The Swedish Benchmark of Linguistic Minimal Pairs

Johan Sjons[1], **Fredrik Heinat**[2], **Murathan Kurfalı**[3]
[1]Department of Linguistics and Philology, Uppsala University, Sweden
[2]Centre for Languages and Literature, Lund University, Sweden
[3]RISE Research Institutes of Sweden, Sweden
johan.sjons@lingfil.uu.se, fredrik.heinat@englund.lu.se, murathan.kurfali@ri.se

Abstract

We introduce the Swedish Benchmark of Linguistic Minimal Pairs, a dataset for evaluating syntactic performance in language models. It includes 2,500 minimal pairs organized into 25 syntactic phenomena, with 100 pairs per phenomenon. Each pair contrasts a well-formed and an ill-formed sentence that differ minimally. For each phenomenon, we manually constructed ten pairs from scratch. We semi-automatically generated the remaining 90 pairs and manually adjusted them. A random sample was assessed by 40 participants, who selected the well-formed sentence in 98.05% of cases. We evaluate eleven state-of-the-art models. Results show that models generally handle local agreement well but struggle with certain long-distance dependencies and word order phenomena. Model size seems to matter less than the training domain. Prompt-based evaluation generally lowers performance. We show that model performance is stable across handcrafted and generated subsets and across sample sizes, suggesting that 100 pairs per phenomenon suffice for reliable evaluation. Future work will expand the number of phenomena.

Keywords: syntactic evaluation, Swedish, language models, minimal pairs, benchmarking

1. Introduction

The output from language models (LMs), and in particular large LMs (LLMs), is remarkably similar to the output of human beings. However, intuitive comparisons of model and human output constitute a crude way of evaluating whether a model is similar to what it is a model of; while models may be mathematically straightforward to define, their realized form after training is not. More specifically, the definition of (L)LMs does not reveal to what extent the parameter settings represent their knowledge of language, syntax being a case in point. One way of approaching this problem is to follow recent calls for methodological rigor in syntactic evaluation of LMs (e.g., Kulmizev and Nivre, 2022), and create targeted evaluation resources.

A prominent example is *The Benchmark of Linguistic Minimal Pairs* (BLiMP; Warstadt et al., 2020), which contains 67,000 sentence pairs covering 67 syntactic phenomena. However, although there are roughly 7,000 languages in the world (Hammarström et al., 2025), following the general pattern of a skewed distribution of language resources (e.g., Belinkov and Glass, 2019; Joshi et al., 2020), only few have resources of this kind.

To expand on previous work in other languages (e.g., Warstadt et al., 2020; Xiang et al., 2021; Someya and Oseki, 2023; Suijkerbuijk et al., 2025), we present *The Swedish Benchmark of Linguistic Minimal Pairs* (Swe-BLiMP).[1] While there is similar previous work for Swedish, it has focused on L2 learner production data (Volodina et al., 2021), been tied to Processability Theory (Lundqvist, 2025), or been part of broader multilingual evaluations with a limited number of agreement phenomena (Jumelet et al., 2025). In terms of syntactic breadth and controlled minimal pairs, Swe-BLiMP is the most challenging Swedish syntax benchmark to date.

In Swe-BLiMP, we include 2,500 minimal pairs across 25 syntactic phenomena, each contrasting a well-formed sentence with a minimally altered ill-formed counterpart. We validate the contrasts using acceptability judgments from 40 non-syntacticians, who preferred the well-formed sentence in 98.05% of cases. We also test whether differences between fully manual and semi-automatic pairs, as well as sample size, affect model performance, and find no substantial impact.

To establish baselines, we evaluate 11 open-source, instruction-tuned models ranging from 0.3B to 70B parameters. A prediction was correct if the well-formed sentence received a lower sentence-level negative log-likelihood than its ill-formed counterpart. Performance varies across phenomena: models generally handle local agreement well but struggle with certain non-local configurations, particularly those involving verb-second word order and resumption. Differences across model families further suggest typologically specific syntactic weaknesses in multilingual systems.

2. Related Work

As LMs have become paramount in computational linguistics and related fields, questions regarding

[1] https://github.com/itsup/swe_blimp

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6783–6794
11–16 May 2026.

the extent to which they are good models of humans' capacity for syntax – the capacity to form larger units from smaller units – have emerged (e.g., Lau et al., 2015; Lakretz et al., 2019; Zhang et al., 2023). What is typically tested is whether sentence perplexity correlates with human acceptability judgments (though there are other methods, such as probing; see e.g., Hewitt and Manning, 2019).

More specifically, our study focuses on a particular form of testing. In essence, the lower perplexity according to the model of either member in minimal pairs of well-formed and ill-formed sentences is used to provide information about the model's knowledge of syntax (e.g., Linzen et al., 2016). Hence, datasets of this kind are needed.

Arguably, the most prominent example is *The Benchmark of Linguistic Minimal Pairs* (BLiMP; Warstadt et al., 2020), consisting of 67,000 English minimal pairs contrasting acceptable and unacceptable sentences, organized into 67 paradigms under 12 high-level phenomena. Pairs were created from grammar templates and tested for acceptability by online participants, which yielded 96.4% agreement. Model evaluation showed that GPT-2, which was the best model, performed well on morphology and basic syntax but did not reach human-like performance in cases involving more complex syntax, in particular in cases involving long-distance dependencies.

The work by Warstadt et al. (2020) has given rise to similar benchmarks. For example, there are corresponding resources for Mandarin Chinese. *The corpus of Chinese linguistic minimal pairs* (CLiMP; Xiang et al., 2021) has 16,000 semi-automatically generated minimal pairs across 16 syntactic contrasts, grouped into nine high-level phenomena, with human agreement at 95.8%. More recently, SLING (Song et al., 2022) introduced 38,000 minimal pairs derived from linguist-annotated sentences in the Chinese Treebank.

Similarly, the *Japanese BLiMP* (JBLiMP; Someya and Oseki, 2023) includes 331 minimal pairs derived from acceptability judgments in theoretical linguistics and organized into 11 categories, making use of both a broad linguistic coverage and the design of BLiMP. The small but high-quality set of examples is the result of *manually* collecting data from linguistic publications.

Moreover, only in 2025, several new BLiMP-style benchmarks were introduced. For Turkish (TurBLiMP; Başar et al., 2025), the dataset includes 16 phenomena with 16,000 minimal pairs, focusing on word order and morphological complexity. For Quebec French (QFrBLiMP; Beauchemin and Khoury, 2025), the benchmark comprises 1,761 pairs across 20 grammatical phenomena drawn from an official linguistic resource, with judgments from native speakers. For Dutch (BLiMP-NL; Suijkerbuijk et al., 2025), the dataset includes 8,400 pairs across 22 phenomena and combines acceptability ratings with reading-time measures.

For Swedish, there are a few datasets whose properties resemble those of BLiMP. Volodina et al. (2021) introduced DaLAJ-GED, a learner corpus curated for grammatical error detection in L2 Swedish. However, while it contains minimal pairs similar in form to those in BLiMP-style datasets, its focus lies on learner production rather than on syntax *per se*, as the pairs are not organized into explicit syntactic phenomena but annotated for error types spanning morphology, lexicon, and syntax. Consequently, while DaLAJ-GED is a high-quality resource for studying learner language or general acceptability, it is less suited for evaluating models' performance specifically on syntax, let alone specific syntactic phenomena.

Another Swedish dataset with a closer focus on syntax is SwePT (Lundqvist, 2025), constructed within the framework of Processability Theory (Pienemann, 1998). It consists of 8,226 minimal pairs in nine syntactic structures. The pairs were automatically generated from the Talbanken and LinES treebanks (De Marneffe et al., 2021). The dataset included a manual evaluation of a sample of pairs that exhibited a high precision of 97%. However, since SwePT is based on a framework for second-language acquisition, its phenomena are not strictly separated.

Swedish is also represented in MultiBLiMP (Jumelet et al., 2025), which covers 101 languages and over 128,000 minimal pairs that were generated using Universal Dependencies (De Marneffe et al., 2021) and UniMorph (Batsuren et al., 2022). However, despite its impressive size for the number of languages and minimal pairs, MultiBLiMP focuses only on subject–verb agreement, providing limited syntactic coverage for Swedish compared to our dataset, where models score near ceiling on Swedish (Jumelet et al., 2025).

In summary, while various BLiMP-style benchmarks have been introduced for several languages, and while Swedish has datasets that are useful in other areas of research, it still lacks a benchmark similar to BLiMP. To address this gap, we introduce Swe-BLiMP.

3. Syntactic Phenomena Selection and Description

The grammatical phenomena included in Swe-BLiMP represent well-established aspects of Swedish syntax. Since both English and Swedish are Germanic languages, there is considerable overlap between our categories and those in Warstadt et al. (2020). Descriptions of most of the phenomena that are specific to Swedish are dealt

with in Teleman et al. (1999), and Holmberg and Platzack (2005). It should be noted that there are many other phenomena that should be included in a more exhaustive benchmark.

3.1. Syntactic Categories

The following is a brief description of the phenomena included. We follow Warstadt et al. (2020) and label general phenomena, such as *word order*, *categories*. Some phenomena have subcategories, and these are called *paradigms*. Every sentence pair belongs to a particular paradigm. Where possible, we created both the well-formed and ill-formed sentences to have the same length.[2] The structures we include can be seen in Table 1, where each label refers to the grammatical violation in the ill-formed sentence.

Argument structure Argument structure (ARG. STR.) is the selectional requirements that verbs put on their arguments (Levin, 1993). These requirements can be both semantic (e.g., the requirement that an argument is +HUMAN), and syntactic (e.g., the requirement that the object of the verb is a noun phrase). The possibility of a verb to take an object or not, irrespective of semantic restrictions, is expressed by its *transitivity*. Verbs that require an object are transitive and verbs that cannot take an object are intransitive. The manipulation in the pairs is whether the verb is transitive or not. The well-formed sentence has a transitive verb and the ill-formed an intransitive verb, and both have a following noun phrase. Whether the semantics of the objects fit with the verbs is not a factor in the manipulation. The reason for this is that we test categorical selection only, and not theta-role violations (Chomsky, 1981), in contrast to Warstadt et al. (2020).

Agreement Swedish shows no subject-verb agreement, but is rich in agreement in the noun phrase (Börjars, 1998). Determiners and premodifying adjectives show gender, number and definiteness agreement with the noun head. Adjectives in predicative function show agreement with the predication base in number and gender. All these different conditions are tested. In the ART. NOUN AGR. paradigm, the agreement between an article / determiner and the head noun in a predicative noun phrase is tested regarding definiteness (DEF), gender (GEN) and number (SG./PL.).

Double definiteness Indefinite nouns take an indefinite determiner (e.g., the indefinite articles *en* and *ett*). Definiteness is marked as an inflection on the head noun in the noun phrase. However, if the noun is premodified, a definite determiner is also required (e.g., the definite articles *den* or *det*). This phenomenon is called *double definiteness* (DOUBLE DEF.) (Börjars, 1994).

Ellipsis Ellipsis refers to the possibility to omit a string of words from a sentence. In order to keep the length the same between the sentences in the pairs, only cases of ellipsis in the noun phrase have been included here (Lobeck, 2006).

Particle verbs In contrast to English, the order between particle and object is strictly particle - object, even if the object is a short noun phrase consisting of only a noun head or a pronoun. In the ill-formed sentence in pair, the order is object - particle (PARTICLE OBJ. ORDER).

Resumption There are usually no resumptive pronouns in Swedish, except in relative clauses when the noun head of the relative clause is the subject in a clause embedded inside the relative clause (Asudeh, 2012). The paradigm NO RESUMPTION tests the only structure where a resumptive pronoun is allowed in Swedish. The well-formed sentence in the pairs has a resumptive pronoun and the ill-formed sentence lacks a resumptive. RESUMPTION1 is an example where the gap in a relative clause is filled by a resumptive pronoun in the ill-formed sentence. RESUMPTION2 is structurally similar to NO RESUMPTION, but here the gap in the embedded clause is not a subject gap, and the ill-formed sentence has a resumptive pronoun in the gap-position.

Word order The category word order tests several different possible word orders in Swedish, both in main and subordinate clauses. Swedish is a verb second (V2) language. This means that the finite verb occupies the second position in main clauses. In the paradigms V3 CL./OBJ./PP/COMPLEX NP, the first position in the sentences is occupied by an object (OBJ.) or an adverbial in the form of a clause (CL.) or a prepositional phrase with a simple noun phrase as complement (PP), or a noun phrase with a modifying clause (COMPLEX NP). The ill-formed sentence in these paradigms keeps the subject-verb order, giving rise to V3 word order, just like well-formed sentences in English.

Subordinate clauses differ from main clauses in not showing properties of V2, where we included *that*-clauses and relative clauses. The first sentence in the pair has SVO order, and the second sentence has main clause word order, that is, V2. In the paradigms EMBEDDED V2 AdvP. and V2 REL. CL. an adverbial in the form of an adverb

[2] All except NO RESUMPTION, RESUMPTION1 and RESUMPTION2.

Category	Well-formed	Ill-formed
ARG. STR.	*Lisa såg vägen i parken*	**Lisa sprang vägen i parken*
DOUBLE DEF.	*Den röda stolen stod i rummet*	**Den röda stol stod i rummet*
ART. NOUN AGR. INDEF.	*En stol stod i rummet*	**En stolen stod i rummet*
ART. NOUN AGR. GEN.	*Det stod en stol i rummet*	**Det stod ett stol i rummet*
ART. NOUN AGR. PL.	*Det stod några stolar i rummet*	**Det stod några stol i rummet*
ART. NOUN AGR. SG.	*Det stod en stol i rummet*	**Det stod en stolar i rummet*
ELLIPSIS	*My åt en vit ko och Bo åt en röd*	**My åt en vit och Bo åt en röd ko*
EMBEDDED V2 AdvP.	*Erik sa att Ulla ofta tappar isen i april*	**Erik sa att Ulla tappar ofta isen i april*
EMBEDDED V3 CL.	*Olle sa att när det blir kväll läser Lisa böcker i köket*	**Olle sa att när det blir kväll Lisa läser böcker i köket*
EMBEDDED V3 OBJ.	*Olle påstod att böcker har Lisa läst sedan hon var liten*	**Olle påstod att böcker Lisa har läst sedan hon var liten*
PARTICLE OBJ. ORDER	*Lisa läste ut boken*	**Lisa läste boken ut*
ADJ. NOUN AGR. GEN.	*Ett äpple var rött*	**Ett äpple var röd*
ADJ. NOUN AGR. DEF.	*Äpplet var rött*	**Äpplet var röd*
ADJ. NOUN AGR. INDEF.	*Ett äpple var rött*	**Ett äpple var röd*
ADJ. NOUN AGR. NUM. PL.	*Några äpplen var röda*	**Några äpplen var rött*
ADJ. NOUN AGR. NUM. SG.	*Ett äpple var rött*	**Ett äpple var röda*
NO RESUMPTION	*Han väntar på en buss som han inte vet när den kommer*	**Han väntar på en buss som han inte vet när kommer*
RESUMPTION1	*Han väntar på en buss som kommer snart*	**Han väntar på en buss som den kommer snart*
RESUMPTION2	*Britta sitter i bilen som hon vet när hon lånade*	**Britta sitter i bilen som hon vet när hon lånade den*
V3 CL.	*När Olle diskar läser Lisa boken*	**När Olle diskar Lisa läser boken*
V3 OBJ.	*Boken ska Lisa läsa imorgon*	**Boken Lisa ska läsa imorgon*
V3 PP	*På kvällen läser Lisa boken*	**På kvällen Lisa läser boken*
V2 REL. CL.	*Lisa läste boken som Olle inte gillar*	**Lisa läste boken som Olle gillar inte*
V3 COMPLEX NP	*I sängen som står i gästrummet läser Lisa boken*	**I sängen som står i gästrummet Lisa läser boken*
V3	*Lisa kanske läser boken*	**Lisa aldrig läser boken*

Table 1: Examples of grammatical and ungrammatical minimal members in pairs per category. The category labels refer to what grammatical phenomenon is violated, except for ELLIPSIS, for which both members of the pair has an ellipsis but that of the ungrammatical member is ill-placed.

phrase indicates that the second sentence has an ill-formed V2 word order.

In addition, there is a limited number of adverbs that can occupy the second position giving rise to V3 word order in main clauses. In the pair of sentences in the V3 paradigm, the first sentence contains one such adverb in the second position, and the second sentence has an adverb that does not allow V3 in second position (Heinat, 2021).

4. Data Construction and Judgments

Swe-BLiMP comprises 25 phenomena with 100 minimal pairs each. Following Warstadt et al. (2020), we let participants choose between the well-formed and ill-formed sentences from four pairs per phenomenon (100 total) in a forced-choice task. We describe each in turn.

Data Construction For each category, we manually constructed ten well-formed sentences and ill-formed counterparts that differed minimally. As an illustration, consider the well-formed and ill-formed sentences in Example (1a-1b).

Because every pair followed a specific pattern, we used lists of words (generated by LLMs and partly manually) and combined the lists with simple shell commands to construct the remaining 90 pairs in a template-based manner.

All template-based pairs were manually checked, and semantically odd sentences (and their counterparts) were revised for higher idiomaticity. Some unnatural cases remain, and we leave quantifying their degree of unnaturalness and distribution for future work.

(1) a. *Han vänta-r på en buss som*
3SG wait-PRS on ART bus rel-comp
han inte vet när den
3SG not know.PRS when it
komm-er
come-PRS

'He is waiting for a bus and doesn't know when it is coming'

b. **Han vänta-r på en buss som han inte vet när komm-er*

Category	L-SV8B	Euro9B	Euro1.7B	L-3.3 70B	Gem3-12B	Oss20B	Gem3-4B	Aya32B	SW20B	SW1.3B	SW0.3B	Avg (SD)	Hum.
ART. NOUN AGR. GEN.	**99.0**	**99.0**	96.0	**99.0**	95.0	95.0	90.0	88.0	**99.0**	94.0	94.0	95.3 (3.7)	98.7
NO RESUMPTION	**100.0**	98.0	88.0	98.0	97.0	**100.0**	96.0	90.0	87.0	90.0	99.0	94.8 (5.0)	98.1
ART. NOUN AGR. PL.	**98.0**	94.0	96.0	97.0	93.0	82.0	87.0	93.0	57.0	50.0	47.0	81.3 (19.9)	99.4
DOUBLE DEF.	**97.0**	92.0	91.0	95.0	94.0	95.0	92.0	92.0	47.0	47.0	51.0	81.2 (21.2)	98.7
ADJ. NOUN AGR. NUM. PL.	91.0	90.0	**94.0**	89.0	92.0	91.0	88.0	82.0	57.0	43.0	42.0	78.1 (20.3)	100.0
V3	77.0	80.0	81.0	72.0	69.0	78.0	75.0	**82.0**	79.0	62.0	80.0	75.9 (6.1)	95.0
ADJ. NOUN AGR. NUM. SG.	84.0	80.0	77.0	79.0	81.0	75.0	80.0	**85.0**	56.0	59.0	72.0	75.3 (9.6)	99.4
PARTICLE OBJ. ORDER	**86.0**	84.0	84.0	81.0	80.0	75.0	75.0	58.0	67.0	64.0	65.0	74.5 (9.6)	98.7
EMBEDDED V2 AdvP.	85.0	**89.0**	83.0	83.0	83.0	87.0	74.0	76.0	45.0	61.0	44.0	73.6 (16.3)	98.1
ELLIPSIS	**92.0**	83.0	69.0	89.0	87.0	83.0	88.0	83.0	45.0	45.0	44.0	73.5 (19.4)	92.5
EMBEDDED V3 CL.	93.0	**94.0**	90.0	71.0	82.0	83.0	65.0	65.0	64.0	42.0	36.0	71.4 (19.6)	98.1
V3 PP	**95.0**	91.0	84.0	75.0	75.0	79.0	80.0	67.0	41.0	48.0	36.0	70.1 (20.0)	98.1
ART. NOUN AGR. SG.	71.0	71.0	**81.0**	71.0	66.0	71.0	59.0	65.0	61.0	67.0	71.0	68.5 (6.0)	100.0
V2 REL. CL.	87.0	89.0	**90.0**	71.0	81.0	89.0	65.0	71.0	38.0	37.0	34.0	68.4 (22.2)	100.0
ARG. STR.	**73.0**	63.0	63.0	71.0	68.0	67.0	**73.0**	66.0	55.0	52.0	56.0	64.3 (7.3)	97.5
V3 CL.	83.0	**85.0**	71.0	65.0	69.0	50.0	55.0	52.0	42.0	56.0	57.0	62.3 (13.6)	100.0
ADJ. NOUN AGR. GEN.	63.0	**71.0**	70.0	69.0	59.0	62.0	63.0	63.0	52.0	55.0	49.0	61.5 (7.2)	98.7
ADJ. NOUN AGR. INDEF.	63.0	66.0	**70.0**	65.0	63.0	59.0	63.0	67.0	50.0	51.0	50.0	60.6 (7.2)	96.8
V3 COMPLEX NP	**85.0**	79.0	60.0	50.0	69.0	49.0	77.0	52.0	45.0	54.0	36.0	59.6 (15.7)	99.4
ADJ. NOUN AGR. DEF.	**69.0**	68.0	52.0	58.0	63.0	48.0	59.0	54.0	51.0	52.0	55.0	57.2 (7.0)	100.0
ART. NOUN AGR. INDEF.	**68.0**	66.0	41.0	63.0	53.0	51.0	48.0	44.0	54.0	55.0	51.0	54.0 (8.6)	99.4
RESUMPTION1	**79.0**	70.0	**79.0**	65.0	41.0	45.0	45.0	58.0	16.0	13.0	2.0	46.6 (26.8)	99.4
V3 OBJ.	**68.0**	63.0	41.0	61.0	55.0	40.0	41.0	39.0	34.0	33.0	25.0	45.5 (14.0)	91.8
EMBEDDED V3 OBJ.	16.0	21.0	9.0	4.0	11.0	10.0	10.0	26.0	**64.0**	52.0	57.0	25.5 (21.7)	95.0
RESUMPTION2	2.0	3.0	14.0	2.0	3.0	3.0	2.0	6.0	**19.0**	16.0	1.0	6.5 (6.6)	98.1
Avg. (SD)	**77.0**	75.6	71.0	69.7	69.2	66.7	66.0	65.0	53.0	51.9	50.2	65.0 (8.9)	98.05

Table 2: Minimal-pair accuracy (%) per category and model. Final columns show per-category average (SD) and human accuracy; last row shows per-model averages with the overall mean in the final cell. (rows sorted by per-category average accuracy). `L` stands for `Llama`, where `L-SV` is the Swedish Llama model and `L-3.3 70B` is the `Llama-3.3 (70B)`, and `SW` refers to the GPT-SW3 models. *Hum.* refers to average human acceptability judgment.

Human judgments Though specific to English, the results presented by Sprouse and Almeida (2012) strongly suggest that individual acceptability judgments by linguists generalize to non-linguistics, which indicates that our procedure is robust for this first collection of sentence pairs of targeted grammatical phenomena in Swedish.

Nonetheless, we randomly sampled two pairs from the well-formed and two pairs of the ill-formed sentences, from both the handcrafted and the semi-automatically constructed pairs, amounting to 100 pairs in total (i.e., four pairs per phenomenon).

In a forced-choice task, 40 participants selected which member of each of the 100 pairs in the same sample sounded better. None of the participants were syntacticians, and only a handful were linguists. No metadata was collected, and apart from knowing who took part, individual responses remain anonymous. Participants received no payment.

We obtained a score of 98.05%, which is the proportion of cases where participants selected the well-formed sentence. Hence, for this sample, the grammatical contrasts were clear even to non-syntacticians.

5. Experiments

We benchmark our dataset on 11 instruct-tuned open-source LLMs. The model selection is justified by our goal to examine both linguistic specialization and model scale: we include Swedish-oriented models (GPT-SW3, Swedish LLaMA), European language-focused models (EuroLLM), and widely-used multilingual models (e.g., Gemma, Aya, LLaMA), spanning 356M to 70B parameters to investigate the effects of size and training languages. Concretely, we evaluate the following models: GPT-SW3 (356M, 1.3B, 20B; Ekgren et al., 2024), Swedish Llama (8B),[3] Gemma-3 (4B, 12B; Team et al., 2024), EuroLLM (1.9B, 12B), gpt-oss (20B; Agarwal et al., 2025), Aya (32B Üstün et al., 2024), and Llama-3.3 (70B) (Dubey et al., 2024).

For each pair $\langle S_{\checkmark}, S_{\times} \rangle$, the model is correct if it assigns higher preference to $S_{\checkmark}$. Since instruction-tuned models can behave differently under raw likelihood versus prompt-based judgments (Hu et al., 2024; Kurfalı and Östling, 2025; Jumelet et al., 2025), we model this preference in three ways:

- *Perplexity (main):* We compute full-sentence negative log-likelihoods using the model's LM head and select the sentence with the lower NLL (i.e., lower perplexity), normalized by token count. This mirrors the *simple LM* scoring of the original BLiMP paper (Warstadt et al., 2020).
- *Grammaticality prompt:* We present the two sentences in random order and ask the model (in Swedish) to choose the *grammatically correct* one by answering with a *single letter* A/B.
- *Acceptability prompt:* Identical to the previous strategy but framed as choosing the sentence that is *most natural in everyday usage*.

[3] https://huggingface.co/AI-Sweden-Models/Llama-3-8B-instruct

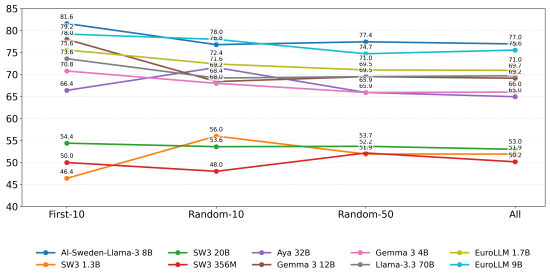

Figure 1: Overall accuracy (macro over categories) for each model across different subset conditions.

For prompting experiments, we used greedy decoding. Each sentence was evaluated in isolation (new chat per pair), with randomized order to avoid positional bias. All experiments ran on two V100 GPUs (≈20 GPU hours).

6. Results and Discussion

Table 2 reports accuracy per category and model. Overall, `Swedish Llama 8B` achieves the highest macro average (77.0%), followed by `EuroLLM 9B` (75.6%) and `EuroLLM 1.7B` (71.0%). Large general-purpose models follow this Swedish- and Europe-centric group: `Llama-3.3 70B` (69.7%), `Gemma-3 12B` (69.2%), `OSS 20B` (66.7%).

Easiest categories. Out of the 25 categories, 4 have a macro average above 80% and 12 above 70%. The easiest categories are found to be gender agreement between determiner and noun (`ART. NOUN AGR. GEN.`) (95.3%, $SD = 3.7$), `NO RESUMPTION` (94.8%, $SD = 5.0$), and `ELLIPSIS` (73.5% $SD = 19.4$), where most models achieve almost perfect score. Number agreement between nouns and determiners were also less challenging (`ART. NOUN AGR. PL.`, 81.3%, $SD = 19.9$), as is double definiteness (DOUBLE DEF., 81.2%,$SD = 21.2$). Yet, it should be noted that the latter category exhibits a strange family split, where most models achieved ≈92–95%, but GPT-SW3 models only ≈47–51%. All these are instances of local (inside the same phrase) morphological agreement. This type of agreement is of the same type that the models tested in Warstadt et al. (2020) found unproblematic. With morphological agreement between a subject noun phrase and phrases outside it, as in `ADJ. NOUN AGR. GEN./ INDEF`, performance drops considerably (cf. Warstadt et al., 2020). Ellipsis was also a category that performed well in Warstadt et al. (2020).

Hardest categories. The most challenging phenomena involve object placement under V2 and two of the resumption categories. `V3 OBJ.` averages 44.3% ($SD = 14.0$), while performance

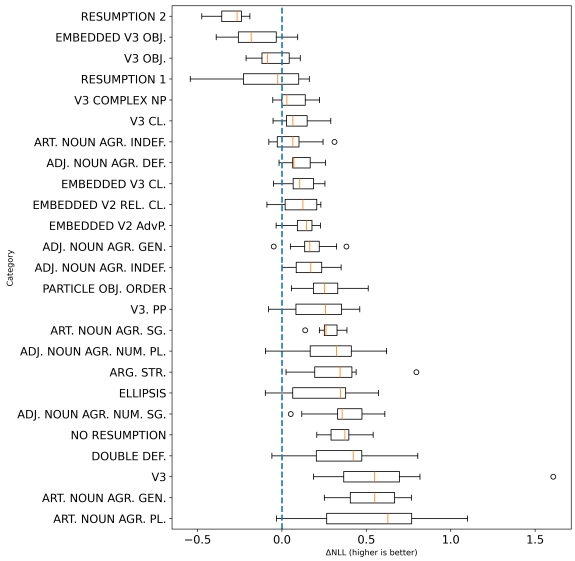

Figure 2: Per-category distribution of the difference in negative log-likelihoods between ungrammatical and grammatical sentences across models.

drops further to 25.5% ($SD = 21.7$) for embedded verb second with a fronted object (`EMBEDDED V3 OBJ.`) – models seem to struggle with V2 contexts, in particular in embedded clauses.

For `V3 OBJ.`, there is a potential confound. While "*Boken Lisa ska läsa imorgon" (*The book, Lisa will read tomorrow*) is not a well-formed sentence in Swedish, it is a well-formed noun phrase, as in the sentence "Boken Lisa ska läsa imorgon har försvunnit" (*The book Lisa will read tomorrow has disappeared*). Even though the sentence boundaries were made explicit to the models, such well-formed phrases might have confused the models.

The same analysis could be made for `EMBEDDED V3 OBJ.`. Consider the ill-formed member of the pair for this phenomenon: "*Olle påstod att böcker Lisa har läst sedan hon var liten" (*Olle claimed that Lisa has read books since she was little*. Part of this ill-formed sentence is: "Lisa har läst sedan hon var liten" (*Lisa has read since she was little*), which, though, is a grammatical declarative sentence. Perhaps the perplexity of *that particular part* decreases the overall perplexity of the sentence.

In other words, because model preference is determined by *average* rather than *local* perplexity, the performance of models under this evaluation method may be questionable.

Resumption exhibits a similar but steeper decline: models achieve near-perfect accuracy on NO RESUMPTION (94.8%, $SD = 5.0$), yet performance falls sharply to 46.6% ($SD = 26.8$) on `RESUMPTION1` and down to 6.5% ($SD = 6.6$) on `RESUMPTION2`. Even the strongest models are far below chance on RESUMPTION2, indicating a *sys-*

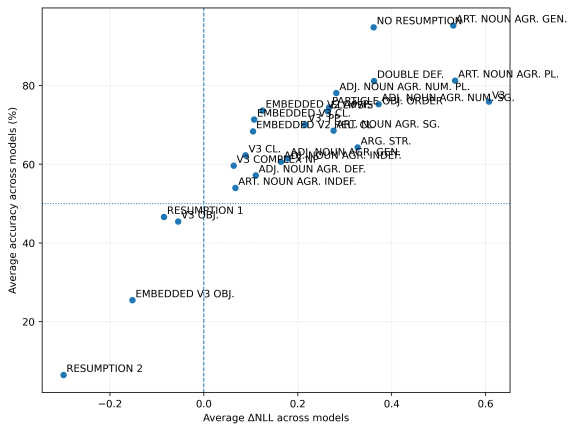

Figure 3: Average per-category confidence (in terms of $\Delta\mathrm{NLL}$) vs. average accuracy across models.

tematic failure. The unifying pattern seems to be simple: models prefer sentences with a resumptive pronoun. Because `NO RESUMPTION` is the only paradigm that requires a resumptive pronoun, performance is high there. In the other paradigms where resumption is disallowed, this same bias results in systematic errors.

At this stage it is unclear why the models prefer structures with resumptive pronouns. Resumptive structures are likely rare in Swedish, suggesting a frequency bias in the models. Another possible explanation for the problem with `RESUMPTION2` is that models have problems with missing arguments, as in relative clauses. In the sentence: "Lisa sitter i bilen som hon vet när hon lånade *-gap-* / *den." (*Lisa is sitting in the car that she knows when she borrowed / *it*), where the gap corresponds to *bilen*. But since the verb "lånade" (*borrowed*) is transitive, an ungrammatical resumptive in the empty position after the verb gives the sequence "när hon lånade den" (*when she borrowed it*), which in other contexts can be a well-formed adverbial clause: "Bilen gick sönder när hon lånade den." (*The car broke down when she borrowed it.*)

Given what we have just discussed, it is reasonable that without carefully investigating both the distribution of structures in the training data of the various models and local perplexity scores, any conclusions regarding the reasons behind model preference are uncertain. We leave such analysis for future research.

A comparison with English resumption (lacking in Warstadt et al. (2020)) would be informative. The type of resumption in `RESUMPTION2` is claimed to be possible in English (Hofmeister and Norcliffe, 2013).

Comparison of models. *Swedish Llama 8B* attains the highest macro average and *leads or ties* in *13/25* categories. *EuroLLM 9B* is the second-best model, leading in 5 categories. In contrast, larger general-purpose models (*Llama-3.3 70B*, *OSS 20B*) do not match these language-centric systems despite their size. These results suggest that language-specific pretraining can outweigh sheer scale when modelling syntactic competence.

However, the *GPT-SW3* models do not follow this pattern: despite their Swedish-centric pretraining, they yield the weakest performances. Yet, considering the prior reports of underperformance on Swedish benchmarks (Kurfalı et al., 2025), this may also point to some other issues with these models that may limit the benefits of language-specific pretraining.

Validation of the data construction. To verify (i) that template-generated items behave similarly to the hand-crafted seed pairs and do not introduce any kind of bias, and (ii) that the number of items per category is sufficient, we evaluated each model under four subset conditions per category: First-10 (the ten hand-crafted seed pairs), Random-10 (ten pairs sampled uniformly from the remaining ninety template-based items), Random-50 (fifty pairs sampled uniformly from the full set), and All (the full set of one hundred pairs).

The results are shown in Figure 1. Overall, performances are stable across subsets. The strongest systems show only small fluctuations (typically within a few percentage points) from First-10 to All, and model rankings are preserved. Accuracy on Random-50 is very close to All, which indicates that our data size is sufficient and that adding more examples would likely not change our conclusions. Similarly, comparing First-10 with Random-10 reveals no systematic "seed effect". The only notable deviation is *GPT-SW3 1.3B*; since this pattern is not observed across other models, it is more likely due to model-specific factors than an artifact of the dataset.

To quantify this observation, we compute the mean pairwise absolute deviation (MPAD), the average absolute accuracy difference across all $\binom{4}{2} = 6$ subset pairs for each model, and obtain a cross-model mean of 2.6 percentage points, which further reinforces the observation neither data construction nor sample size substantially affects performance.

Analysis of confidence. In addition to accuracy, we examine models' *confidence* on minimal pairs. For each minimal pair $\langle S_\checkmark, S_\times \rangle$ we define confidence as the difference in the negative log-probabilities:

$$\Delta\mathrm{NLL} \;=\; \ln\big(\mathrm{PPL}(S_\times)\big) - \ln\big(\mathrm{PPL}(S_\checkmark)\big),$$

– $\Delta\mathrm{NLL} > 0$ indicates that models assign higher probability to the well-formed sentence. Figure 2

summarizes these margins across models with one box per category. Each box shows where the middle half of models fall (25th–75th percentile; horizontal lines being the median). The whiskers extend to the smallest and largest values that are within $1.5\times$ the interquartile range from the box; and the circles beyond the whiskers are the outlier models.

Most medians in Figure 2 lie in the 0.2–0.6 range which corresponds to roughly $1.2\times$–$1.8\times$ higher perplexity, indicating a clear preference for the well-formed sentence. Box widths quantify how much different models differ within a category. Narrow boxes, such as in EMBEDDED V2 ADVP and NO RESUMPTION, indicate a strong agreement across models. In contrast, wider boxes, e.g., DOUBLE DEF., V2 PP, show a wide spread, suggesting that different models have different preferences.

Most notably, all models prefer the ill-formed alternative for RESUMPTION2, where the across-model median is around -0.2. This matches the extremely low accuracies for RESUMPTION2. A similar, yet less extreme, pattern appears in V3 OBJ. and EMBEDDED V3 OBJ., where medians are close to zero.

We show that the confidence aligns with accuracy at the category level. Figure 3 plots the average $\Delta\mathrm{NLL}$ against average accuracy per category. Categories with larger positive margins tend to achieve higher accuracy, while those with medians around 0 are at chance level. This indicates that $\Delta\mathrm{NLL}$ is a reliable proxy for syntactic competence and models' confidence is in line with their performance.

Comparison of evaluation strategies. Finally, we compare the three evaluation strategies: perplexity scoring and two prompting setups (grammaticality, acceptability).[4] Table 3 shows macro accuracy and Δ values, where positive Δ indicates improvement over perplexity. The results show that prompting largely degrades the performance. Models achieving high accuracies through perplexity-based evaluation (*SV-Llama-3 8B*, *EuroLLM 9B/1.7B*) lose 14–26 points under prompting. In contrast, some large general models (*Llama-3.3 70B*, *Gemma3 12B*) gain slightly with the *Grammar* prompt. Acceptability prompt, on the other hand, tends to underperform compared to the grammar prompt for the same model.

The findings echo the general observation that prompt-based judgments can diverge from likelihood-based preferences (Song et al., 2025; Hu et al., 2024) and highlight how evaluation strategy affects probing models' linguistic knowledge.

[4]Models with over 10% invalid responses were excluded.

Summary. The total number of minimal pairs in Swe-BLiMP (2,500) is considerably smaller than the average size of BLiMP-like benchmarks ($\sim$19k pairs), although the spread is large (SD $\sim$23k). However, as we have shown, increasing the number of pairs does not necessarily improve reliability. The number of phenomena (25) is close to the mean ($\sim$25) across benchmarks.

As for our findings, there are three things worth pointing out. Firstly, we expected model performance to drop as a function of long-distance agreement, in line with previous results (e.g., Warstadt et al., 2020; Someya and Oseki, 2023; Suijkerbuijk et al., 2025), despite exceptions to this pattern (e.g., van Schijndel et al., 2019). However, this was only true to some extent in our experiments. For example, while models *did* struggle on `EMBEDDED V3 OBJ.` with a model average of 25.5% ($SD = 21.7$), with `GPT-SWE` (20B) being somewhat of an outlier with an accuracy of 64.0%, most models also struggled with simpler structures, such as `ADJ. NOUN AGR. INDEF.`, with a model average of 54.0 % ($SD = 8.6$). Yet again, the seemingly strong bias toward resumptive pronouns across models indicates that surface patterns often override structural constraints.

Secondly, model size did not strongly correlate with accuracy; architecture and training domain appeared more important. An overrepresentation of English in large-scale pretraining corpora may bias models toward English representations (cf. Wendler et al., 2024), which could partly explain weaker performance on Swedish-specific phenomena.

Lastly, prompt-based evaluation mostly lowered performance compared to perplexity, though a few of the larger general models improved slightly with the grammar prompt. Model performance was otherwise stable across handcrafted and semi-automatically generated subsets, as well as across sample sizes.

In summary, going back to one of the first points we made, that is, that (L)LMs are by definition difficult to if not evaluate, then at least understand. Put differently, in our case, it is not easy without further manipulations of contrasting well-formed and ill-formed sentences to say what, exactly, causes which difficulties of specific cases for models. A closer investigation of the problems with word order, for example, seems like an important next step in the development of Swedish BLiMP. On the same note, other categories that should be included in future studies include quantifiers (Klingvall and Heinat, 2022b,a, 2024b; Heinat and Klingvall, 2020), negation (Klingvall and Heinat, 2024a), negative polarity items (NPIs) (Klingvall and Heinat, 2021) and island extractions. Contrary to English, extraction out of relative clauses is possible in Swedish (Heinat and Wiklund, 2015).

Model	PPL	Δ(Grammar)	Δ(Acceptability)
SV-Llama-3 8B	77.0	-25.9	-26.4
EuroLLM 9B	75.6	-14.3	-21.1
EuroLLM 1.7B	71.0	-20.6	-21.8
Llama-3.3 70B	69.7	+6.5	-4.0
Gemma-3 12B	69.2	+4.5	+3.8
Gemma-3 4B	66.0	+0.4	-1.4
Aya 32B	65.0	+0.6	-1.2
GPT-SW3 20B	53.0	-3.4	-4.1
Avg.	68.3	-6.5	-9.5

Table 3: Comparison of different evaluation strategies in terms of accuracy (%) on the dataset. Δ columns indicate performance difference between the prompt and the perplexity-based accuracy.

7. Conclusion and Future Work

We have introduced *The Swedish Benchmark of Linguistic Minimal Pairs* (Swe-BLiMP), a new resource for evaluating language model performance on Swedish syntax. The dataset covers 25 grammatical categories through 2,500 minimal pairs and was validated with human judgments, with participants preferring the well-formed sentence in 98.05% of cases. We tested eleven state-of-the-art models. The results showed, in general, that local dependencies could be captured by models, whereas some long-distance relations – particularly verb-second structures and resumption – were more difficult. Our analysis suggests that model size alone does not guarantee improved performance; rather, architecture and training domain appear to be more important. Performance was stable across handcrafted and semi-automatically generated subsets and across sample sizes, and prompt-based evaluation generally did not outperform perplexity. Future work will expand Swe-BLiMP with additional categories.

Limitations

The most important limitation of the study is that the number of phenomena (i.e., 25) is smaller than half of the original BLiMP paper (Warstadt et al., 2020). Addressing this is part of our future plans. As for the dataset itself, we have not investigated the distribution of naturalness across well-formed or ill-formed sentences. Additionally, while 100 sentence pairs per phenomenon might be viewed as a possible limitation of the study, we provide reasons why this is not necessarily the case.

Acknowledgements

We thank the 40 participants who took part in the acceptability judgment task. We also thank the three anonymous reviewers for constructive and helpful comments.

8. Bibliographical References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Arshia Asudeh. 2012. *The logic of pronominal resumption*. Oxford University Press, Oxford.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. Turblimp: A turkish benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2506.13487*.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Mar-

seille, France. European Language Resources Association.

David Beauchemin and Richard Khoury. 2025. Qfrcola: A quebec-french corpus of linguistic acceptability judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 119–130, November 4–9, 2025. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Kersti Börjars. 1994. Swedish double determination in a European typological perspective. *Nordic Journal of Linguistics*, 17(2):219–252.

Kersti Börjars. 1998. *Feature distribution in Swedish noun phrases*. Blackwell, Oxford.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Studies in Generative Grammar 9. Foris, Dordrecht.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An autoregressive language model for the Scandinavian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7886–7900, Torino, Italia. ELRA and ICCL.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. Glottolog 5.2.

Fredrik Heinat. 2021. The Swedish perfect and periphrasis. In Kristin Melum Eide and Marc Fryd, editors, *The Perfect Volume: Papers on the perfect*, Studies in Language Companion Series 217, chapter 14, pages 344–364. John Benjamins, Amsterdam.

Fredrik Heinat and Eva Klingvall. 2020. Set focus and anaphoric reference: An ERP study. *Brain and Language*, 206:104808.

Fredrik Heinat and Anna-Lena Wiklund. 2015. Scandinavian relative clause extractions: apparent restrictions. *Working Papers in Scandinavian syntax*, 94:36–50.

John Hewitt 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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Philip Hofmeister and Elisabeth Norcliffe. 2013. Does resumption facilitate sentence comprehension? In *The Core and the Periphery: Data-driven Perspectives on Syntax inspired by Ivan A Sag*. Centre for the Study of Language and Information.

Anders Holmberg and Christer Platzack. 2005. *The Scandinavian Languages*. Oxford University Press.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs.

Eva Klingvall and Fredrik Heinat. 2021. Negative properties of quantifiers in English and Swedish. *Studii de Lingvistică*, 11:149–166.

Eva Klingvall and Fredrik Heinat. 2022a. The effects of quantifier size on the construction of discourse models. *Journal of Neurolinguistics*, 63(101066):1–15.

Eva Klingvall and Fredrik Heinat. 2022b. Referential choices. A study on quantification and discourse prominence in sentence production in Swedish. *Journal of Pragmatics*, 193:122–138.

Eva Klingvall and Fredrik Heinat. 2024a. The effects of negation on discourse structure. *Journal of Pragmatics*, 235:115–131.

Eva Klingvall and Fredrik Heinat. 2024b. Lexical cues and discourse integration: an ERP study of the N400 and P600 components. *Cortex*, 178:91–103.

Artur Kulmizev and Joakim Nivre. 2022. Schrödinger's tree—on syntax and neural language models. *Frontiers in Artificial Intelligence*, 5:796788.

Murathan Kurfalı and Robert Östling. 2025. Conflicting needles in a haystack: How LLMs behave when faced with contradictory information. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Murathan Kurfalı, Shorouq Zahra, Evangelia Gogoulou, Luise Dürlich, Fredrik Carlsson, and Joakim Nivre. 2025. Swesat-1.0: The Swedish university entrance exam as a benchmark for Large Language Models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 331–339.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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)*, pages 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1618–1628, Beijing, China. Association for Computational Linguistics.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

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

Anne Lobeck. 2006. *Ellipsis in DP*, chapter 22. John Wiley & Sons, Ltd.

Stella Lundqvist. 2025. Do large language models and humans follow similar learning stages?: Assessing GPT-2's order of Swedish grammar acquisition within the Processability Theory framework. Master's thesis, Uppsala University.

Manfred Pienemann. 1998. *Language Processing and Second Language Development: Processability Theory*, volume 15. John Benjamins Publishing.

Taiga Someya and Yohei Oseki. 2023. Jblimp: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. 2025. Language models fail to introspect about their knowledge of language. *arXiv preprint arXiv:2503.07513*.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jon Sprouse and Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's core syntax1. *Journal of Linguistics*, 48(3):609–652.

Michelle Suijkerbuijk, Zoë Prins, Marianne De Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. Blimp-nl: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, 51(4):1267–1301.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Ulf Teleman, Staffan Hellberg, and Erik Andersson. 1999. *Svenska Akademiens Grammatik*. Norstedts Ordbok, Stockholm.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837, Hong Kong, China. Association for Computational Linguistics.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021. DaLAJ – a dataset for linguistic acceptability judgments for Swedish. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 28–37, Online. LiU Electronic Press.

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.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Yuhan Zhang, Edward Gibson, and Forrest Davis. 2023. Can language models be tricked by language illusions? easier with syntax, harder with semantics. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 1–14, Singapore. Association for Computational Linguistics.