Do large language models and humans follow similar L2 learning stages? Assessing GPT-2's Swedish grammar acquisition within the Processability Theory framework

Stella Lundqvist[1] **Murathan Kurfalı**[2] **Johan Sjons**[1]
[1]Department of Linguistics and Philology, Uppsala University, Sweden
[2]RISE Research Institutes of Sweden
stellazoelouise.lundqvist.0915@student.uu.se murathan.kurfali@ri.se
johan.sjons@lingfil.uu.se

Abstract

We investigate whether GPT-2 acquires Swedish grammatical structures in the same implicational order as for human second language (L2) learners, as predicted by Processability Theory (PT). We present SwePT – a minimal pair dataset 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 using an acceptability classification task throughout fine-tuning with different input orders in regards to the grammatical structures identified in the data. 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, we do not find evidence for PT-like grammatical development in our experiments.

1 Introduction

Despite language models' (LLMs) ability to acquire complex linguistic patterns and generate coherent and human-like language, the mechanisms underlying their grammatical development remain poorly understood. Research on these capabilities often draws inspiration from studies of child language development (e.g., McCoy et al., 2018; Choshen et al., 2021; Warstadt et al., 2023; Evanson et al., 2023; Yedetore et al., 2023), yet insights from second language acquisition (SLA) research remain largely unexplored. One promising framework is Processability Theory (PT; Pienemann, 1998b, 2005), one of the most influential theories on second language (L2) grammatical development in SLA research. This psycholinguistic theory posits that a learner can acquire only those linguistic forms and functions that they are cognitively prepared to process, constrained by the processing procedures available at their current stage of language development. The procedures are accessed in a predictable order sequence, supported by cross-linguistic evidence from numerous empirical studies of speech production and grammatical perception (e.g., Norrby and Håkansson, 2007; Kawaguchi, 2008; Mansouri, 2008; Ellis, 2008; Keatinge and Keßler, 2009; Wang, 2011; Spinner, 2013; Buyl and Housen, 2015). To date, however, no study has investigated the hypotheses of PT on artificial learners.

In this study, we evaluate an LLM against PT's developmental sequence, allowing us to examine whether the emergence of grammatical patterns in artificial systems follows trajectories similar to those observed in human second language learners. We approximate Swedish L2 learners by fine-tuning a GPT-2 model pretrained on English, on Swedish text data organized in three different curricula: randomized input order, order of increasing complexity and order of decreasing complexity, as hypothesized by PT.

We present the Swedish Processability Theory Minimal Pair Dataset (SwePT), consisting of nine subsets of minimal pairs containing different grammatical phenomena that represent four stages of the Swedish PT developmental hierarchy. We test the models' grammatical knowledge development using acceptability judgments (AJT) on SwePT at regular intervals throughout fine-tuning, to examine whether the model exhibits a developmental trajectory in acquiring Swedish grammatical structures that aligns with the trajectory that has been observed within human L2 learners. We 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, and analyze the learning

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 52–76

trajectories.

The main contributions of this study are: (i) introducing a novel approach and bridging the fields of SLA and research on LLMs 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 AJT with curriculum learning. By evaluating whether human developmental patterns are shared by artificial learners such as GPT-2, our study 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 and Related Work

2.1 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 (Saphra and Lopez, 2018; Chiang et al., 2020; Liu et al., 2021; Choshen et al., 2021; Blevins et al., 2022; La Fiandra et al., 2025). Choshen et al. (2021) found a systematic learning trajectory across LLMs with different initializations, architecture and training data, albeit at different speeds, and 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 our study.

2.2 Measuring Linguistic Competence within the PT Framework

PT is built upon Levelt's (1989) model of speech *production*, which inherently views the cognitive processes of production and reception as separate. It is thus not surprising that the vast majority of PT studies concern speech production, with only a few (e.g. Norrby and Håkansson, 2007) including written data. Four studies (Ellis, 2008; Keatinge and Keßler, 2009; Spinner, 2013; Buyl and Housen, 2015) have previously investigated grammatical *comprehension* within the PT framework. The lack of unity in the methodological approaches and findings in these studies highlight the need for additional research to determine whether PT can predict receptive processing sequences with the same reliability as productive sequences.

2.3 AJT and Minimal Pair Benchmarks

A common method for inferring linguistic knowledge of language models is using acceptability judgment tests (AJT) with benchmarks of *minimal pairs*, where the learner is presented with one grammatical and one ungrammatical sentence that differ from each other on a single linguistic aspect and is tasked to determine which one of them is grammatical. Significant contributions to this practice include the Corpus of Linguistic Acceptability (CoLA, (Warstadt, 2019), The Benchmark of Linguistic Minimal Pairs for English (BLiMP Warstadt et al., 2020) and the Russian Benchmark of Linguistic Minimal Pairs (RuBLiMP, Taktasheva et al., 2024). The most significant contribution to Swedish AJT are the Dataset for Linguistic Acceptability Judgments (DaLAJ and DaLAJ-GED,

Volodina et al., 2021, 2023). While these benchmarks include a large number of minimal pairs and cover a broad range of linguistic phenomena allowing for evaluating general performance on AJT, SwePT is designed specifically to test the developmental sequence as predicted by PT, including linguistic phenomena not covered in pre-existing Swedish datasets.

3 Methodology

3.1 SwePT: A Swedish PT minimal pairs dataset

We present the Swedish Processability Theory Minimal Pair Dataset (SwePT), consisting of nine subsets of minimal pairs containing different grammatical phenomena that represent four stages of the Swedish PT developmental hierarchy, namely SVO (canonical word order SVO, 2nd stage), PLU (plural, 2nd stage), TNS (tense, 2nd stage), ATT (attributive agreement, 3rd stage), PR_a (predicative agreement, 4th stage), PR_b (predicative agreement with attractors, 4th stage), INV (inversion after topicalization, 4th stage), NEG (preverbal negation, 5th stage) and INQ (Non-inversion in indirect questions, 5th stage). Examples of the minimal pairs representing each subset are presented in Table 1.

Processing Pipeline. SwePT was constructed with an automated approach similar to that of RuBLiMP (Taktasheva et al., 2024). The grammatical sentences of each minimal pair were selected from the Swedish Talbanken and LinES treebanks from UD (De Marneffe et al., 2021) after processing the sentences through a custom pipeline identifying the target linguistic structures using rule-based Python scripts.[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 Appendix A.

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 (see Section 3.2). To form the minimal pairs of the syntactic structures (SVO, INV, INQ and NEG), relevant grammatical constituents and arguments were identified and had their positions switched with respect to the target structure. The alteration of the morphological structures (PLU, TNS, ATT and PR_a) was performed by converting the conjugated target structures into their neutral form (lemma). The alteration process for the PR_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 process are described in Appendix A.

3.2 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.

Labeling the Fine-tuning Data. Each sentence in the data was first parsed, annotated and converted into CoNLL-U format using Stanza (Qi et al., 2020). The parsing performance was evaluated by comparing the distribution of linguistic categories in the parsed OSCAR subset with those in the gold-standard Talbanken and LinES corpora (Table 5 and Figure 2, Appendix B). 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. 500k sentences were then randomly sampled from the 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). See Appendix A for details on the identification criteria.

Grouping the Fine-tuning Data. After parsing the raw text from OSCAR and labeling the training data, the sentences were grouped into four subsets

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

[2]https://huggingface.co/datasets/oscar-corpus/OSCAR-2201

Structure	n of pairs	Example
5 NEG	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 PR_a	226	De flesta u-länder har varit [koloniserade] *De flesta u-länder har varit [koloniserad] *(Most developing countries have been colonized)*
4 PR_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 ATT	213	Han har inget [civiliserat] ansikte. *Han har inget [civiliserad] ansikte. (*He does not have a civilized face.*)
2 TNS	2000	Jag [är] min fars dotter. *Jag [vara] min fars dotter. (*I am my father's daughter.*)
2 PLU	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 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.

representing each of the developmental stages (2–5) in the PT hierarchy. The subsets were populated in decreasing order, and the sentences in each subset thus only contain 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., none of the target 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 4 in Appendix B. Observe that the subsets are different in size, since each developmental stage are represented by different numbers of linguistic structures that occur in varying frequencies in the training data.

3.3 Fine-tuning

Models. We fine-tuned and evaluated four small (124 M parameters) GPT-2 model instances[3] pre-trained on English. As a causal (unidirectional) transformer, GPT-2 estimates the probability of the next word given its previous context (Radford et al., 2019). This aspect is similar to the incremental processing of humans (e.g., Altmann and Kamide, 1999; Kuribayashi et al., 2025), which is one of 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. We employed the method of curriculum learning (Bengio et al., 2009) during fine-tuning, where models are initially fine-tuned

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

on simpler concepts and gradually move on to more complex concepts. 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`, seed=42), one in reverse order (`GPT-reverse`, seed=42) and two on all four subsets concatenated into one dataset (`GPT-mixed`, seed=42 and `GPT-mixed_2`, seed=123), 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. Checkpoints were saved at each 100th time step and named according to their indices. Training arguments are specified in Appendix B.3.

3.4 Evaluation

Acceptability Judgment Test. We 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 of each pair. 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 accuracy is calculated as the percentage of the pairs where the grammatical sentence was given a higher score than its ungrammatical counterpart.

Acquisition Time and the Emergence Criterion. While most SLA 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. insufficient 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).[4]

The emergence criterion has been 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), as AJT and not language output are used for measuring the models' acquisition in our study. Before reaching a certain threshold during training, the model is expected to distribute the probabilities over the grammatical and ungrammatical sentences somewhat randomly, and the model's acquisition of the structure cannot be inferred from a single correctly identified grammatical sentence without the context of the cumulative probabilities of the entire subset. To account for some noise around the chance level mark, we thus set the acquisition threshold at 60% accuracy. Following the approach of Buyl and Housen (2015), we also evaluated at 50% and 80%, and calculated the k number of sentences per subset that must be correct in order to ensure acquisition at each threshold. The acquisition thresholds are displayed in Table 6 in Appendix C.

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).

Learning Trajectories. For the purpose of examining learning trajectories, we perform a rank

[4] 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.

correlation permutation test,[5] 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.

4 Results and Discussion

4.1 Performance on SwePT

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[6] model. GPT-SW3 was pretrained on 320B tokens of text in Scandinavian languages, mainly Swedish, and thus functions as a skyline. Table 2 displays the final accuracies from the AJT on SwePT of all models. GPT-SW3's average accuracy score 95.38% roughly aligns with the manually calculated precision score of 97.11% (see Appendix A.3).

The accuracies across all structures and fine-tuned models, with the exception of NEG 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, 20M Swedish tokens in the fine-tuning data cannot compare in size to the 320B tokens that GPT-SW3 was trained on and is likely insufficient to reach maximum performance.

Accuracy on PR_a and PR_b remains below chance for all fine-tuned models. GPT-SW3's higher performance on these structures suggests that attractor-sensitive hierarchical generalization in PR_b is weaker than heuristic-based strategies, though this conclusion is tentative given the small number of PR_b minimal pairs.

The low NEG accuracy in `GPT-reverse` is likely due to catastrophic forgetting (McCloskey and Cohen, 1989): NEG appears only during the first 500 time steps, after which post-negations dominate training. The superior overall performance of `GPT-order` supports this explanation, as earlier structures are repeatedly reintroduced later in training, reducing forgetting.

4.2 Acquisition Time

Acquisition time is defined as the checkpoint at which accuracy first exceeds a given threshold. Table 8 in Appendix E displays the acquisition times across models and thresholds. The hypothesized PT implicational order can be inferred from the acquisition times of `GPT-order`, where at least one structure per stage is acquired before or simultaneously as higher-stage structures, with the exception of INQ, which is expected given the input order.

There is a noticeable variability in acquisition times within stages. PLU emerges over 1000 time steps later than SVO and TNS at the 50% and 60% thresholds and never reaches 80%. INV crosses the 50% threshold early, while PR structures are acquired late or not at all. Although SVO and TNS (Stage 2) are generally acquired early, in `GPT-reverse` they emerge after NEG (Stage 5), which is introduced first to `GPT-reverse`. This sensitivity to input order suggests that PT predictions are not robust under curriculum manipulation, consistent with the implicational scaling results. A rank-correlation permutation test shows consistent relative acquisition orders across all models (Table 9, Appendix E), despite differences in absolute timing.[7]

4.3 Implicational Patterns

Table 7 presents collapsed implicational scales using thresholds at 50%, 60% and 80% accuracy.

While the observed order differs slightly between the three scales, all observed patterns deviate from the predicted order as hypothesized by PT. In all scales, higher-stage structures emerge before lower-stage ones; for instance, INQ (Stage 5) precedes PR_a, PR_b, and PLU. There is also significant variability within each scale. The IR (index of reproducibility) coefficients across all three scales are far below the 0.93 scalability threshold (Rickford, 2004). This implies that the observed order is not implicational.

4.4 Learning Trajectories

Figure 1 displays the acquisition trajectories of all linguistic structures tested through the AJT. The

[5] e.g., the Spearman's coefficient of rank correlation, or Spearman's ρ (Gibbons and Chakraborti, 2014).

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

[7] Note that that the rank correlation test only measures the rank order at the predefined acquisition thresholds, meaning that the rank order across the entire fine-tuning sequence is not reflected in these results. See Figure 1 for the ordering of the grammatical structures across the entire sequence.

Model	SVO	PLU	TNS	ATT	PR_a	PR_b	INV	NEG	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 2: 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.

results from the `GPT-mixed_2`, which follows a very similar pattern to the `GPT-mixed` model, is presented in Figure 3 in Appendix E.1.

`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 PR_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. NEG 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 PR subsets never reach above the 60% threshold, but rather decrease in accuracy throughout training. Notably, in all models SVO, TNS, INQ and INV have already reached an above-chance accuracy at the first checkpoint.

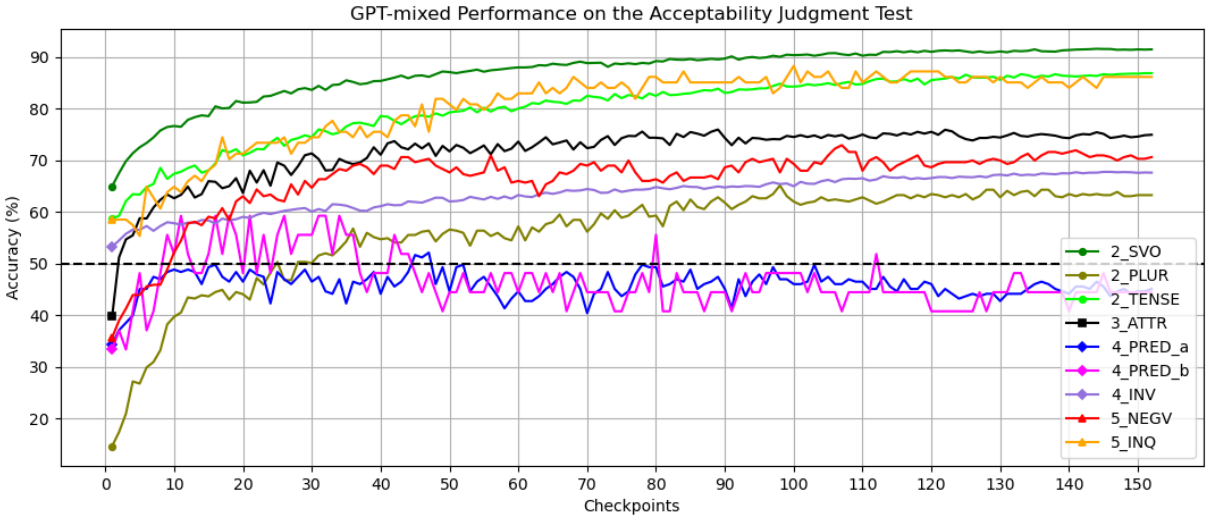
In the curriculum models (Figures 1b and 1c), accuracy closely tracks the distribution of structures in the training data (Table 4). NEG illustrates this clearly: in `GPT-reverse`, it exceeds 90% accuracy during early Stage-5 training, then steadily declines to 30% by the final checkpoint. The reverse pattern appears in `GPT-order`. ATT shows a similar curriculum-dependent rise, with accuracy increasing sharply when its corresponding stage is introduced.

4.5 Curriculum Learning Effects

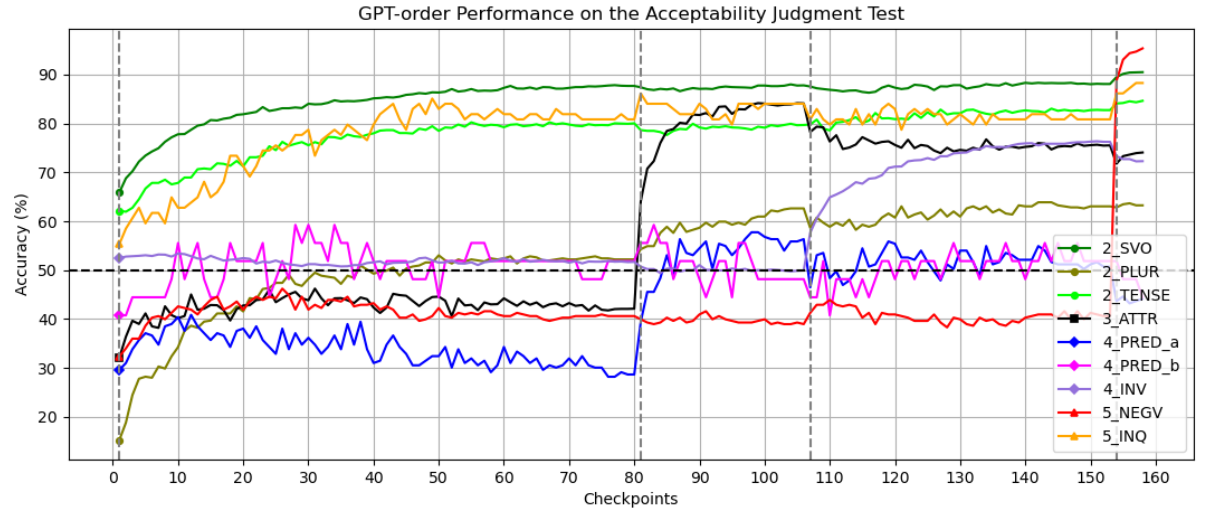
Although the acquisition order contradicts PT, trajectories are systematic and align with prior findings (see Section 2.1) that LLMs behave similarly to bag-of-words models during initial training stages (Choshen et al., 2021). To examine this pattern of unigram statistics, we calculated the distribution of the contrasting morphological features in each minimal pair of SwePT, by searching the dependency trees using simple rule-based scripts. The distribution (see Appendix A.2) reveals that 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 PLU 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 ATT, since attributive adjectives in common form (lemma) are in majority in the training data. For TNS, 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 PLU and ATT 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 (Saphra and Lopez, 2018; Choshen et al., 2021).

Interestingly, the accuracy on the PR subsets stays around chance-level in the mixed model, while performance on PR_a fluctuates in predictable patterns for `GPT-order` and `GPT-reverse` with regards to the training data. The minimal pairs for PR_a were constructed using the same principles as for ATT, where neuter gendered or plural form of the adjective is contrasted to the common, singular form of the same adjective in the ungrammatical sentence. This implies that the grammatical sentence in the PR_a subset will be subject to the same frequency-based bias as in ATT, as previously discussed, thus favoring the ungrammatical version of each pair. PR_b, while smaller and manually constructed without this bias, also shows no improvement, suggesting that the models have not reached training stages where global hierarchical cues dominate, such as agreement beyond the NP (Stage 4 in PT).

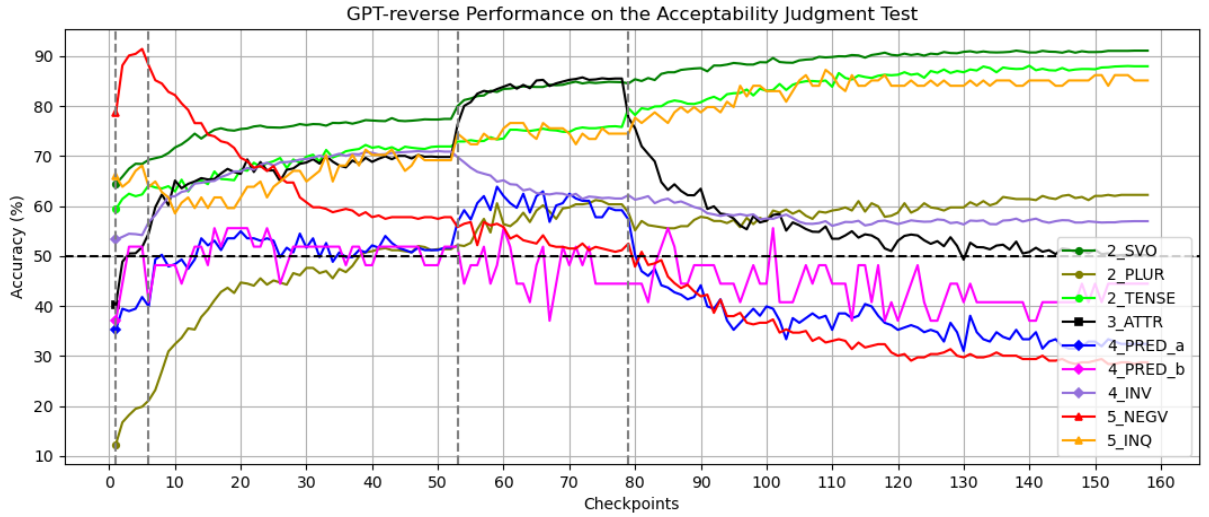
INQ consistently shows high accuracy across all



(a) Results from the AJT on all linguistic structures across checkpoints for the `GPT-mixed` model trained on all four data subsets concatenated and shuffled.



(b) 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.



(c) 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.

Figure 1: AJT results across checkpoints for GPT-2 fine-tuned under three different conditions. The vertical dashed lines in `GPT-reverse` and `GPT-order` mark where training commences on data from a new Stage subset.

models despite its hypothesized difficulty in PT. This can be explained by the fact that interrogative subclauses share syntactic structure with declarative subclauses and are likely found across all stages in the training data. As a result, models can generalize INQ word order early. The similarity between INQ and Stage-2 structures (SVO, TNS) suggests that GPT-2 relies on syntactic regularities rather than semantic overgeneralization. Given evidence that semantic information is learned later than syntax (Blevins et al., 2022; Saphra and Lopez, 2018), a single epoch of fine-tuning may be insufficient for models to adopt human-like semantic strategies, even though simpler n-gram-based heuristics prove effective for INQ.

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 would 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.

It is also possible that GPT-2 does not possess the same processing constraints as humans do, that such constraints are induced later in training, that different learning strategies are applied or that PT truly cannot be tested through receptive skills, as discussed by Pienemann (1998b) and Dyson and Håkansson (2017). Alternatively, if the relevant distinction is not between production and reception but rather between implicit and explicit knowledge, as argued by Ellis (2008), it may be the case that the "cognitive processes" behind GPT-2's next-word prediction, guiding 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.

5 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 can largely 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 AJT.

Limitations and Future Work

Adapting the emergence criterion. As previously addressed by authors of PT studies focusing on receptive skills, the emergence criterion cannot be seamlessly adapted to align with the evaluation of the AJT. A fundamental conceptual distinction that this study fails to resolve remains: the fact that accuracy and emergence reflect separate aspects of acquisition. Since the accuracy rates of different structures develop with 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 inherently 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 and requires further addressing.

Furthermore, evidence of rule application in its obligatory context does not only encompass grammatically correct target-like forms, but any application that can demonstrate that the learner is able to process that grammatical rule[8]. By limiting the evaluation to the binary classification of a minimal pair, valuable information from non-target con-

[8]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, serving as evidence for access to the procedure despite its incorrect surface form (Flyman Mattsson, 2022).

structions may be lost (see e.g. Schönström, 2014).

Future work could avoid these methodological issues by focusing on the language *production* of LLMs, where the emergence criterion could be implemented without the need for adaptation to accuracy scores, and potential "interlanguage" constructions of the target grammatical rule could be taken into account, offering insights beyond the limitations of minimal pairs.

Minimal pair dataset generation. The minimal pairs of SwePT were identified and altered on the basis of a single metric of complexity: the presupposed processing constraints of its target linguistic structure. Each sentence is thus assumed to be equally as complex, regardless of aspects such as token frequency and sentence length that are confounding factors to the acceptability of a sentence as predicted by probability. While sentence length does not impact the difference in score between sentences of a single minimal pair that are equal in length, it may impact the overall processability and make the acceptability scores less reliable, as may differences in parse-depth across pairs.

Furthermore, the identification constraints for the morphological structures could allow more variety of obligatory context to avoid tying the performance on the AJT to a single heuristic. For example, the TNS subset may elicit different performance on the AJT with more variation of verb forms across the pairs, and with the obligatory context for a specific tense based on semantic or syntactic cues rather than context-independent occurrences. Within pairs, the difference in semantic plausibility after altering the grammatical sentence is not accounted for. Manual refinement of the scripts where length, token frequencies and an extended set of obligatory contexts are controlled for may increase the interpretability of the AJT results.

The observed number of false positives could also be reduced by utilizing native-speaker crowdsourcing and/or a Swedish LM such as GPT-SW3 for evaluation of the minimal pairs.

Parsing evaluation. 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.

Additional models. With additional computing resources, including additional models beside GPT-2 would increase the relevance of the study. In order to offer additional insights on transfer effects and the modeling of first and second language acquisition, a Swedish model such as GPT-SW3 could be pretrained and evaluated, providing a comparison between the acquisition order of a fine-tuned "L2 learner" and itself as an "L1 learner".

References

Gerry TM Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

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

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

Aafke Buyl and Alex Housen. 2015. Developmental stages in receptive grammar acquisition: A processability theory account. *Second Language Research*, 31(4):523–550.

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

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

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

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

Rod Ellis. 2008. Investigating grammatical difficulty in second language learning: Implications for second language acquisition research and language testing. *International journal of applied linguistics*, 18(1):4–22.

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

Anna Flyman Mattsson. 2022. Rethinking textbook grammar introduction. *Instructed Second Language Acquisition*, 6(2):196–218.

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

Gisela Håkansson and Catrin Norrby. 2010. Environmental influence on language acquisition: Comparing second and foreign language acquisition of swedish. *Language learning*, 60(3):628–650.

Fredrik Heinat. 2012. Finiteness in swedish. *Working papers in Scandinavian syntax*, 90:81–110.

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

Satomi Kawaguchi. 2008. Argument structure and syntactic development in japanese as a second language. In *Cross-linguistic aspects of processability theory*, pages 253–298. John Benjamins Publishing Company.

Dagmar Keatinge 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*, pages 67–92.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *arXiv preprint arXiv:2502.01615*.

Olivia La Fiandra, Nathalie Fernandez Echeverri, Patrick Shafto, and Naomi Feldman. 2025. Large language models and children have different learning trajectories in determiner acquisition. In *Proceedings of the First BabyLM Workshop*, pages 100–108.

Willem Levelt. 1989. Speaking-from intention to articulation. *A Bradford book*.

Leo Z Liu, 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*.

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

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

R Thomas McCoy, 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*.

Catrin Norrby 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*, 45:45–68.

Manfred Pienemann. 1998a. Developmental dynamics in l1 and l2 acquisition: Processability theory and generative entrenchment. *Bilingualism: Language and cognition*, 1(1):1–20.

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

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

Peng Qi, 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*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

John R Rickford. 2004. Implicational scales. *The handbook of language variation and change*, pages 142–167.

Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with svcca. *arXiv preprint arXiv:1811.00225*.

Krister Schönström. 2014. Visual acquisition of swedish in deaf children: An l2 processability approach. *Linguistic approaches to bilingualism*, 4(1):61–88.

Patti Spinner. 2013. Language production and reception: A processability theory study. *Language Learning*, 63(4):704–739.

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

Elena Volodina, 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*, pages 94–101.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021. Dalaj-a dataset for linguistic acceptability judgments for swedish: Format, baseline, sharing. *arXiv preprint arXiv:2105.06681*.

Xiaojing Wang. 2011. *Grammatical development among Chinese L2 learners: From a processability account*. Ph.D. thesis, Newcastle University.

A Warstadt. 2019. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Alex Warstadt, 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*.

Alex Warstadt, 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*, 8:377–392.

Aditya Yedetore, 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*.

A Details on SwePT

A.1 Processing Pipeline Details

The Swedish Talbanken and LinES treebanks from UD were merged into a single CoNLL-U data file before processing. 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. 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 scripts allow no duplicates and only returns the minimal pair if the grammatical and ungrammatical sentence are not identical.

The following section contains examples and explanations of each grammatical structure of the Swedish PT hierarchy, as well as the criteria for identification of each grammatical structure.

A.1.1 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 examples 1 and 2 below.

(1) Jag läser boken
I read.FIN book

'I read the book'

(2) *Jag boken läser
I book.FIN read

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.[9]
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.1.2 Plural (PLU)

Access to the category procedure also enables encoding lexical information, which is necessary for marking plural on nouns and tense on verbs. Plural

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

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 4 below.

(3) En grå elefant
one gray.SG elephant.SG

'One gray elephant'

(4) Två grå-a elefant-er
two gray-PL elephant-PL

'Two gray elephants'

(5) *Två grå-a elefant
two gray-PL elephant.SG

The selection of sentences for the PLU 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.[10]

4. The lowercased form of the noun must not be the same as the noun's lemma.[11]

5. The noun must not be an abbreviation.[12]

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.

A.1.3 Tense (TNS)

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 examples 6 and 7 below.

(6) Jag såg ett träd
I see.PST a tree

'I saw a tree'

(7) Jag se ett träd
I see.INF a tree

The selection of sentences for the TNS 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.

[10] 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.

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

[12] 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.1.4 Definiteness (N/A)

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.

A.1.5 Non-inversion with topicalization (N/A)

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 model cannot be tested on the *XSV structure of this stage, which is why this structure is excluded in the present study.

A.1.6 Attributive agreement (ATT)

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.

(8) Den stor-a hund-en
DET.COM.SG.DEF big-SG.DEF dog.SG-DEF.COM

'The big dog'

(9) Många stor-a hund-ar
many big-PL.IND dog-PL.COM.IND

'Many big dogs'

(10) *Många stor hund-ar
many big.SG.COM.IND dog-PL.COM.IND

The selection of sentences for the ATT 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.1.7 Predicative agreement (PR_a and PR_b)

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 11 and 12 illustrate 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.

(11) Hund-ar-na [på gård-en] är stor-a
dog-PL-DEF.COM on yard.SG-DEF.COM be big-PL

'The dogs on the yard are big'

(12) *Hund-ar-na [på gård-en] är
dog-PL-DEF.COM on yard.SG-DEF.COM be
stor
big.SG.COM

The PR subset is separated into PR_a and PR_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 PR_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 PR_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,[13]

2. the adjective inflection makes no distinction between gender in its form,[14] 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 PR_a are also in place for PR_b, there was an overlap of sentences in both subsets after processing. Thus, as a last step, the sentences from PR_a that also occurred in PR_a were removed from PR_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 PR_b subset, but may occur in the PR_a subset.

A.1.8 Inversion after topicalization (INV)

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 13b 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 15 below.

(13) Jag ska läsa boken imorgon
I will.FIN read book tomorrow

'I will read the book tomorrow'

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

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

(14) Imorgon ska jag läsa boken
tomorrow will.FIN I read book

'Tomorrow I will read the book'

(15) *Imorgon jag ska läsa boken
tomorrow I will.FIN read book

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 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.[15] 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.1.9 Pre-verbal negation in subclauses (NEG)

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 16 and 17 below). Before this process is accessed, learners generalize the SVneg word order from main clauses to subclauses, resulting in ungrammatical sequences such as example 18.

[15]Compare the sentences "Kanske är jag hungrig" and "Kanske jag är hungrig", which are both acceptable in Swedish (see e.g., Heinat, 2012).

(16) 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'

(17) 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'

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

The python script used for selecting and processing sentences for the NEG 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 "kräver" (*demands*). It is possible that such cases, if present in the training data, may confuse the model in its acceptability judgments.

A.1.10 Non-inversion in indirect questions (INQ)

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,[16] in indirect questions the subject precedes the verb in the interrogative subclause.[17] 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 20 and 22 below.

(19) Jag undrar vad hon inte har gjort
I wonder what she not have.FIN do

'I wonder what she hasn't done'

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

(21) Jag undrar om hon kommer
I wonder if she come.FIN

'I wonder if she's coming'

(22) *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.

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.[18]

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.[19]

5. If the conjunction has the lemma "om" or "huruvida", the embedded verb must have the clausal complement relation.[20]

6. The embedded clause must contain a nominal subject or an expletive which must not be a relative pronoun.

[16] e.g. "Vad äter du?" (*lit. What eat you (What are you eating)?*)) or "Äter du?" (*lit. Eat you (do you eat)?*)

[17] e.g. "Jag undrar vad/om du äter" (*lit. I wonder what/if you eat*)

[18] 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.

[19] 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.

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

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,[21] the negation and finite verb will also swap positions.

A.2 Distribution of Morphological Forms in SwePT

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 PLU subset, 79% of non-genitive nouns (not counting abbreviations) were in singular (lemma) form, and 21% in plural form. Relevant to the TNS subset, 28% of verbs were in infinitive (lemma) form, and 72% in tensed form.

A.3 Evaluation of SwePT

The minimal pair generation was evaluated manually, identifying 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. Observe that the PR_b subset does not contain any false positives by default, since the ungrammatical sentences were manually generated.

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

Subset	SVO	PLU	TNS	ATT	PR_a
FP	2/50	2/50	0/50	2/50	3/50
Error rate	4%	4%	0%	4%	6%
Subset	PR_b	INV	NEG	INQ	Avg.
FP	0/50	0/50	2/50	2/50	
Error rate	0%	0%	4%	4%	2.89%

Table 3: 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%.

B Fine-tuning Details

B.1 Fine-tuning Data

Structure	Stage-2	Stage-3	Stage-4	Stage-5	Total
INQ				1 362	1 362
NEG				10 645	10 645
PR_b			51	1	52
PR_a			7 855	217	8 072
INV			111 195	3 156	114 351
ATT		64 569	26 289	2 848	93 706
TNS	194 702	49 118	110 559	11 833	366 212
PLU	13 329	4 001	6 992	397	24 719
SVO	47 764	15 241	15 604	3 004	81 613
N/A	55 268	17 878	31 982	3 295	108 423
n of sents.	254 875	82 447	147 490	15 188	500 000

Table 4: Distribution of the linguistic structure labels of sentences in all four training subsets (stages 2-5).

B.2 Preprocessing of Fine-tuning Data

We used Stanza with the `tokenize`, `pos`, `lemma` and `depparse` tools in our pipeline, with batches of 50 on one GPU. The dataset was first partitioned into ten separate files for parallel processing. The sentences 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 timeouts and processed separately based on indexing, after which the files were concatenated into a single CoNLL-U file. Due to unforeseen issues during this process, some of the data was skipped unintentionally. This should be taken into account if replicating this study. Through this process, the dataset was reduced to 25,758,263 sentences/examples, 478,895,789 words and 1,021,799,273 tokens (after truncating with max_length=512).

A Chi-square test shows that the parsed OSCAR data differs significantly from the two gold-standard corpora, in that the syntactic categories (SVO, INV, NEG, 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 (ATT, TNS, PLU, PR). See Table 5 and Figure 2 for a visualization of the distribution. Stanza has a reported 87.85 labeled attachment score (LAS)[22] 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 assume that the output is good enough for the present purposes.

B.3 Training Arguments

The GPT-2 model instances 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. 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. 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.

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

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

Table 5: 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, $p < 0.001$)

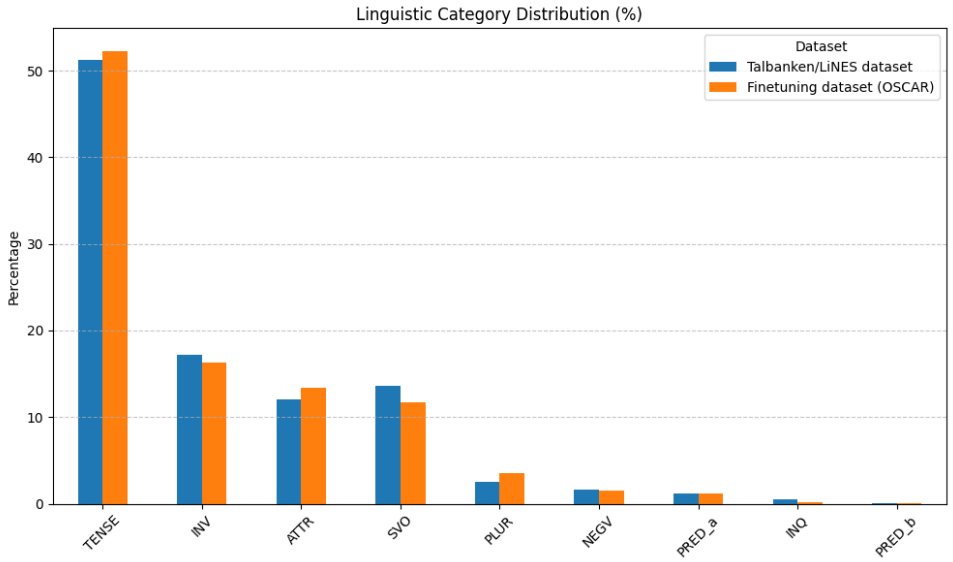


Figure 2: Distribution of linguistic categories between the evaluation dataset (SwePT) and the training data (subset from OSCAR) after parsing. The syntactic categories (SVO, INV, NEG, INQ) are consistently more frequent in the evaluation data, compared to in the training data. The opposite relationship is observed among the morphological structures (ATT, TNS, PLU, PR) which are independent from grammatical sentence structure.

C Acquisition thresholds

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 $P(X \geq k) = \text{BinomSF}(k-1, n, \pi)$.

Structure	n	π=50%	π=60%	π=80%
SVO	2519	1302	1553	2049
PLU	479	258	306	398
TNS	2000	1038	1237	1630
ATT	2268	1174	1400	1847
PR_a	213	119	140	181
PR_b	27	19	21	26
INV	2581	1333	1590	2099
NEG	303	167	197	255
INQ	94	56	65	82

Table 6: Acquisition thresholds k per subset where $n =$ number of minimal pairs, and the remaining columns showing $k =$ the number of pairs needed to be correct to ensure acquisition above different thresholds π (with $\alpha = 0.05$).

D Additional Result Details

D.1 Implicational Scales

In our study, each checkpoint across all model instances 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.

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)[23]. The rows (the learners) are ordered from more advanced to less advanced in terms of how many PT structures they have acquired.

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. 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.

[23] 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.

n	PR_b	PR_a	NEG	PLU	ATT	INV	INQ	SVO	TNS
4	-	+	+	+	+	+	+	+	+
235	-	-	+	+	+	+	+	+	+
21	-	+	-	+	+	+	+	+	+
3	-	+	+	-	+	+	+	+	+
99	-	-	-	+	+	+	+	+	+
96	-	-	+	-	+	+	+	+	+
11	-	+	-	+	+	-	+	+	+
15	-	-	-	-	+	+	+	+	+
26	-	-	-	+	-	+	+	+	+
15	-	-	-	+	+	-	+	+	+
4	-	-	+	-	-	+	+	+	+
1	-	-	+	-	+	+	-	+	+
49	-	-	-	-	-	+	+	+	+
6	-	-	-	-	+	+	-	+	+
30	-	-	-	-	-	-	+	+	+
5	-	-	-	-	-	+	-	+	+

(a) Acq. threshold: 50%. IR = 0.393

n	PR_a	PR_b	PLU	NEG	INV	ATT	INQ	TNS	SVO
30	-	-	+	+	+	+	+	+	+
194	-	-	-	+	+	+	+	+	+
3	-	-	+	-	+	+	+	+	+
81	-	-	-	-	+	+	+	+	+
32	-	-	-	+	-	+	+	+	+
17	-	-	-	+	+	+	-	+	+
58	-	-	-	-	-	+	+	+	+
12	-	-	-	-	+	+	-	+	+
3	-	-	-	+	-	+	-	+	+
1	-	-	-	+	+	-	-	+	+
131	-	-	-	-	-	-	+	+	+
19	-	-	-	-	-	+	-	+	+
5	-	-	-	+	-	-	-	+	+
28	-	-	-	-	-	-	-	+	+
2	-	-	-	+	-	-	-	-	+
4	-	-	-	-	-	-	-	-	+

(b) Acq. threshold: 60%. IR = 0.292

n	PLU	PR_a	PR_b	INV	NEG	INQ	ATT	TNS	SVO
3	-	-	-	-	+	+	-	+	+
12	-	-	-	-	-	+	-	+	+
2	-	-	-	-	+	-	-	+	+
254	-	-	-	-	-	-	-	+	+
41	-	-	-	-	-	-	+	-	+
195	-	-	-	-	-	-	-	-	+
7	-	-	-	-	+	-	-	-	-
106	-	-	-	-	-	-	-	-	-

(c) Acq. threshold: 80%. IR = 0.697

Table 7: 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. IR = Index of Reproducibility

E Acquisition times

Model	SVO	PLU	TNS	ATT	PR_a	PR_b	INV	NEG	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	PLU	TNS	ATT	PR_a	PR_b	INV	NEG	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	PLU	TNS	ATT	PR_a	PR_b	INV	NEG	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 8: The time of acquisition per structure and model, calculated at thresholds of 50%, 60% and 80% accuracy (see Table 6 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).

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

Table 9: 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.

E.1 Acquisition trajectories

E.2 Model confidence on acceptability scores

Figures 4 and 5 plot the absolute differences between the log-likelihood scores assigned to the grammatical vs. the ungrammatical sentence of each minimal pair from the AJT, i.e. the models' "confidence" 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.

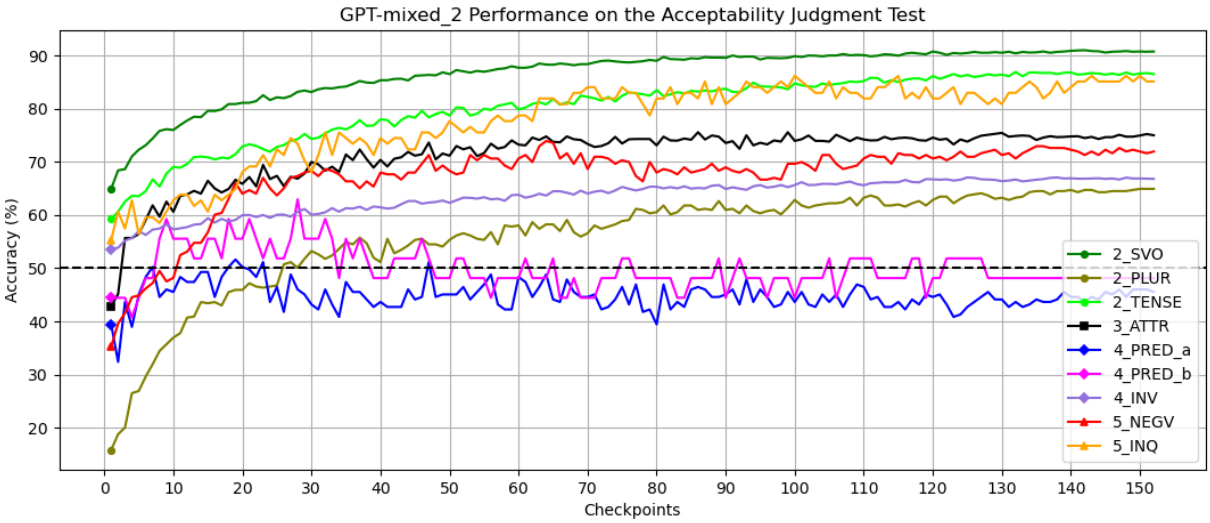
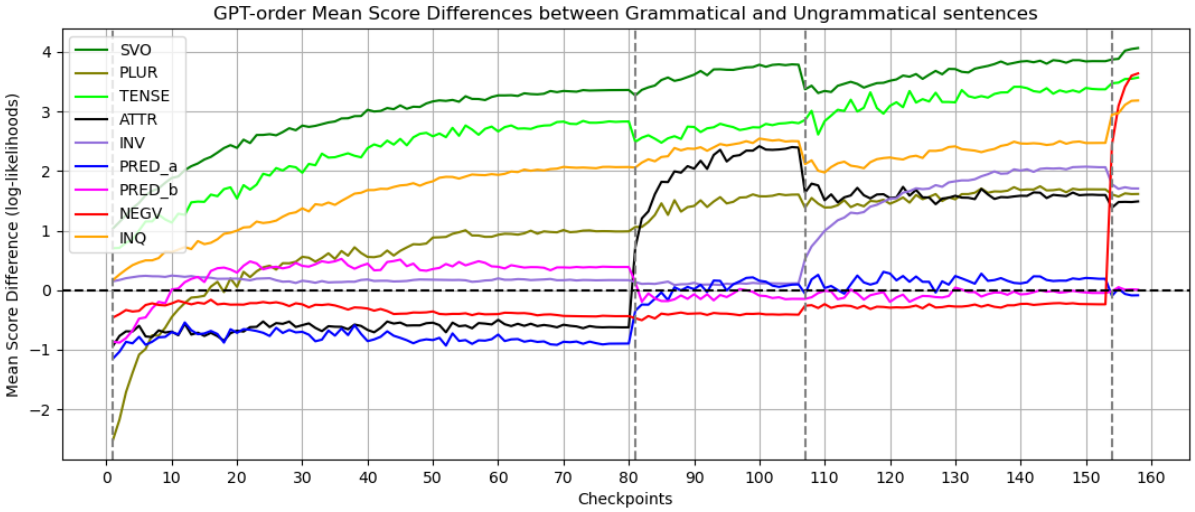
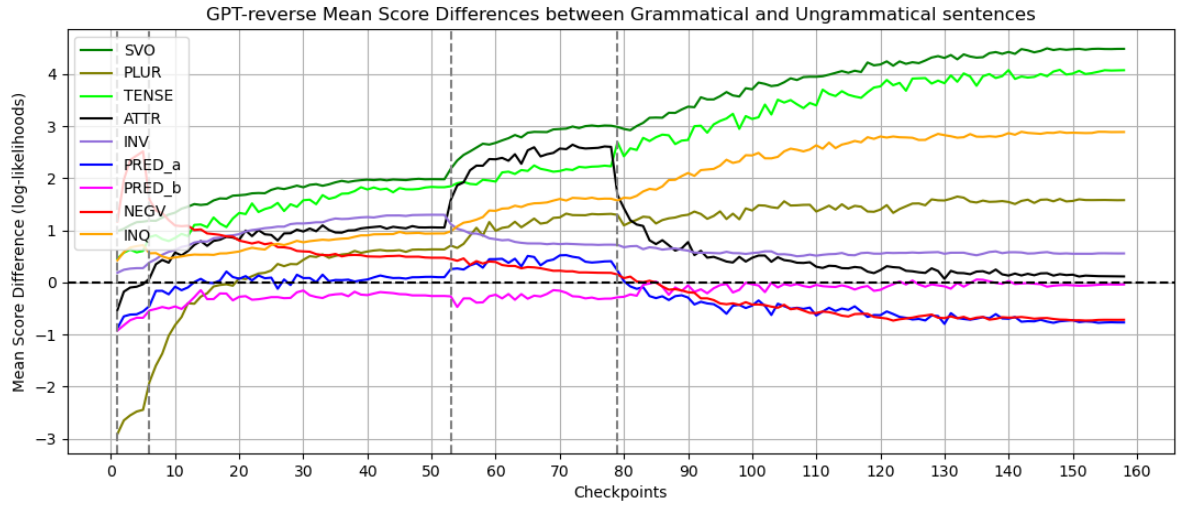


Figure 3: Results from the AJT on all linguistic structures across checkpoints for the model trained on all four data subsets concatenated and shuffled (seed=123). The acquisition trajectories follow a relatively parallel pattern, and the acquisition order deviates from that predicted by PT.

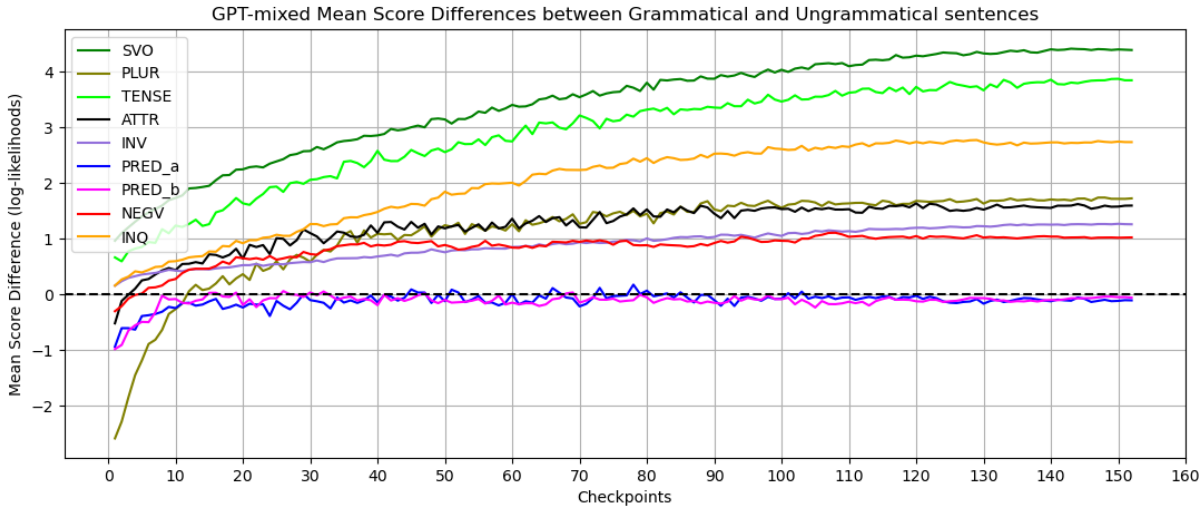


(a) Absolute differences for the GPT-order model.

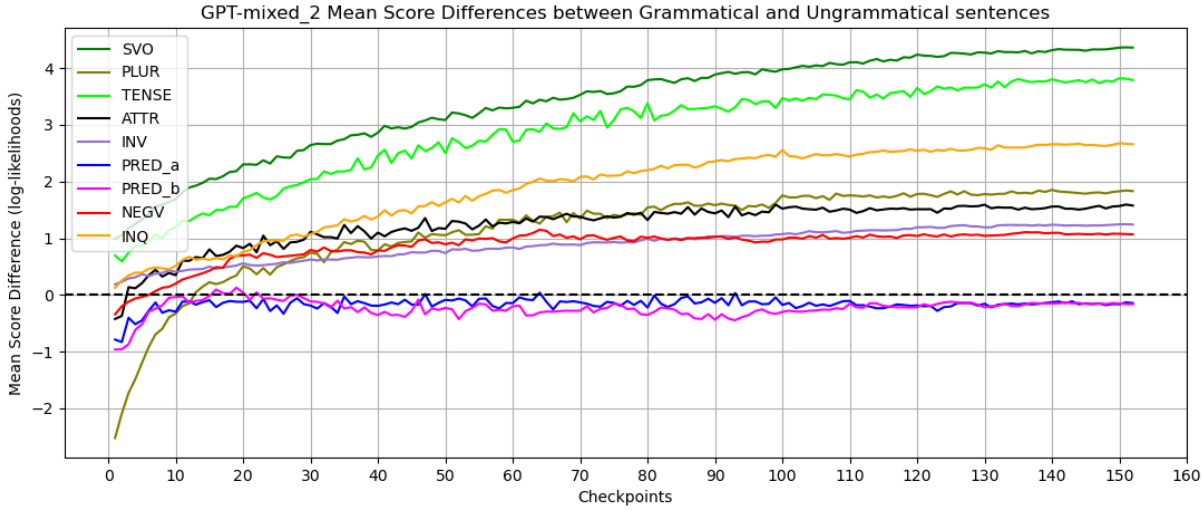


(b) Absolute differences for the GPT-reverse model.

Figure 4: 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. The vertical dashed lines mark where training commences on data from a new Stage subset.



(a) Absolute differences for the GPT-mixed model.



(b) Absolute differences for the second GPT-mixed model.

Figure 5: Absolute differences between the log-likelihood scores assigned to the grammatical and ungrammatical sentences of the minimal pairs from the AJT for the mixed models.