while the pre-existing

[1] https://huggingface.co/AI-Sweden-Models/gpt-sw3-126m

knowledge in the GPT-2 models fine-tuned in the present study to some extent must be overwritten in order for the model to learn the Swedish data.

While the accuracy on PRED_a and PRED_b remained below chance level for all fine-tuned models at the end of training, the difference in performance on the two from the GPT-SW3 model indicates that the presence of attractors in the PRED_b dataset, which requires hierarchical generalization abilities, is not as refined as the models' heuristics. This should be interpreted with caution, however, giving the minimal amount of minimal pairs in the PRED_b subset.

The comparably low accuracy on NEGV in `GPT-reverse` can be explained by the fact that the only NEGV structures in the training data were introduced during the first 500 time steps of training, and likely forgotten throughout later training where post-negations dominate. This can be attributed to *catastrophic forgetting* (McCloskey and Cohen, 1989, p. 110), the observed phenomena that "[t]raining on a new set of items may drastically disrupt performance on previously learned items". The observation that `GPT-order` achieved the highest average accuracy across structures among all fine-tuned models strengthens this notion, as structures from earlier stages are included in the datasets introduced to the model later in training, thereby mitigating the effect of catastrophic forgetting.

4.2. Acquisition time

Acquisition time refers to the checkpoint (100th time step) where the accuracy on each structure emerges above a predetermined threshold of acquisition. As discussed in Section 3.3.2, three thresholds are set, namely at chance level (50%), 60% and 80% average accuracy. Table C.1 in Appendix C.1 displays the acquisition times across fine-tuned models and acquisition thresholds. The implicational order as hypothesized by PT can be inferred from the acquisition times of the `GPT-order` model, where at least one structure per stage is acquired before or simultaneously as structures from higher stages, with the exception of INQ. This is expected since the structures are exposed to the model in this order.

There is a noticeable variability in acquisition times across structures from the same stage, with PLUR emerging above the 50% and 60% acquisition thresholds more than 1000 time steps after SVO and TENSE, and never above the 80% threshold. Similarly, accuracy on INV quickly emerges above the 50% threshold, while the PRED structures on the same stage are acquired late or not at all. SVO and TENSE from the 2nd developmental stage are acquired early across models and thresholds, albeit not before NEGV on the 5th stage in the `GPT-reverse` model, which was initially trained on NEGV data. This implies that PT is not robust to the input order in the curriculum and support the results from the implicational scaling, which defies the predictions of the PT.

The rank correlation permutation test shows that the relative ordering of acquisition times is consistent across all four models.[2] The numbers are presented in Table C.2 in Appendix C. It should be noted that the rank correlation test only takes the rank order at the aforementioned acquisition thresholds into account, meaning that the rank order across the entire fine-tuning sequence is not reflected in these results. The ordering of the grammatical structures across the entire fine-tuning sequence can be read from the plots presented in Section 4.4.

[2] $p < 0.000$ for the 50% and 60% accuracy thresholds and $p < 0.01$ for the 80% accuracy threshold

4.3. Implicational patterns

Implicational scaling, as discussed in section 3.3.3, is traditionally used in PT studies to examine whether the observed acquisition order aligns with PT, and whether this order is implicational (i.e., structures from lower stages must be acquired before structures from higher stages). Table 4.2 presents collapsed implicational scales using acquisition thresholds at 50%, 60% and 80% accuracy. The columns represent the PT structures (9 in total), the rows represent the number of checkpoints sharing that particular acquisition order, and the cell values are coded as '+' (structure acquired) or '-' (structure not acquired). The columns (PT structures) are ordered according to their "overall rank order", i.e. from the most empirically difficult (the structure that has been acquired by the least learners) to the easiest (the structure that has been acquired by the most learners),[3]. The rows (the learners) are ordered from more advanced to less advanced in terms of how many PT structures they have acquired.

While the observed order differs slightly between the three scales, all observed patterns deviate from the predicted order as hypothesized by PT, since structures from later developmental stages emerge above the acquisition thresholds before structures from earlier developmental stages. For example, INQ (stage 5) emerges before PRED_a, PRED_b and PLUR in all scales.

There is also significant variability within each scale. The IR (index of reproducibility) coefficients across all three scales are far below 0.93 (the threshold of scalability Rickford, 2004). This implies that the observed order is not implicational.[4]

4.4. Learning trajectories

Figures 4.1, 4.2a and 4.2b display the acquisition trajectories of all linguistic structures tested through the AJT. The results from the `GPT-mixed_2` which follows a very similar pattern to the `GPT-mixed` model is presented in Figure C.1 in Appendix C.3.

Figures C.3 and C.2 in Appendix C.4 plot the absolute difference between the log-likelihood scores assigned to the grammatical vs. the ungrammatical sentence of each minimal pair from the AJT, thus displaying the "confidence" of the models in their classifications across checkpoints. While following similar envelopes as the learning curves, the confidence curves are smoother, indicating that confidence in grammatical distinctions improves more gradually and is less sensitive to noise or outliers, than raw loss.

`GPT-mixed`, which was trained without curriculum learning, displays the most consistent trajectory, with higher average accuracies compared to the two curriculum models. With the exception of PRED_b, which should be considered an outlier due to its minimal dataset size, the acquisition of all structures follow a visibly parallel acquisition pattern in an order that defies the predictions of PT. NEGV and INQ which are hypothesized to be the most difficult for the model to learn since they require the processing procedure at the 5th developmental stage, quickly reach above the 60% acquisition threshold before structures from the 4th and 3rd stages. The PRED subsets never reach above the 60% threshold, but rather decrease in accuracy throughout

[3] In some PT studies (e.g., Spinner, 2013) the overall rank order is based on the order predicted by PT instead, where structures that belong to the same developmental stage are grouped together in the same category. However, due to the experimental nature of our study where the PT structures cannot be assumed but only hypothesized to fall within the same developmental stages, we use the standard approach previously described in favor of a more fine-grained analysis.

[4] According to standard practice when using the PT framework, IR must be interpreted with consideration to the fact that not all structures of each stage must be emerged in order to consider that stage as acquired. However, this must be considered less relevant in the present study due to the large amount of data points in the evaluation dataset and the large number of checkpoints tested.

n	PRED_b	PRED_a	NEGV	PLUR	ATTR	INV	INQ	SVO	TENSE
4	-	+	+	+	+	+	+	+	+
235	-	-	+	+	+	+	+	+	+
21	-	+	-	+	+	+	+	+	+
3	-	+	+	-	+	+	+	+	+
99	-	-	-	+	+	+	+	+	+
96	-	-	+	-	+	+	+	+	+
11	-	+	-	+	+	-	+	+	+
15	-	-	-	-	+	+	+	+	+
26	-	-	-	+	-	+	+	+	+
15	-	-	-	+	+	-	+	+	+
4	-	-	+	-	-	+	+	+	+
1	-	-	+	-	+	+	-	+	+
49	-	-	-	-	-	+	+	+	+
6	-	-	-	-	+	+	-	+	+
30	-	-	-	-	-	-	+	+	+
5	-	-	-	-	-	+	-	+	+

(a) Acquisition threshold: 50%. Index of reproducibility (IR) = 0.393

n	PRED_a	PRED_b	PLUR	NEGV	INV	ATTR	INQ	TENSE	SVO
30	-	-	+	+	+	+	+	+	+
194	-	-	-	+	+	+	+	+	+
3	-	-	+	-	+	+	+	+	+
81	-	-	-	-	+	+	+	+	+
32	-	-	-	+	-	+	+	+	+
17	-	-	-	+	+	+	-	+	+
58	-	-	-	-	-	+	+	+	+
12	-	-	-	-	+	+	-	+	+
3	-	-	-	+	-	+	-	+	+
1	-	-	-	+	+	-	-	+	+
131	-	-	-	-	-	-	+	+	+
19	-	-	-	-	-	+	-	+	+
5	-	-	-	+	-	-	-	+	+
28	-	-	-	-	-	-	-	+	+
2	-	-	-	+	-	-	-	-	+
4	-	-	-	-	-	-	-	-	+

(b) Acquisition threshold: 60%. Index of reproducibility (IR) = 0.292

n	PLUR	PRED_a	PRED_b	INV	NEGV	INQ	ATTR	TENSE	SVO
3	-	-	-	-	+	+	-	+	+
12	-	-	-	-	-	+	-	+	+
2	-	-	-	-	+	-	-	+	+
254	-	-	-	-	-	-	-	+	+
41	-	-	-	-	-	-	+	-	+
195	-	-	-	-	-	-	-	-	+
7	-	-	-	-	+	-	-	-	-
106	-	-	-	-	-	-	-	-	-

(c) Acquisition threshold: 80%. Index of reproducibility (IR) = 0.697

Table 4.2.: Implicational scales across all checkpoints from the four models evaluated on SwePT, based on acquisition times calculated at three different acquisition thresholds. A '+' mark indicates that that specific structure is acquired. A '-' mark indicates that the structure is not acquired. The scales are collapsed, meaning that checkpoints with identical acquisition order are counted together in the same row. *n* denotes this number of checkpoints. The structures are ordered from left to right, with the structure that is learned at the most checkpoints to the left, and the structure learned by the least checkpoints to the right.

training. Notably, in all models SVO, TENSE, INQ and INV have already reached an above-chance accuracy at the first checkpoint.

`GPT-order` and `GPT-reverse` were trained on two different curricula, with `GPT-order` being exposed to the data in an order of increasing difficulty, and `GPT-reverse` in an order of decreasing difficulty. The accuracy across checkpoints is plotted in Figures 4.2a and 4.2b. There is a visible correlation between the accuracy and the proportion of a structure in the training data, which are displayed in Table 3.3. The most prominent example is NEGV, which is not processable before procedures of earlier stages have been acquired, but yet reaches an accuracy above 90% during the the `GPT-reverse` model's first 5 checkpoints (Table 4.2b), where the model is trained on the Stage-5 dataset. This accuracy declines steadily throughout later training steps, ending on a confident (Table C.2b) 30% accuracy at the last checkpoint. This performance is mirrored in the `GPT-order` model, where accuracy increases drastically during the last 5 checkpoints where the Stage-5 dataset is introduced. The performance on ATTR follows a similar pattern, with a clear increase in accuracy between checkpoints 81-106 in `GPT-order` and checkpoints 53-78 in `GPT-reverse`.

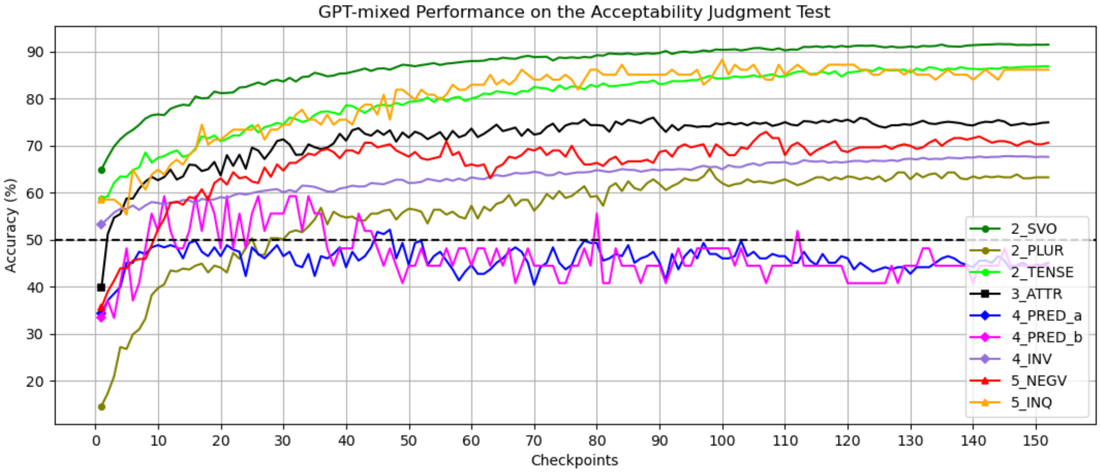
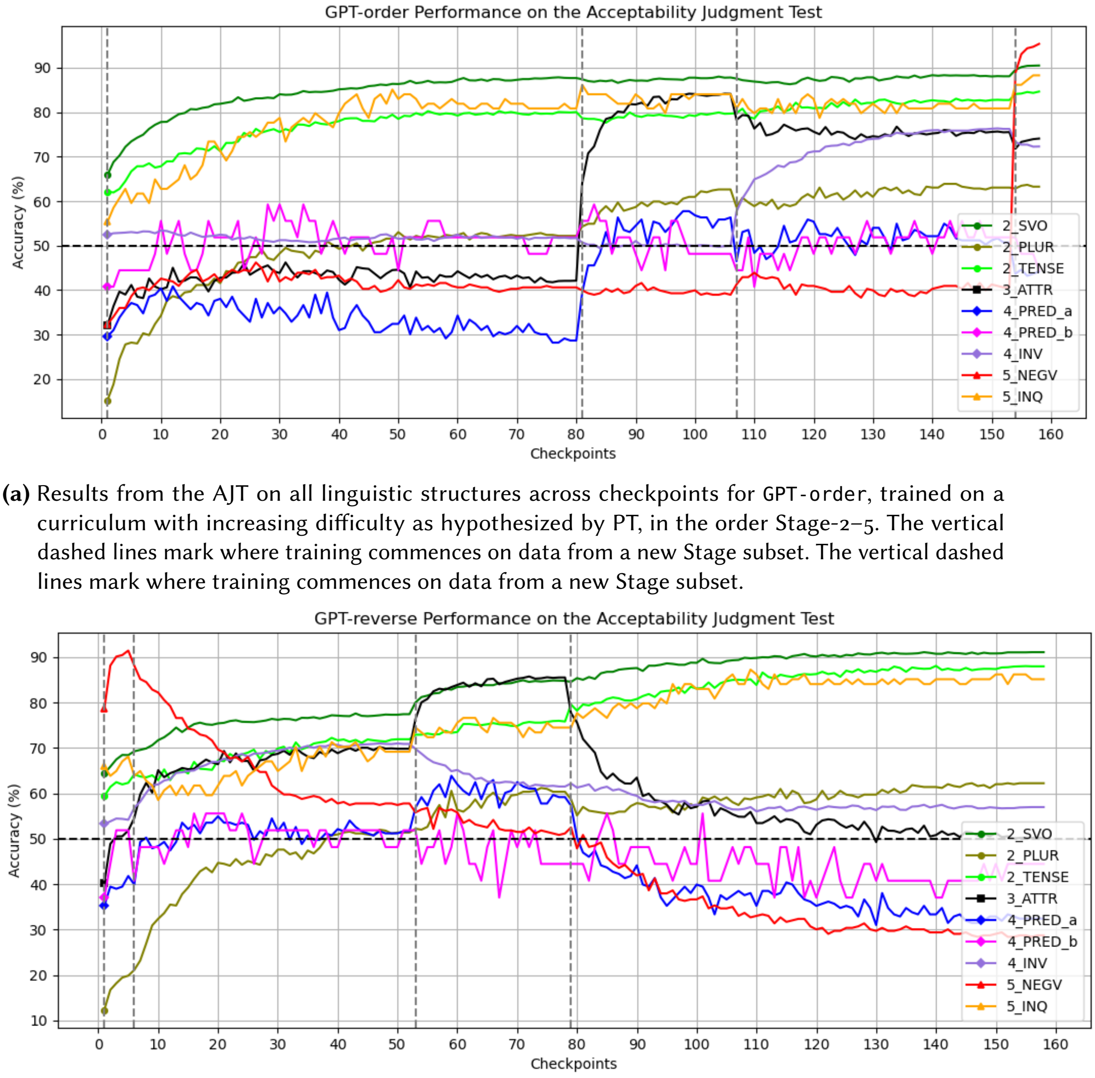


Figure 4.1.: Results from the AJT on all linguistic structures across checkpoints for the `GPT-mixed` model trained on all four data subsets concatenated and shuffled. The acquisition trajectories follow a parallel pattern, and the acquisition order deviates from that predicted by PT.

4.5. Effects from curriculum learning

Although the acquisition order of the linguistic structures does not follow the trajectory as hypothesized by PT, the observed trajectories were found to be systematic and may be explained by additional theories from previous research, as presented in section 2.3.3. The finding that LLMs behave similarly to bag-of-word models during initial training stages (Choshen et al., 2021) provide one possible explanation to the trajectories across morphological structures during the initial training stages observed in the present study. In order to examine this, we calculated the distribution of the contrasting morphological features in the grammatical versus ungrammatical sentences in the minimal pairs of SwePT, by searching the CoNLL-u training data (dependency trees) using simple rule-based scripts. 72% of all attributive adjectives in the training data were in common (lemma) form, 27% in neuter and 0.92% in invariant form (gender-agnostic). Among all predicative adjectives, 65% were in common (lemma) form, 35% in neuter and 0.12% in invariant form. Relevant to the PLUR subset, 79% of non-genitive



(a) Results from the AJT on all linguistic structures across checkpoints for `GPT-order`, trained on a curriculum with increasing difficulty as hypothesized by PT, in the order Stage-2–5. The vertical dashed lines mark where training commences on data from a new Stage subset. The vertical dashed lines mark where training commences on data from a new Stage subset.

(b) Results from the AJT on all linguistic structures across checkpoints for `GPT-reverse`, trained on a curriculum with decreasing difficulty as hypothesized by PT, in the order Stage-5–2. The vertical dashed lines mark where training commences on data from a new Stage subset.

Figure 4.2.: Results from the AJT on all linguistic structures across checkpoints for the models trained with curriculum learning. The vertical dashed lines mark where training commences on data from a new Stage subset.

nouns (not counting abbreviations) were in singular (lemma) form, and 21% in plural form. Relevant to the TENSE subset, 28% of verbs were in infinitive (lemma) form, and 72% in tensed form.

Considering this distribution, the models are seemingly favoring the sentence of each minimal pair that contains the more frequent form of the target structure. The initial 10% accuracy of the PLUR structure roughly correlates with the 80-20 ratio between singular (lemma) and plural nouns in the training data. A similar pattern is detected for the initial below-chance accuracy of ATTR, since attributive adjectives in common form (lemma) are in majority in the training data. For TENSE, the initial *high* accuracy aligns with the observation that tensed verbs are more frequent than their respective infinitive forms (lemma) in the training data. As training progresses, the accuracy curves of PLUR and ATTR in all models rise quickly, suggesting that the classifications deviate more and more from the observed unigram distribution. This is in line with previous research indicating that LLMs in later learning stages start to resemble n-gram models that are sensitive to word order, and eventually start relying more on structural cues in context (Choshen et al., 2021; Saphra and Lopez, 2018).

Interestingly, the accuracy on the PRED subsets stays around chance-level in the mixed model, while performance on PRED_a fluctuates in predictable patterns in the `GPT-order` and `GPT-reverse` models with regards to the training data. The minimal pairs for PRED_a were constructed using the same principles as for ATTR, where neuter gendered or plural form of the adjective is contrasted to the common, singular form of the same adjective in the ungrammatical sentence of the pair. This implies that the grammatical sentence in the PRED_a subset will be object to the same frequency-based bias as in the ATTR subset, as previously discussed, thus favoring the ungrammatical version of each pair. While the PRED_b subset is constituted by a significantly smaller amount of minimal pairs than its PRED_a counterpart, it can still be speculated that the lack of ATTR structures in the Stage-2 dataset which clearly affects the performance on PRED_a (see the first portion of Figure 4.2a and the last portion of Figure 4.2b), does not have the same impact on the PRED_b subset that was constructed manually without a bias toward common adjectives (lemmas) in the ungrammatical sentences. The fact that accuracy does not improve throughout training for the PRED subsets may suggest that the models have not yet reached the stages of training where they are found to rely on more global cues. Possibly, such strategies include the hierarchical generalizations required for agreement beyond the NP, which corresponds to the procedures at the fourth developmental stage of the PT hierarchy.

There is a notable difference between the hypothesized difficulty of non-inversion in indirect questions and the consistently high performance on INQ across all fine-tuned models. The consistent high performance throughout training in all models, despite the INQ structures as labeled after parsing only occurring in the Stage-5 subset, can be explained by the fact that interrogative subclauses are syntactically identical to subclauses of declarative sentences (see further explanation of this structure in Section 2.2.2.). Subclauses as a linguistic category is not mutually exclusive with any of the PT structures, meaning that subclauses are likely evenly distributed across the Stage subsets in the training data (save for the addition of NEGV in Stage-5), allowing the models to accurately generalize the word order in interrogative subclauses. The performance on INQ is very similar to the stage 2 structures SVO and TENSE. This indicates that while humans are hypothesized within PT to struggle with the non-inversion in indirect questions due to overgeneralization of direct question word order (see Section 2.2.2), the GPT-2 models do not seem to rely on the semantic similarities between direct and indirect questions, but instead focus on their syntactic features, regardless of whether they are encoded hierarchically or heuristically. In light of previous research indicating that semantic information is learned later than e.g. syntactic information (Blevins et al., 2022; Saphra and Lopez, 2018), it is possible that the models assessed during a single epoch of fine-tuning do not yet have access to the generalization strategies that allow humans to focus on salient semantic patterns, although the models' n-gram-like strategies seemingly happen to be more effective for the case that INQ represents.

5. Discussion

There are several explanations to the observed violations of the PT hypotheses. The first explanation is that the hypotheses of PT are wrong altogether. This is unlikely, considering its wide empirical support across languages from numerous studies (e.g. Eklund Heinonen, 2009; Glahn et al., 2001; Håkansson and Norrby, 2006; Pienemann and Håkansson, 1999). Assuming that PT does hold for human L2 learners, a second explanation to the results is that neural language models such as GPT-2 do not possess the same processing constraints as humans do, or that they apply different learning strategies. It is possible that such constraints are induced later in training, as suggested by the previous research presented in Section 2.3.3.

A third explanation is that PT truly cannot be tested through receptive skills, as discussed by Pienemann (1998b) and Dyson and Håkansson (2017)[1]. Alternatively, if the relevant distinction is not between production and reception but rather between implicit and explicit knowledge, as argued by R. Ellis (2008)[2], it may be the case that the "cognitive processes" behind GPT-2's next-word prediction, that guide its classifications and text generation, align more with the "explicit knowledge" utilized by humans during e.g. an untimed AJT than with "implicit knowledge" used for timed AJT and speech production. Since the AJT performance partly could be explained on the basis of one-gram frequencies, and since there was little difference in performance (beyond what was explainable by the data) between PRED_a and PRED_b despite the presence of attractors in the latter, one could speculate that the window of time during which the models were evaluated in this study only reflects performance from the 2nd developmental stage, which does not reject the hypothesis that the models enter further developmental stages later in training. The observed high accuracy on the structures from the 3rd, 4th and 5th developmental stages could then be explained through surface-level heuristics, which could be argued to require only the lemma and category procedures available at the 1st and 2nd developmental stages. The finding from previous research that LLMs grow more similar to humans in terms of performance in later learning stages (Choshen et al., 2021) supports the speculation that alignment with PT order would appear late in training. Testing such a claim would only be possible if an alternative method that allows the emergence criterion to be applied according to standard PT practice was found, and if the fine-tuning was extended with more data across more epochs.

5.1. Limitations and future work

Adapting the emergence criterion

As authors of similar PT studies focusing on receptive skills[3] have previously addressed, the emergence criterion cannot be seamlessly adapted to align with the evaluation of the acceptability classification task. A fundamental conceptual distinction remains, which this study fails to resolve: the fact that accuracy and emergence reflect separate aspects of acquisition. Since the accuracy rates of different structures develop with

[1] See section 2.2.3

[2] See section 2.2.3

[3] See Section 2.2.3

different gradients, the inferred acquisition order is sensitive to the predefined acquisition threshold, which reduces reliability. While the plotted learning trajectories in this study offer transparency regarding this aspect, by modeling the change in accuracy across time, the gradient property in and of itself does not align with the categorical logic of the emergence criterion. This has a significant impact on the validity of the results as evaluated within the framework of PT. This methodological issue requires further addressing.

Linked to the difficulties of evaluating receptive knowledge within PT is the limitation of using minimal pairs to test the order of acquisition within the PT framework. Evidence of rule application in its obligatory context does not only encompass grammatically correct target-like applications, but any application that can demonstrate that the learner is able to process that grammatical rule. For example, the past tensed form of the Swedish irregular verb "gå" (*walk*) is "gick", while a majority of regular verbs are realized in past tense with the *-de* suffix. Thus, the interlanguage form "*gådde" often emerge in the speech output of Swedish L2 learners that have access to the processing procedure on the 2nd developmental stage, and serves as evidence for application of the grammatical rule even though the surface form is incorrect (Flyman Mattsson, 2022). By limiting the GPT-2 models to the binary classification of a minimal pair, valuable information from non-target constructions may be lost (see e.g. the study by Schönström, 2014).

Future work could avoid these methodological issues by focusing on the language *production* of LLMs. By examining the generative text output at different checkpoints throughout training, the emergence criterion could be implemented without the need for adaptation to accuracy scores, and any "interlanguage" constructions of the target grammatical rule, if applicable, could be taken into account, offering insights beyond the limitations of the minimal pairs in the present study.

Minimal pair dataset generation

Although the manually calculated error rate and the high performance of GPT-SW3 on SwePT both indicate that the minimal pairs offered adequate quality for the purposes of this study, there are several concerns and areas of improvement related to the dataset curation process. Firstly, the observed number of false positives could be reduced by utilizing native-speaker crowdsourcing to validate the output. An even more effective solution could be to use a Swedish language model such as GPT-SW3 for evaluation of the minimal pairs. Aside from the concern of false positives, the filtering constraints in the scripts used to identify and alter the Talbanken and LinES sentences may also filter out false negatives. This can be mitigated in future work by allowing expert linguists to validate the scripts.

Secondly, the sentences of SwePT were not normalized for length, token frequencies or additional metrics of complexity. In contrast to the previous studies examining developmental stages of human receptive linguistic skill acquisition within the PT framework (presented in section 2.2.3), the minimal pairs in SwePT used for evaluation in the present study were curated using automatically applied rule-based heuristics, in favor of increasing replicability by generating on a large amount of pairs. The minimal pairs of SwePT were identified and altered on the basis of a single metric of complexity, namely the presupposed processing constraints of its target linguistic structure. Each sentence across each subset is thus assumed to be equally as complex. This has several implications.

As discussed in section 2.3.2, token frequency and sentence length are confounding factors to the acceptability of a sentence as predicted by probability. While sentence length does not have an impact on the difference in score between sentences of a single minimal pair that are equal in length (in terms of the number of words per sentence) beyond the distribution between different morphological forms, it may impact the overall processability and make the acceptability scores less reliable, as may differences in parse-depth across pairs. Within pairs, the difference in semantic plausibility after altering the grammatical sentence is not taken into account. Manual curation or refinement of the scripts where all of the aforementioned aspects are controlled for may increase the interpretability of results from tasks such as AJT.

As the performance on the morphological categories of SwePT were found to align with the frequencies of the target structures in either sentence of the minimal pair, the constraints used for the curation of these structures should allow more variety of obligatory context in order to not tie the performance on the AJT to a single heuristic. For example, the TENSE subset may have elicited different performance scores on the AJT if the verb forms were varied across the pairs, and if the obligatory context for a specific tense was based on semantic or syntactic cues rather than context-independent occurrences.

Parsing evaluation

It is possible that the results from the AJT may have been affected by poor fine-tuning data. While the quality of large common crawl datasets such as OSCAR is hard to control for, in future work, it is recommended to evaluate the fine-tuning data parsing process more rigorously in order to quantify the noise. KL divergence may be a better choice than the chi-square test when quantifying how much the parsed fine-tuning data diverges from the gold-standard distributions. Processing in earlier steps, including the splitting of sentences before parsing, should also be evaluated systematically to ensure its precision and avoid propagating errors further down the pipeline.

Fine-tuning

In terms of the fine-tuning, a significant limitation to this study is the fact that accuracy of many PT structures seemingly emerged above the chance-level mark before the first checkpoint. The loss of valuable information thereof could be mitigated in future research by saving checkpoints at smaller intervals, possibly in logarithmic intervals since the learning curves even out later in training (see e.g., Bunzeck and Zarrieß, 2024).

With additional computing resources, a model such as GPT-SW3 could also be pretrained and evaluated throughout, providing a comparison between the acquisition order of a fine-tuned "L2 learner" and itself as an "L1 learner". Such a study could offer additional insights on transfer effects and the modeling of first and second language acquisition.

6. Conclusion

The present study examined whether GPT-2's acquisition order of Swedish grammatical structures follows the order sequence as stipulated by Processability Theory, and to what extent this acquisition order is affected by the input sequencing of the training data (i.e., different curricula). The results indicated that while the observed acquisition order was found to be robust to the order sequencing of the training data as measured with rank correlation tests at the thresholds of acquisition defined in this study, the acquisition order of the fine-tuned models did not align with the implicational order sequence as hypothesized by PT. Observations of the performance on the AJT and the frequency distribution of the contrasting features in the minimal pairs suggested that the performance largely can be explained by unigram and n-gram heuristics. These findings suggest that the grammatical development predicted by PT does not naturally emerge from next-word prediction objectives. These results should be interpreted with caution, however, due to inherent incompatibilities between the PT framework and the methodology required for testing grammatical receptive skills with acceptability judgment tests.

Bibliography

Adger, David (2003). *Core syntax: A minimalist approach.* Oxford University Press.

Altmann, Gerry TM and Yuki Kamide (1999). “Incremental interpretation at verbs: Restricting the domain of subsequent reference”. *Cognition* 73.3, pp. 247–264.

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston (2009). “Curriculum learning”. In: *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48.

Blevins, Terra, Hila Gonen, and Luke Zettlemoyer (2022). “Analyzing the mono-and cross-lingual pretraining dynamics of multilingual language models”. *arXiv preprint arXiv:2205.11758.*

Brown, Roger (1973). *A first language: The early stages.* Harvard University Press.

Bunzeck, Bastian and Sina Zarrieß (2024). “Fifty shapes of BLiMP: syntactic learning curves in language models are not uniform, but sometimes unruly”. In: *Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning*, pp. 39–55.

Buyl, Aafke and Alex Housen (2015). “Developmental stages in receptive grammar acquisition: A Processability Theory account”. *Second Language Research* 31.4, pp. 523–550.

Chater, Nick and Christopher D Manning (2006). “Probabilistic models of language processing and acquisition”. *Trends in cognitive sciences* 10.7, pp. 335–344.

Chiang, Cheng-Han, Sung-Feng Huang, and Hung-yi Lee (2020). “Pretrained language model embryology: The birth of ALBERT”. *arXiv preprint arXiv:2010.02480.*

Chomsky, Noam (1957). *Syntactic Structures.* The Hague: Mouton and Co.

Chomsky, Noam (1980). “Rules and representations”. *Behavioral and brain sciences* 3.1, pp. 1–15.

Choshen, Leshem, Guy Hacohen, Daphna Weinshall, and Omri Abend (2021). “The grammar-learning trajectories of neural language models”. *arXiv preprint arXiv:2109.06096.*

Crocker, Matthew W and Steffan Corley (2008). “Modular architectures and statistical mechanisms: The case from lexical category disambiguation”. *The Lexical Basis of Sentence Processing: Formal, computational and experimental issues*, pp. 157–180.

Crocker, Matthew W and Frank Keller (2006). “Probabilistic grammars as models of gradience in language processing”. *Gradience in grammar: Generative perspectives*, pp. 227–245.

De Marneffe, Marie-Catherine, Christopher D Manning, Joakim Nivre, and Daniel Zeman (2021). “Universal dependencies”. *Computational linguistics* 47.2, pp. 255–308.

Dyson, Bronwen Patricia and Gisela Håkansson (2017). *Understanding second language processing: A focus on Processability Theory.* Vol. 4. John Benjamins Publishing Company.

Eklund Heinonen, Maria (2009). “Processbarhet på prov: Bedömning av muntlig språkfärdighet hos vuxna andraspråksinlärare”. PhD thesis. Universitetsbiblioteket.

Ellis, Nick C (2002). “Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition”. *Studies in second language acquisition* 24.2, pp. 143–188.

Ellis, Rod (2005). “Measuring implicit and explicit knowledge of a second language: A psychometric study”. *Studies in second language acquisition* 27.2, pp. 141–172.

Ellis, Rod (2008). "Investigating grammatical difficulty in second language learning: Implications for second language acquisition research and language testing". *International journal of applied linguistics* 18.1, pp. 4–22.

Ellis, Rod and Shawn Loewen (2007). "Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): Responding to Isemonger". *Studies in second language acquisition* 29.1, pp. 119–126.

Evanson, Linnea, Yair Lakretz, and Jean-Rémi King (2023). "Language acquisition: do children and language models follow similar learning stages?" *arXiv preprint arXiv:2306.03586*.

Fine, Alex B, T Florian Jaeger, Thomas A Farmer, and Ting Qian (2013). "Rapid expectation adaptation during syntactic comprehension". *PloS one* 8.10, e77661.

Flyman Mattsson, Anna (2022). "Rethinking textbook grammar introduction". *Instructed Second Language Acquisition* 6.2, pp. 196–218.

Fossum, Victoria and Roger Levy (2012). "Sequential vs. hierarchical syntactic models of human incremental sentence processing". In: *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics (CMCL 2012)*, pp. 61–69.

Frank, Stefan L and Rens Bod (2011). "Insensitivity of the human sentence-processing system to hierarchical structure". *Psychological science* 22.6, pp. 829–834.

Frank, Stefan L, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco (2013). "Reading time data for evaluating broad-coverage models of English sentence processing". *Behavior research methods* 45, pp. 1182–1190.

Gibbons, Jean Dickinson and Subhabrata Chakraborti (2014). *Nonparametric statistical inference: revised and expanded*. CRC press.

Glahn, Esther, Gisela Håkansson, Björn Hammarberg, Anne Holmen, Anne Hvenekilde, and Karen Lund (2001). "Processability in Scandinavian second language acquisition". *Studies in second language acquisition* 23.3, pp. 389–416.

Goodkind, Adam and Klinton Bicknell (2018). "Predictive power of word surprisal for reading times is a linear function of language model quality". In: *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pp. 10–18.

Grosjean, François (1980). "Spoken word recognition processes and the gating paradigm". *Perception & psychophysics* 28.4, pp. 267–283.

Håkansson, Gisela and Catrin Norrby (2006). "Processability Theory applied to written and spoken L2 Swedish". English. In: *Second language acquisition research: theory-construction and testing*. Ed. by Fethi Mansouri. The information about affiliations in this record was updated in December 2015. The record was previously connected to the following departments: Linguistics and Phonetics (015010003). United Kingdom: Cambridge Scholars Publishing, pp. 81–94. ISBN: 1-84718-051-5.

Håkansson, Gisela and Catrin Norrby (2010). "Environmental influence on language acquisition: Comparing second and foreign language acquisition of Swedish". *Language learning* 60.3, pp. 628–650.

Hale, John (2001). "A probabilistic Earley parser as a psycholinguistic model". In: *Second meeting of the north american chapter of the association for computational linguistics*.

Han, Youngju and Rod Ellis (1998). "Implicit knowledge, explicit knowledge and general language proficiency". *Language teaching research* 2.1, pp. 1–23.

Hewitt, John and Christopher D Manning (2019). "A structural probe for finding syntax in word representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138.

Josephson, Olle (2020). *Grammatik, ord, texttyper: Svenska med fokus på form*.

Kaplan, Ronald M, Joan Bresnan, et al. (1981). *Lexical-functional grammar: A formal system for grammatical representation.* Massachusetts Institute Of Technology, Center For Cognitive Science.

Kawaguchi, Satomi (2008). "Argument structure and syntactic development in Japanese as a second language". In: *Cross-linguistic aspects of processability theory.* John Benjamins Publishing Company, pp. 253–298.

Keatinge, Dagmar and Jörg-U Keßler (2009). "The acquisition of the passive voice in L2 English: Perception and production". *Research in second language acquisition: Empirical evidence across languages*, pp. 67–92.

Kuhl, Patricia K (2004). "Early language acquisition: cracking the speech code". *Nature reviews neuroscience* 5.11, pp. 831–843.

Kuribayashi, Tatsuki, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin (2025). "Large Language Models Are Human-Like Internally". *arXiv preprint arXiv:2502.01615*.

Lau, Jey Han, Alexander Clark, and Shalom Lappin (2017). "Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge". *Cognitive science* 41.5, pp. 1202–1241.

Levelt, Willem (1989). "Speaking-From Intention to Articulation". *A Bradford book.*

Levy, Roger (2008). "Expectation-based syntactic comprehension". *Cognition* 106.3, pp. 1126–1177.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg (2016). "Assessing the ability of LSTMs to learn syntax-sensitive dependencies". *Transactions of the Association for Computational Linguistics* 4, pp. 521–535.

Liu, Leo Z, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith (2021). "Probing across time: What does RoBERTa know and when?" *arXiv preprint arXiv:2104.07885*.

Luka, Barbara J and Lawrence W Barsalou (2005). "Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension". *Journal of Memory and Language* 52.3, pp. 436–459.

Manning, Christopher D, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy (2020). "Emergent linguistic structure in artificial neural networks trained by self-supervision". *Proceedings of the National Academy of Sciences* 117.48, pp. 30046–30054.

Mansouri, Fethi (2008). "Agreement morphology in Arabic as a second language: Typological features and their processing implications". In: *Cross-linguistic aspects of Processability Theory.* John Benjamins Publishing Company, pp. 117–153.

McCloskey, Michael and Neal J Cohen (1989). "Catastrophic interference in connectionist networks: The sequential learning problem". In: *Psychology of learning and motivation.* Vol. 24. Elsevier, pp. 109–165.

McCoy, R Thomas, Robert Frank, and Tal Linzen (2018). "Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks". *arXiv preprint arXiv:1802.09091*.

McCoy, R. Thomas, Robert Frank, and Tal Linzen (2020). "Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks". *Transactions of the Association for Computational Linguistics* 8. Ed. by Mark Johnson, Brian Roark, and Ani Nenkova, pp. 125–140. DOI: 10.1162/tacl_a_00304. URL: https://aclanthology.org/2020.tacl-1.9.

Mueller, Aaron, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster (2022). "Coloring the Blank Slate: Pre-training Imparts a Hierarchical Inductive Bias to Sequence-to-sequence Models". In: *Findings of the Association for Computational Linguistics: ACL 2022.* Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1352–1368.

DOI: 10.18653/v1/2022.findings-acl.106. URL: https://aclanthology.org/2022.findings-acl.106.

Norrby, Catrin and Gisela Håkansson (2007). "The interaction of complexity and grammatical processability: The case of Swedish as a foreign language". *Iral-international Review of Applied Linguistics in Language Teaching - IRAL-INT REV APPL LINGUIST* 45 (Jan. 2007), pp. 45–68. DOI: 10.1515/IRAL.2007.002.

Oh, Byung-Doh, Christian Clark, and William Schuler (2022). "Comparison of structural parsers and neural language models as surprisal estimators". *Frontiers in Artificial Intelligence* 5, p. 777963.

Pienemann, Manfred (1998a). "Developmental dynamics in L1 and L2 acquisition: Processability theory and generative entrenchment". *Bilingualism: Language and cognition* 1.1, pp. 1–20.

Pienemann, Manfred (1998b). *Language processing and second language development: Processability theory*. Vol. 15. John Benjamins Publishing.

Pienemann, Manfred (2005). *Cross-linguistic aspects of processability theory*. John benjamins publishing company.

Pienemann, Manfred and Gisela Håkansson (1999). "A unified approach toward the development of swedish as L2: A processability account". *Studies in second language acquisition* 21.3, pp. 383–420.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning (2020). "Stanza: A Python natural language processing toolkit for many human languages". *arXiv preprint arXiv:2003.07082*.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). "Language models are unsupervised multitask learners". *OpenAI blog* 1.8, p. 9.

Rickford, John R (2004). "Implicational scales". *The handbook of language variation and change*, pp. 142–167.

Saphra, Naomi and Adam Lopez (2018). "Understanding learning dynamics of language models with SVCCA". *arXiv preprint arXiv:1811.00225*.

Schönström, Krister (2014). "Visual acquisition of Swedish in deaf children: An L2 processability approach". *Linguistic approaches to bilingualism* 4.1, pp. 61–88.

Smith, Nathaniel J and Roger Levy (2013). "The effect of word predictability on reading time is logarithmic". *Cognition* 128.3, pp. 302–319.

Spinner, Patti (2013). "Language production and reception: A processability theory study". *Language Learning* 63.4, pp. 704–739.

Taktasheva, Ekaterina, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov (2024). "RuBLiMP: Russian benchmark of linguistic minimal pairs". *arXiv preprint arXiv:2406.19232*.

Tanenhaus, Michael K, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy (1995). "Integration of visual and linguistic information in spoken language comprehension". *Science* 268.5217, pp. 1632–1634.

Tomasello, Michael (2005). *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Tsvetkov, Yulia, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer (2016). "Learning the curriculum with bayesian optimization for task-specific word representation learning". *arXiv preprint arXiv:1605.03852*.

Volodina, Elena, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, et al. (2019). "The SweLL language learner corpus: From design to annotation". *Northern European Journal of Language Technology (NEJLT)* 6, pp. 67–104.

Volodina, Elena, Yousuf Ali Mohammed, Aleksandrs Berdičevskis, Gerlof Bouma, and Joey Öhman (2023). "DaLAJ-GED-a dataset for Grammatical Error Detection tasks on Swedish". In: *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pp. 94–101.

Volodina, Elena, Yousuf Ali Mohammed, and Julia Klezl (2021). "DaLAJ-a dataset for linguistic acceptability judgments for Swedish: Format, baseline, sharing". *arXiv preprint arXiv:2105.06681.*

Wang, Xiaojing (2011). "Grammatical development among Chinese L2 learners: From a processability account". PhD thesis. Newcastle University.

Warstadt, A (2019). "Neural Network Acceptability Judgments". *arXiv preprint arXiv:1805.12471.*

Warstadt, Alex, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, et al. (2023). "Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora". In: *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman (2020). "BLiMP: The benchmark of linguistic minimal pairs for English". *Transactions of the Association for Computational Linguistics* 8, pp. 377–392.

Wilcox, Ethan, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy (2020). "On the predictive power of neural language models for human real-time comprehension behavior". *arXiv preprint arXiv:2006.01912.*

Yang, Charles (2016). *The price of linguistic productivity: How children learn to break the rules of language.* MIT press.

Yedetore, Aditya, Tal Linzen, Robert Frank, and R Thomas McCoy (2023). "How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech". *arXiv preprint arXiv:2301.11462.*

Appendices

A. Identification and perturbation of grammatical structures

A.1. Canonical word order (SVO)

The selection of sentences for the SVO subset was made based on the following constraints:

1. The first token of the sentence (or second in case of initial quotation marks) must not be a relative pronoun, an interrogative pronoun, a subjunction or a verb.
2. The sentence must contain a subject which must be a noun, a proper noun, a pronoun or a determiner.
3. The subject must be governed by a root verb.
4. The subject must precede the root.[1]
5. The sentence must contain an object or clausal complement which must be a dependent of the root.

To form the ungrammatical sentence for the minimal pair, all dependents of the subject and object phrases were identified, and the object phrase was positioned in front of the finite verb in the ungrammatical sentence, forming *SOV word order.

A.2. Plural (PLUR)

The selection of sentences for the PLUR subset was made based on the following constraints:

1. The sentence must contain a numeric modifier of which the lemma is not "1", "en" or "ett".
2. The sentence must contain a plural noun that governs the numeral.
3. The noun must not be marked with the genitive case.[2]
4. The lowercased form of the noun must not be the same as the noun's lemma.[3]
5. The noun must not be an abbreviation.[4]

Note that the constraints in the script allow plural nouns that occur in definite form, e.g. "de fem nedersta trappstegen" (*the five lowest steps*). In the PT hierarchy, definite form and plural are found on the same developmental stage.

To form the ungrammatical version of the minimal pair, the noun in each sentence was modified into its lemma, i.e. into its singular form.

[1] This ensures that no sentences with topicalization or subjunctive or imperative root verbs are included.

[2] Genitive in Swedish is marked with an "s" at the tail of the word and does not change other inflections of the noun. Thus, this is not an obligatory constraint, but simply serves to decrease the number of matches and simplify the subset.

[3] This excludes cases where there is no distinction between the singular and plural form, e.g. "ett hus, två hus" (*one house, two houses*).

[4] This excludes cases such as "kr" (the abbreviation of the Swedish currency *kronor*) which makes no distinction between the singular and plural but where the form differs from the lemma "kronor".

A.3. Tense (TENSE)

The selection of sentences for the TENSE subset was made based on the following constraints:

1. The sentence must contain a tensed verb or an auxiliary.
2. The verb must be in active form.
3. The verb form must be distinct from its lemma.

To form the ungrammatical version of the minimal pair, the verb in each sentence was modified into its lemma, which corresponds to its infinitive form.

A.4. Attributive agreement (ATTR)

The selection of sentences for the ATTR subset was made based on the following constraints:

1. The sentence must contain an adjectival modifier.
2. The adjective must be in positive and indefinite form.
3. If in singular form, the adjective must not be in common/utrum gender, and the form must be distinct from its lemma.

To form the ungrammatical version of the minimal pair, the adjective in each sentence was modified into its lemma (and capitalized when the form was capitalized), which corresponds to its common and singular form.

A.5. Predicative agreement (PRED)

The PRED subset is separated into PRED_a and PRED_b, where the former is created by altering the predicative adjective in the ungrammatical sentence to correspond to its lemma, and the latter includes an attractor noun that differs from the subject in number or gender, and that is linearly closer to the predicative adjective than the subject.

The sentences for PRED_a were selected based on the following constraints:

1. The subject must be a noun.
2. The sentence must contain a copula.
3. The predicate agreement must occur between the subject and an adjective in positive form that governs the subject.
4. The subject must precede the copula.
5. The lowercase form of the adjective must differ from its lemma.

For PRED_b, the sentence should contain a second noun that operates as an attractor. The following constraints were applied:

1. The noun subject must be governed by a root adjective,
2. The sentence must contain a second noun (attractor) which is governed by the subject,
3. The attractor must have the noun modifier relation.

In order to ensure that the adjective explicitly agrees with the subject and can be modified into a form that agrees with the attractor, sentences are excluded if...

1. the adjective inflection makes no distinction between neither gender nor grammatical number,[5]
2. the adjective inflection makes no distinction between gender in its form,[6] and the attractor is in singular,
3. the attractor and the subject are marked with the same gender and are both marked with singular,
4. the subject and the attractor are both marked with plural,
5. the subject and attractor differ in gender but the adjective does not have a singular lemma.

After applying the constraints, only 28 sentences were elicited. Due to this scarce number, the duplicated sentences were *manually* altered into their ungrammatical form by modifying the form of the adjective to modify the attractor. In some cases where the attractor consisted of multiple nouns, the noun phrase was simplified to contain only the first noun.Since the constraints in the script for PRED_a are also in place for PRED_b, there was an overlap of sentences in both subsets after processing. Thus, as a last step, the sentences from PRED_a that also occurred in PRED_a were removed from PRED_a.

It is important to note that the script does not take into account the occasions where the predicative agreement is governed by semantics rather than grammar. One systematized example of this is the phenomenon of the singular neuter form of an adjective being used to describe an abstract noun, regardless of the grammatical gender or number of that noun (e.g., "skatteberäkning[`COM`] kan vara jobbigt[`NEU`] att utföra" (*Tax calculations can be difficult to perform*, or "En avromantisering[`COM`] av äktenskapet är nödvändigt[`NEU`] för kvinnans egen skull" (*A deromantization of marriage is neccessary for the woman's own benefit*)). Nouns that are conceptually interpreted as an entity or situation are generally referred to by the general neuter determiner "det" (*it*), which triggers the acceptability of the neuter agreement on the adjective. Other examples where multiple adjective markings may be acceptable, albeit not grammatical, include the singular vs. plural agreement with nouns representing groups of people, such as "Nämnden[`COM, SING`] var inte beredda[`PLU`] att ta ett beslut i frågan" (*The committee was not prepared to make a decision on the matter*).

The abovementioned phenomena serve as examples of the fact that grammaticality and acceptability are not equivalent concepts. Sentences that violate these grammatical aspects were manually removed in the PRED_b subset, but may occur in the PRED_a subset.

A.6. Inversion after topicalization (INV)

The selection of sentences for the INV subset was made based on the following constraints:

1. The first constituent (after any punctuation or conjunctions) before the finite verb must not be a subject (passive or active), expletive, interrogative pronoun or imperative verb.
2. The sentence must contain a subject (active or passive) or expletive.

The script operates by identifying the root verb or auxiliary dependent of the root verb and identifying all constituents that precede this verb, as well as all the

[5] Examples are the adjectives "skrämmande" or "bra", that can modify nouns of any grammatical gender or number.

[6] An example is the adjective "indiskret", which retains its form when modifying neuter nouns.

subject dependents. The position of the subject phrase is then switched with the finite verb/auxiliary to form the ungrammatical sentence of the minimal pair.

Observe that the adverb "kanske" (*maybe*) in the initial position in the main clause sometimes is found with a directly succeeding subject.[7] In other words, it is realized grammatically as a conjunction rather than an adverb. Although it is tempting to filter out such occurrences, all scripts operate based on grammar principles, not acceptability principles, and an exception should not be made here.

A.7. Pre-verbal negation in subclauses (NEGV)

The python script used for selecting and processing sentences for the NEGV subset used the following constraints:

1. The sentence must contain a negation with the lemma "inte".
2. The negation head must either be a clausal subject, a clausal complement, an adverbial clause modifier, a clausal modifier of a noun or a relative clause modifier.
3. The negation must precede its head.
4. The negation must not be topicalized.

The script functions by identifying the embedded negation, the embedded verb (and its dependent auxiliary, if applicable) and the embedded subject. The duplicated sentence is then altered into its ungrammatical form by switching the positions of the negation and the finite verb/auxiliary. In cases where the subject succeeds the negation rather than preceding it, the subject is also moved in front of the verb in order to form canonical SVneg(O) word order.

It should be noted that the script also allows for sentences where the precedes the subject in a dependent clause, such as "kostnaderna" in the following Talbanken sentence: "Där kan hyrorna i stort sett inte ändras såvida inte kostnaderna[SUBJ] ökar[...]." (*There, the rents can't be changed much unless the costs[SUBJ] increase.*) This word order results in an ungrammatical equivalent in the subset where the verb precedes the subject ("såvida ökar kostnaderna inte"), i.e. XVSneg main clause word order. Although using a subjunction as a topicalized constituent is not grammatical, with a context window that excludes the first constituent "såvida", the VSneg sequence is grammatical.

It is common that multiple clausal adverbials (such as "inte", "alltid", "fortfarande" (*not, always, still*)) are placed in juxtaposition in a sentence. The position of the negation "inte" in relation to other clausal adverbials can differ. In some cases, this entails that the clausal adverbials will be separated from each other in the ungrammatical sentence, e.g. in "[...]eftersom jag bara inte har[...]" -> "[...]eftersom jag bara har inte[...]" (*[...]because I just don't have[...]*). When the negation precedes the second clause adverbial however, it results in another word order in the ungrammatical sentence. Ponder the Talbanken sentence "Det är ett jobb som inte[neg] bara kräver[VERB] en eller två föräldrar utan insatser från så många olika håll[...]." (*It's a job that doesn't just require one or two parents, but input from so many different sides.*) Observe that the swapping of the negation and verb in this sentence results in an ungrammatical sentence where the negation doesn't immediately follow the verb. For this particular sentence, the change in word order ("som kräver bara inte en eller två föräldrar, utan[...]") results in an arguably acceptable interpretation of the dependencies, where "bara" (*just*) is related to the noun phrase "en eller två föräldrar" (*one or two parents*) rather than the verb

[7] Compare the sentences "Kanske är jag hungrig" and "Kanske jag är hungrig", which are both acceptable in Swedish.

"kräver" (*demands*). It is possible that such cases, if present in the training data, may confuse the model in its acceptability judgments.

A.8. Non-inversion in indirect questions (INQ)

The python script used for selecting and processing sentences for the INQ subset applied the following constraints:

1. The sentence must contain a matrix verb whose lemma corresponds to "undra", "fråga", "fundera", "undersöka", "veta", "gissa", "förklara", "diskutera" or "beskriva".
2. The sentence must include either a question word or the lemma "om" (*if*) or "huruvida" (*whether*) with the marker relation.
3. The question word must not be the subject.[8]
4. The sentence must contain an embedded verb that governs the question word or marker, and if applicable an auxiliary that is governed by such a verb.[9]
5. If the conjunction has the lemma "om" or "huruvida", the embedded verb must have the clausal complement relation.[10]
6. The embedded clause must contain a nominal subject or an expletive which must not be a relative pronoun.

The script functions by identifying the embedded verb or auxiliary dependent of the embedded verb, as well as identifying the subject phrase in the embedded clause. The position of the verb or auxiliary is then switched with the subject to form the ungrammatical sentence in the minimal pair. If the embedded clause contains a negation that precedes the subject,[11] the negation and finite verb will also swap positions.

[8] as in e.g. "Jag undrar vem som kommer." While this is a valid indirect question, the question word must be separate from the subject in order to generate the ungrammatical equivalent.

[9] This excludes indirect questions where the question word is a dependent of a noun, e.g. in "Jag undrar vilken bok han läser". Since such examples are in minority, and already excluded if containing a subject relative pronoun (e.g. "Jag undrar vilken bok *som* är bra" (*I wonder which book* that *is good*), this constraint was applied in favor of simplicity in the identification of the embedded verb.

[10] This constraint separates indirect questions from conditional clauses.

[11] e.g. "Man kan fråga sig om [inte]`NEG` [detta antagande]`SUBJ` är felaktigt" (*One may wonder whether (if not) this assumption is incorrect)*

B. Comparison of training and evaluation dataset populations

Stage	Structure	Talbanken/LinES		OSCAR	
2	SVO	2,581	13,65%	4,208,314	11,66%
	PLUR	479	2,53%	1,275,606	3,53%
	TENSE	9,698	51,28%	18,871,751	52,29%
3	ATTR	2,268	11,99%	4,816,125	13,34%
4	INV	3,258	17,23%	5,888,655	16,32%
	PRED_a	226	1,20%	413,598	1,15%
	PRED_b	4	0,02%	2,823	0,01%
5	NEGV	304	1,61%	543,445	1,51%
	INQ	94	0,50%	70,230	0,19%

Table B.1.: Comparison of the distribution between PT structures in the Talbanken/LinES dataset and the training data (OSCAR after parsing 9%). Raw data is presented in the left columns, and the percentage of sentences annotated with each respective category in the right columns. A Chi-Square Test shows that the populations are significantly different. (X^2 (8, 17,874 = sample size) = 251.62, p < 0.001)

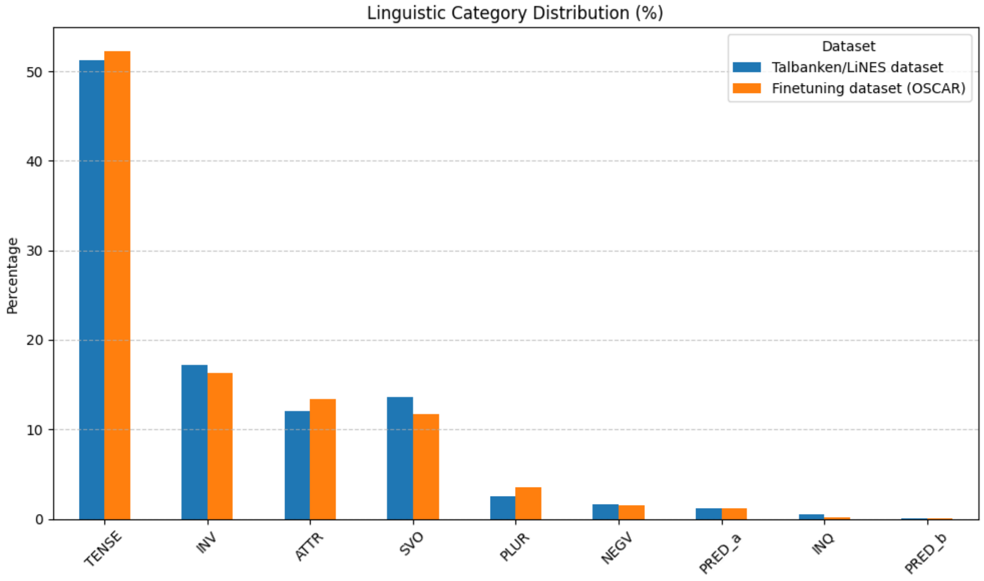


Figure B.1.: Distribution of linguistic categories between the evaluation dataset (SwePT) and the training data (subset from OSCAR) after parsing. The syntactic categories (SVO, INV, NEGV, INQ) are consistently more frequent in the evaluation data, compared to in the training data. The opposite relationship is observed among the morphological structures (ATTR, TENSE, PLUR, PRED) which are independent from grammatical sentence structure.

C. Additional AJT results

C.1. Acquisition times

Model	SVO	PLUR	TENSE	ATTR	PRED_a	PRED_b	INV	NEGV	INQ
GPT-mixed	1	100	1	10	-	-	1	100	10
GPT-mixed_2	1	100	1	10	-	-	1	100	10
GPT-order	1	100	1	100	100	-	1	154	10
GPT-reverse	1	100	1	10	53	-	1	1	1

(a) Acquisition threshold set at 50% accuracy.

Model	SVO	PLUR	TENSE	ATTR	PRED_a	PRED_b	INV	NEGV	INQ
GPT-mixed	1	128	10	10	-	-	100	100	100
GPT-mixed_2	1	126	10	100	-	-	100	100	100
GPT-order	1	142	1	100	-	-	109	154	100
GPT-reverse	1	-	10	10	-	-	10	1	100

(b) Acquisition threshold set at 60% accuracy.

Model	SVO	PLUR	TENSE	ATTR	PRED_a	PRED_b	INV	NEGV	INQ
GPT-mixed	100	-	100	-	-	-	-	-	100
GPT-mixed_2	100	-	100	-	-	-	-	-	-
GPT-order	100	-	117	100	-	-	-	154	156
GPT-reverse	100	-	100	56	-	-	-	2	109

(c) Acquisition threshold set at 80% accuracy.

Table C.1.: The time of acquisition per structure and model, calculated at thresholds of 50%, 60% and 80% accuracy (see Table 3.5 for the calculation of each structure-specific threshold). The numbers indicate the checkpoints, which are saved at intervals of 100 time steps. A dash indicates that that structure has never reached the respective threshold (not acquired).

C.2. Rank correlation

Threshold	Mean correlation	p-value
50%	0.8373	0.0000
60%	0.7214	0.0000
80%	0.8833	0.0110

Table C.2.: Results from the rank correlation permutation test across all four models. The high mean correlation scores indicate that the models acquire the grammatical structures in a consistent order. The low p-values indicate high significance of this correlation. The high correlation but low significance for the 80% accuracy threshold is explained by the lower amount of data points in that group.

C.3. Acquisition trajectories

C.4. Model confidence on acceptability scores

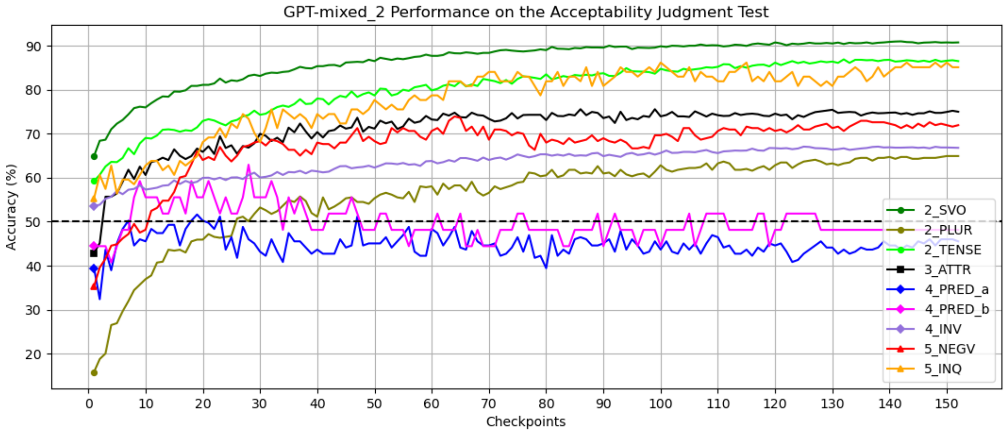
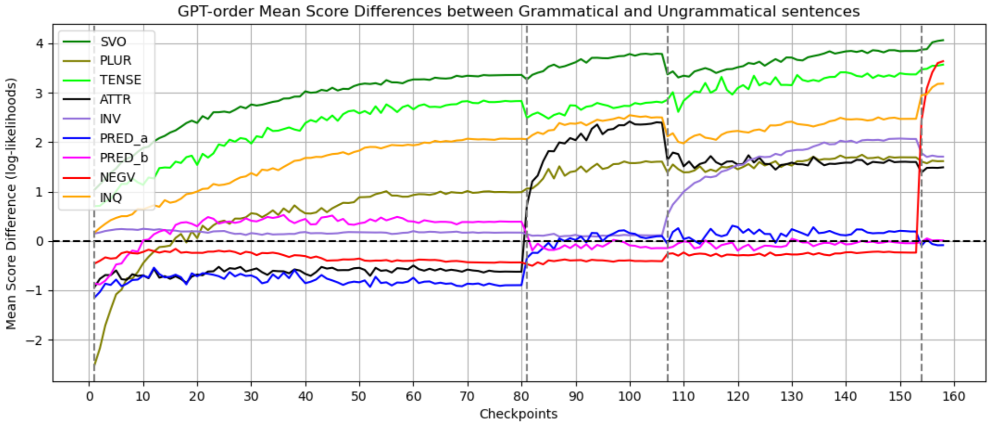
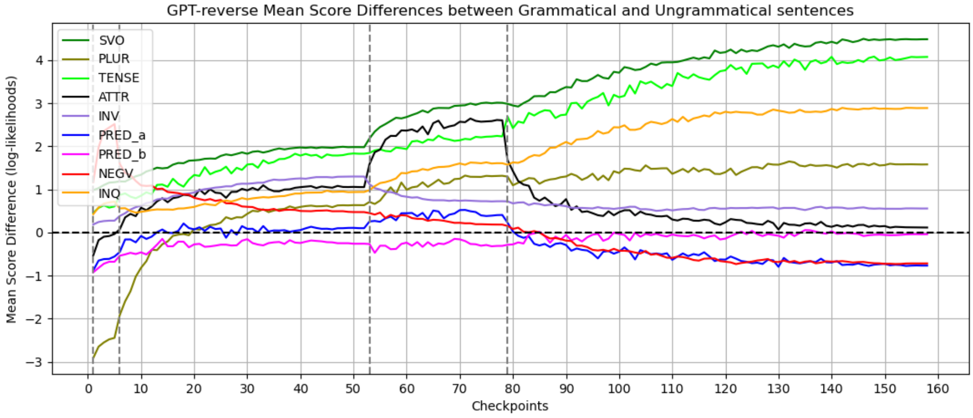


Figure C.1.: Results from the AJT on all linguistic structures across checkpoints for the model trained on all four data subsets concatenated and shuffled (seed=123). The acquisition trajectories follow a relatively parallel pattern, and the acquisition order deviates from that predicted by PT.

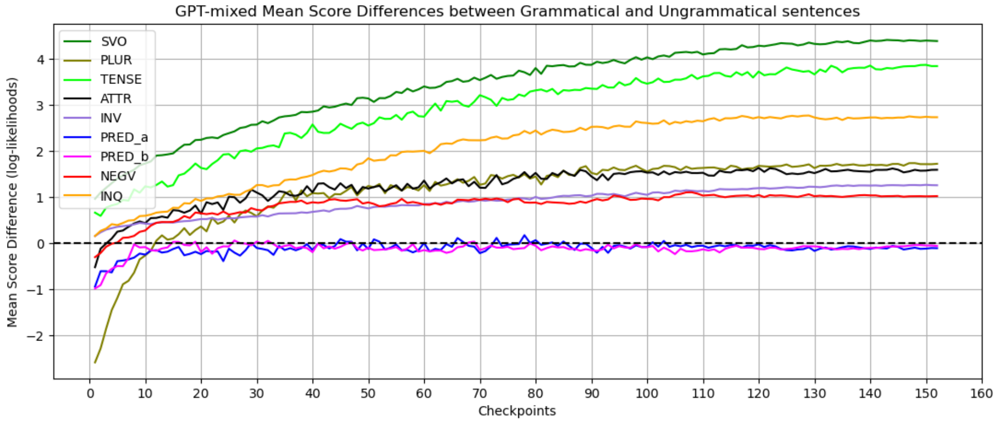


(a) Absolute differences between the log-likelihood scores assigned to the grammatical and ungrammatical sentences of the minimal pairs from the AJT, for the GPT-order model. The vertical dashed lines mark where training commences on data from a new Stage subset.

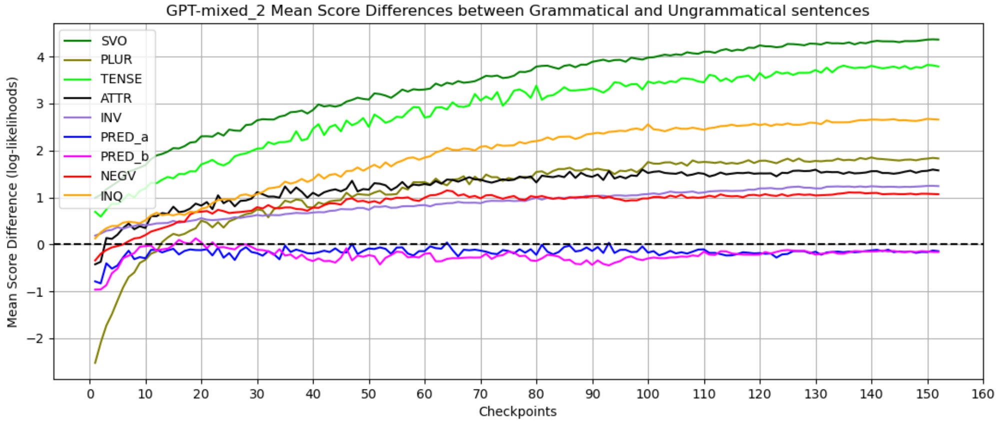


(b) Absolute differences between the log-likelihood scores assigned to the grammatical and ungrammatical sentences of the minimal pairs from the AJT, for the GPT-reverse model. The vertical dashed lines mark where training commences on data from a new Stage subset.

Figure C.2.: Absolute differences between the log-likelihood scores assigned to the grammatical and ungrammatical sentences of the minimal pairs from the AJT, for the curriculum models.



(a) Absolute differences between the log-likelihood scores assigned to the grammatical and ungrammatical sentences of the minimal pairs from the AJT, for the GPT-mixed model.



(b) Absolute differences between the log-likelihood scores assigned to the grammatical and ungrammatical sentences of the minimal pairs from the AJT, for the second GPT-mixed model.

Figure C.3.: Absolute differences between the log-likelihood scores assigned to the grammatical and ungrammatical sentences of the minimal pairs from the AJT for the two mixed models.