Just drop the subject: A controlled rearing study of Subject Drop in English

Aug 15, 2025

Thomas Morton

> *…And thus by constantly hearing words, as they occurred in various sentences, I collected gradually for what they stood; and having broken in my mouth to these signs, I thereby gave utterance to my will.*
> (St. Augustine, *Confessions*, c. 400AD)

It's one of the primary goals of linguistics to look at complex phenomena in language and ask how children acquire such behavior. Along the way to adult performance, children exhibit linguistic behavior unlike the language they are learning. Before mastering some phenomena, children's language can exhibit behavior similar to other languages in content. The question becomes, do children exhibit linguistic behavior unlike their target language because they have yet to settle on the single target generalization, or that they have learned the target language but processing constraints shape linguistic outputs in ways that appear outside of distribution.

A classic case of this appears in the study of children's acquisition of non-null-subject languages. In these languages, like English, omitting a subject pronoun is unacceptable in nearly all cases, unlike Italian which allows for dropped subjects in many, if not most, cases. In Languages like Italian, children tend to preference towards omitting subjects rather quickly, the optional preference seeming fairly easy to learn. On the other hand, English children persist in omitting subjects (despite being illicit in their language) up until nearly three years old. Although English-speaking children drop their subjects at rates much lower than Italian-speaking children, the appearance of subject dropping in English children begs the question whether they content-fully believe they speak a language that allows for subject-dropping; or, despite knowing such a rule that dropped subjects are illicit in English, such artifacts appear as a result of their language processor. In this case a processor that omits subjects as a result of capacity constraints or economic principles overtaking grammatical principles.

There is a rich literature investigating both sides of this question, one side investigating the path that children taken in acquir-

ing grammatical generalizations and another that looks to explain the artifacts of early learning as processing effects. The study proposed here looks to ask these questions again, using Large Language Models (LLMs) as a new tool in the psycholinguistic arsenal for asking questions about the learning and processing of language. Specifically, we wish to investigate what role specific sources of linguistic evidence contribute to the direct and indirect learning of the English overt subject constraint.

LLMs are specifically well-served to ask these questions, as it is prohibitively difficult to know what a child's linguistic input looks like, or even more to manipulate the kind of input available to children while investigating their learning. In this way, we can manipulate the sources of information available to models and compare the causal effect that different sources of linguistic information (or their lack there-of) have on acquire linguistic generalizations.

Further, in asking questions of children's linguistic processing, exposing children to linguistic stimuli longitudinally to test changes in children's performance would surely introduce confounds of exposure and learning. Meanwhile, LLMs offer us the ability to sample a model's performance throughout training without influencing future performance of the model. Further, we can investigate the model on a wider range of evaluation stimuli without worrying about effects of fatigue.

The goal then, is to investigate large language models as candidate learners of the overt subject constraint and compare theories of learning and processing to see which theories best capture the performance of models when learning and processing sentences with and without subjects.

Statistical Language Models in Psycholinguistics

Jeffrey Elman's 1990 article 'Finding Structure in Time,' introduces the Recursive Neural Network (RNN) architecture. This architecture is the first to introduce a hidden layer within its architecture, that when trained on a task like next-word prediction develops task-specific weightings. In addition, in the RNN, the hidden layer maps temporal information (positional information of linguistic/items/tokens) over time onto the context. This allows for a model that is highly effective at finding solutions within the serial ordering of tokens (they include nonlinguistic tasks as

well), but also creating solutions that are sensitive to temporal (positional) structure across the context, which creates a network more capable of maintaining long-distance dependencies. In this work, Elman finds within the representational space of the word embeddings groupings structure that seem to align with intuitions of superordinate and subordinate lexical heirarchy, as well as animate/inanimate splits, among others. He suggests that connectionist models of this kind are capable of holding rich linguistic representations that it can derive from the data. He states, "One of the things which feedforward PDP models have shown is that simple networks are capable of discovering interesting internal representations of many tasks ... and representations in tasks which unfold over time."

Elman's [1] RNN model and Rumelhart and McCleland's [2] connectionist program brought a new perspective to looking at language modeling. Such neural network approaches became very important tools to creating interpretable statistical models of language learning and language processing [3–5]. However, despite the effectiveness of such approaches, they still struggle to overcome simple n-gram or bag-of-word approaches to language modeling [6, 7].

The overwhelming success of Large Language Models (LLMs), in the form of causal and bi-direction autoregressive Transformers, like BERT[8] and GPT [9], has revolutionized the field of computational psycholinguistics. LLMs seem capable of acquiring linguistic generalizations in emergent, robust, and surprising ways [10, 11].

Despite the fact that transformers are general enough in their learning, that building a model to learn language doesn't require engaging with linguistic theory. In some cases, authors have proposed that linguistic theory shouldn't play a role in our investigation of LLMs[12]. However, many still argue that traditional analyses of language should still inform how we investigate LLMs, and that even more, LLMs can be used as tools to investigate human learning and cognition [13–15]. And work since the emergence of Transformers has continued to integrate linguistic theory into the study of LLMs to better understand human behavior [15–25]

We can use LLMs as tools to investigate learning, but the question remains how to link human learning and model learning — if they are to serve as candidate learners. One way of thinking about this problem is in terms how humans and LLMs may both converge towards similar representations, and what that means for both's learning. That is, maybe humans and LLMs converge

[2]: Rumelhart et al. (1986), *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations*

[3]: Levelt et al. (1999), "A theory of lexical access in speech production"

[4]: Chang (2002), "Symbolically speaking: a connectionist model of sentence production"

[5]: Chang et al. (2006), "Becoming syntactic"

[6]: Wang et al. (2012), "Baselines and bigrams: Simple, good sentiment and topic classification"

[7]: Arora et al. (2017), "A simple but tough-to-beat baseline for sentence embeddings"

[8]: Tenney et al. (2019), "BERT rediscovers the classical NLP pipeline"

[10]: Manning et al. (2020), "Emergent linguistic structure in artificial neural networks trained by self-supervision"

[11]: Potts (2025), *Finding linguistic structure in large language models*

[12]: Piantadosi (2024), "Modern language models refute Chomsky's approach to language"

towards similar, simple representations of grammatical theories. Biased learners are at times necessary to make new inductions from data [26]. One way that a learner can be biased is to be biased towards a reductive center, by looking to form simplest explanations for problems. One can think that introducing explicit architectures like the kind found in early computational linguistics models might be one way towards this goal [27, 28]. However recent work investigating the learning of neural networks shows that, counter to intuition, less restrictive networks tend towards simpler solutions [29–35]

These simplicity-first ideas are not new in the study of linguistics or psycholinguistics. In some areas of linguistics, 'minimalist,' innate representations and operations make up simple-as-possible combinatorial systems [36]. Under the views of construction grammars, linguistic structures reduce through use into idiomatic phrases which are iteratively and recursively constructed into a usage-based representational system of language [37]. In psycholinguistics, work on the shape of language production proposes that one's one preference to maintain short distances between dependencies as a pressure both necessitates optimal linguistic representations in use, but also introduces such biases into linguistic data that can be processed and guide processing [38]. As well as computation models that explicitly introduce such parameters [39–42]

This also helps to account for the fact that across many different training contexts, model configurations, languages, etc. LLMs seem to converge towards similar representations between each other. This has become a popular view of learning in LLMs called the 'Platonic Representational Hypothesis. [43]' Essentially, content that models can learn on are inherently rich with information, this information leads general learning models towards solutions that are similarly simple. Work on LLMs as they exist now fairly agree that many of the representations that LLMs demonstrate are similar to their human counterparts [10, 44–47]. Modern techniques investigate LLMs using similar measures as human participants [25, 48, 49].

[26]: Mitchell (1980), *The Need for Biases in Learning Generalizations*

[29]: Valle-Pérez et al. (2018), "Deep learning generalizes because the parameter-function map is biased towards simple functions"

[30]: Belkin et al. (2019), "Reconciling modern machine-learning practice and the classical bias-variance trade-off"

[31]: Zhang et al. (2021), "Understanding deep learning (still) requires rethinking generalization"

[32]: Henighan et al. (2023), *Superposition, Memorization, and Double Descent*

[33]: Attias et al. (2024), "Information complexity of stochastic convex optimization: Applications to generalization and memorization"

[34]: Maloney et al. (2022), "A solvable model of neural scaling laws"

[35]: Goyal et al. (2022), "Inductive biases for deep learning of higher-level cognition"

[43]: Huh et al. (2024), "The platonic representation hypothesis"

Similarly, modern studies investigating LLMs as candidate learners of language demonstrate that LLMs utilize direct and indirect sources in human-like ways [50–55]. It is important to ask questions about learning with LLMs, as they are currently at the center around discussions about what is learnable from linguistic input and what is not via general statistical learning system [12, 56–61].

This study seeks to use LLMs as tools to investigate human language learning, under the thesis that the kinds of information available for human learners to acquire specific theories of linguistic grammar are also available to LLMs as evident in their success at a broad range of linguistic tasks. Our goal is to use such models as candidate learners to investigate how linguistic information guides models in a causal way towards developing rules. Our hope is that these models can give us insight into how human learners utilize linguistic information available to them in their environment, and to ask also what is not available in their environment that might bias language users to certain generalizations. We do not express that language learners and LLMs are the same, but that we can see both as fairly general learners within similar problems spaces, and that investigating one can give us insights into the other.

[50]: Jumelet et al. (2021), "Language models use monotonicity to assess NPI licensing"
[51]: Ahuja et al. (2024), "Learning syntax without planting trees: Understanding hierarchical generalization in transformers"
[52]: Patil et al. (2024), "Filtered Corpus Training (FiCT) shows that language models can generalize from indirect evidence"
[53]: Misra et al. (2024), "Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs"
[54]: Feng et al. (2024), "Is child-directed speech effective training data for language models?"
[55]: Yao et al. (2025), "Both direct and indirect evidence contribute to dative alternation preferences in language models"

Accounts of subject drop

Corpus work has widely attested subject drop in English children's speech[62, 63]. Examples like:

(1) Shake hands.
Turn light off.
Want go get it.
Show mommy that.
Now making muffins

Bloom [62] put forward the claim that English children drop their subjects for performance reason. Bloom more subjects were dropped in sentences with negation. Under this account, when children encounter contexts with heavy cognitive load, they are prone to omit information that may be less important or otherwise contextually recoverable. Evidence for such cases come from cases where children from even very early ages, before when they could have possibly acquired such a rule, show distributional features of their target language. For example Valian [64] reports that English children still produce substantially

[62]: Bloom (1970), *Language Development*

[64]: Valian (1991), "Syntactic subjects in the early speech of American and Italian children"

more overt subjects than Italian children before averaging over a mean length utterance of 2. One caveat to this account is the asymmetry between subject and object pronouns, such that subject pronouns are much more likely to be dropped than object pronouns.

[65] proposes that this can be contended for if such processing asymmetries can be found in orthogonal contexts. For example, longer names are omitted more than shorter names, across both the subject and object positions. The case isn't necessarily that such accounts are not about learning, except that there is some learning process that is occurring during children's development that interacts with this processing element and that accounts of this effect that are purely grammatical in explanation lose out explanatorily. The idea that our linguistic output is determined by the interplay between resource-limited processors guided by more abstract representations is not a new one [66].

[65]: Bloom (1990), "Subjectlees sentences in child language"

[66]: Bock (1982), *Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation*

They propose that there must be some particular processing difficulty in speaking a subject as compared to speaking the object. This kind of account is criticized by Hyams and Wexler [67, 68] argue that there is not sufficient evidence to suggest that there should be a difference in the processing of a subject that should lead to such start assymetries between the subject and object position. They suggest that children are not considering grammars where objects can be dropped, but they are where subjects can be dropped, and so this is an artifact of their competence, not performance.

For example, one assumption that must be made by a performance account allowing for this asymmetry is subjects are inherently more difficult to produce than objects, whereas work in the area of sentence production would suggest that in-fact the subject should require less work to process, because it can be planned separately from the verb [69]. Further, initiating speech altogether requires fairly little planning, with some experimental evidence suggesting that adults can begin speaking before even fully planning the first word [70, 71]. Further, work by McDaniel [72] on the development of children's language planning suggests that children restart more in the lower half of the sentence than the first half, counter to P. Bloom's [65] predictions.

[69]: Momma et al. (2018), "Unaccusativity in sentence production"

Hyams et. al. [**yams1993-zk**, 67, 73] proposed an alternative account to this phenomena. Hyams claimed that this asymmetry suggests that children's behavior in these cases are reflective of children's learning of grammatical rules. Their work falls under the popular of-the-time principles and parameters framework in generative linguistics (see [74], c.f. [75]). Under this account, chil-

[73]: Hyams (1986), *Language acquisition and the theory of parameters*

[67]: Hyams (1989), "The null subject parameter in language acquisition"

[**yams1993-zk**]: **yams1993-zk** (**yams1993-zk**), **yams1993-zk**

dren initially set the null-subject parameter to the positive value (allowing null subjects), and must learn from positive evidence in the input that their language requires overt subjects. Specifically, Hyams' Triggering theory [68] proposes that it is the non-uniform nature of English verbal morphology that serves as the crucial trigger for resetting this parameter. Languages with uniform verbal agreement (either consistently rich like Italian, or consistently poor like Mandarin) allow null subjects, while English's inconsistent system triggers the obligatory subject requirement.

Building on parametric approaches, Yang [76, 77] developed the Variational Learning theory, which provides a probabilistic account of parameter setting. Under this theory, children entertain multiple grammatical hypotheses simultaneously and update their probabilities based on input frequency. Yang argues that expletive subjects (like "it" and "there") serve as the critical unambiguous evidence for the [-null subject] parameter in English. The relative rarity of expletives in child-directed speech explains why English-learning children take longer to converge on the adult grammar compared to children learning null-subject languages.

[76]: Yang (2003), *Knowledge and learning in natural language*
[77]: Yang (2004), "Universal Grammar, statistics or both?"

More recently, Duguine [78] proposed an Inverse approach that shifts focus from verbal morphology to the nominal domain. This account suggests that the crucial evidence comes from the interaction between determiner richness and verbal agreement weakness, among other factors. In Duguine's framework, a rich determiner system combined with weak verbal agreement (as in English) provides indirect evidence against null subjects, while languages with poor determiner systems or rich verbal agreement allow subject drop.

[78]: Duguine (2017), "Reversing the approach to null subjects: A perspective from language acquisition"

Bertolino [79] extends this line of reasoning by examining the role of bare singular count nouns as potential evidence for partial subject drop, suggesting that even subtle distributional patterns in the input may influence children's hypotheses about their target grammar.

Despite decades of research, the debate between performance-based and competence-based accounts remains unresolved. A key challenge has been the difficulty of manipulating children's linguistic input to test causal hypotheses about what evidence drives the acquisition of the overt subject constraint. The current study addresses this limitation by using Large Language Models as experimental models of language acquisition, allowing us to systematically manipulate different sources of linguistic evidence and measure their causal contribution to learning the English subject requirement.

Methods

Materials

Large Language Models will be trained on the BabyLM dataset [80, 81]. The BabyLM dataset is a 100 million word corpus designed to train human-sized models on a linguistically diverse sample which includes a larger-than-average proportion of child-directed speech. The corpus is sized roughly to model the linguistic input of a ten to fourteen year old child. Seperate from the 100 million word training corpus, a 10 million word test set is held out to test the model's memorization of the dataset.

[80]: Warstadt et al. (2023), "Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora"
[81]: Warstadt (2022), "Artificial neural networks as models of human language acquisition"

Table 1: Word counts for the strict track of the 2nd BabyLM Challenge.

Dataset	Description	# Words (strict track)
CHILDES	Child-directed speech	29M
British National Corpus (BNC), dialogue portion	Dialogue	8M
Project Gutenberg (children's stories)	Written English	26M
OpenSubtitles	Movie subtitles	20M
Simple English Wikipedia	Written Simple English	15M
Switchboard Dialog Act Corpus	Dialogue	1M
Total		**100M**

(2) *Third person singular and plural (English is a non-pro-drop language)*
   a. Anna finished the book. She/*Ø thinks the ending is perfect.
   b. The clients saw the proposal. They/*Ø think the budget is acceptable.

(3) *Second person singular and plural*
   a. Marco, you read the email. You/*Ø think we need more time.
   b. Students, you heard the news. You all/*Ø think the decision is fair.

(4) *First person singular and plural*
   a. I reviewed the agenda. I/*Ø think the schedule is too tight.
   b. My team and I saw the demo. We/*Ø think the product has potential.

(5) *Subject and Object Control (PRO in non-finite clauses)*
   a. Maria convinced her brother Ø/*him to leave the party early.
   b. The director promised the actors Ø/*he to revise the script.

(6) *Expletive constructions*
   a. *Ø/It seems that the students passed the exam easily.

(7) *Distant antecedent in embedded finite clauses*
   a. The waiter mentioned that *Ø/he had waited over an hour.

(8) *Coordinate structures with and without topic shift*

a. Giovanni woke up late and Ø/he missed the train completely.

b. Anna called Mark and *Ø/he refused to answer her questions.

Stimulus Manipulations for Testing Processing Accounts In addition to the ablative interventions that test grammatical-learning theories, we will manipulate features of the evaluation stimuli to test processing-based accounts of subject drop. These manipulations allow us to investigate whether models exhibit the same processing constraints that have been proposed to explain children's early subject omissions, following work by Michaelov and Bergen [82] on processing biases in language models.

Each evaluation stimulus pair will be tested under multiple processing conditions:

[82]: Michaelov et al. (2022), "Do language models make human-like predictions about the coreferents of Italian anaphoric zero pronouns?"

Context Complexity Manipulation To test whether increased processing load leads to more subject drop preferences (as predicted by Bloom 1990), we will vary the complexity of the context preceding our target sentences:

(9) *Simple Context*

a. The dog barked. He/*Ø scared the mailman away.

(10) *Complex Context (longer NPs)*

a. The large brown dog with the red collar barked. He/*Ø scared the mailman away.

(11) *Complex Context (embedded clauses)*

a. The dog that lived in the house at the end of the street barked. He/*Ø scared the mailman away.

Negation Manipulation Following Bloom's 1970 observation that negation increases subject drop in child speech, we will test whether negation in either the target sentence or context affects subject realization:

(12) *Target Sentence Negation*

a. Anna finished the book. She/*Ø doesn't think the ending is perfect.

b. The clients saw the proposal. They/*Ø don't think the budget is acceptable.

(13) *Context Sentence Negation*

a. Anna didn't finish the book. She/*Ø thinks the ending is perfect.

b. The clients didn't see the proposal. They/*Ø think the budget is acceptable.

(14) *Double Negation*

a. Anna didn't finish the book. She/*Ø doesn't think the ending is perfect.

Im total, for each item group (2-8), 12 base pairs, a preferred and unprepared sentences are constructed. Each sentence consists of a context sentence, and a target sentence. This made 13 different item groups and to a total of 153 language pairs. Sentences were generated using Deepseek AI to generate pairs, and were hand-checked by the author as a native English speaker. Further, each of the 153 languages pairs was manipulated with 5 processing manipulations, again, the pairs were run through Deepseek's transformer model to generate the proper manipulations. These were then checked and edited by hand by the author. In total, 918 sentence pairs were constructed across the 13 item groups.

Ablative Interventions

In this study, we will use experimental ablation interventions on LLM training corpora in order to derive the causal role that individual linguistic evidence has on learning [50–55, 83]. Each of these ablative techniques are designed to alter the English dataset to make it like a language unlike English. During and after training, performance is assessed on evaluation stimuli designed to target knowledge of grammatical constructions involved in preferences for null and overt subjects, expletives, and determiner morphology.

Following other studies [53, 55, e.g.], we will perform ablations before training by breaking up the training corpora into sentences using the spaCy [84] library POS and sentence parser. From there, each ablative task has a specific method of performing the target ablation. After ablation, a subset of modified sentences will be checked by the researcher to ensure that the implementation is correct. For interventions where words are being removed, an appropriate amount of additional stimuli will be added back to the training set to allow for equal amounts of training tokens for each model. Those additional stimuli will be intervened on and the process will be repeated until a complete dataset is constructed for each model.

[53]: Misra et al. (2024), "Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs"
[55]: Yao et al. (2025), "Both direct and indirect evidence contribute to dative alternation preferences in language models"
[50]: Jumelet et al. (2021), "Language models use monotonicity to assess NPI licensing"
[54]: Feng et al. (2024), "Is child-directed speech effective training data for language models?"
[51]: Ahuja et al. (2024), "Learning syntax without planting trees: Understanding hierarchical generalization in transformers"
[52]: Patil et al. (2024), "Filtered Corpus Training (FiCT) shows that language models can generalize from indirect evidence"
[83]: Leong et al. (2023), "Language models can learn exceptions to syntactic rules"

No Expletives

Pleonastic subjects, or Expletives, like *it* or *there* are required in certain contexts where a clause lacks an appropriate subject, but one is nonetheless required. In this case, a dummy pronoun that has no direct reference is required. We will use the SpaCy

Table 2: Ablative Controlled Rearing Study Design

Ablation	Un-ablated Example	Modified Example
No Expletives	It is raining and there is a puddle on the street.	Is raining and a puddle is on the street.
Poor Determiner Morphology	Some people saw the one car and a truck.	The people saw the the car and the truck.
No Articles	A dog chased the cat up the tree.	Dog chased cat up tree.
Infinitive Verbal Morphology	She walked to the store because he is driving.	She walk to the store because he be drive.
No Spoken Pronominal Subjects	He went to the park after she finished the work.	Went to the park after finished the work.

POS parser, which very accurately marks expletive pronouns, to detect expletives in context, see Procedure.

Procedure 1: FindDummyPronouns(*corpus*)

```
1.  Load SpaCy NLP model with a dependency parser
2.  Initialize D ← an empty list for dummy pronouns
3.  for each sentence in corpus do
4.    doc ← process(sentence, NLP model)
5.    for each token in doc do
6.      if token.dep_label = 'expl' and token.head.pos_tag = 'VERB' then
7.        Add token to D
8.      end if
9.    end for
10. end for
11. return D
end procedure
```

Then, in order to be sure that those pronouns are not in fact referential, we will use SpaCy's experimental coreferrant component to determine whether, given the now detected expletives, if they belong to a reference cluster, if so, we do not omit them as they may not be empty subjects, but if they do, we ablate them. We will use the previous two sentences of the detected dummy words to determine whether they have potential referential properties.

Procedure 2: ConfirmNonReferential(*corpus*)

```
Load SpaCy NLP model with parser and coreference resolver
Initialize D_confirmed ← an empty list
potential_dummies ← FindDummyPronouns(corpus)    // Call Procedure 1
for each token in potential_dummies do
    context ← sentence containing token + preceding sentence
    doc ← process(context, NLP model)
    clusters ← doc.coreference_clusters
    has_referent ← False
    for each cluster in clusters do
        if token is in cluster then
            has_referent ← True
            break
        end if
    end for
    if not has_referent then
        Add token to D_confirmed
    end if
end for
return D_confirmed
end procedure
```

Poor Determiner Morphology

English itself has fairly poor determiner morphology, in the sense that it contains little information about the nominal features of its associated noun. For instance, English lacks the kind of gender concord found in gendered language, or markings for plurality, instead largely marking definiteness. We propose an ablation that removes all further richness from the determiner morphology, marking all determiners as 'the.' Very simply we will use SpaCy to find all determiners, and replace them with a single token 'the'.

Procedure 3: ImpovershDeterminers(*text*)

```
Load spaCy NLP model with a POS tagger
Initialize modified_parts ← an empty list
doc ← process(text, NLP model)
for each token in doc do
  if token.pos_ = 'DET' then
    append 'the' to modified_parts
  else
    append token.text to modified_parts
  end if
end for
result ← join_with_spaces(modified_parts)
return result
end procedure
```

No Articles

Some languages lack articles altogether. English finds determiners optional in circumstances such as with plural subjects and mass nouns. This intervention finds all definite and indefinite basic articles such as 'a' or 'the' and removes them entirely. This leaves articles like 'some,' 'all,' 'these,' etc. but those remain in much lower frequency. In this case SpaCy uses a POS tagger to find all tokens marked as determiners and removes basic determiners in the modified corpus[1].

1: While this is a fairly rough cutting of fairly basic parts of the corpus, you could potentially run a similar intervention on a smaller subset of linguistic content. Some work suggests specifically bare singular count nouns could be evidence for learners that their language allows for partial subject drop [79].

Procedure 4: RemoveArticles(`text`)

```
Load spaCy NLP model with a POS tagger
Initialize modified_parts ← an empty list
doc ← process(text, NLP model)
for each token in doc do
  is_article ← token.pos_ = 'DET' and token.lower_ in ['a', 'an', 'the']
  if not is_article then
    append token.text_with_ws to modified_parts
  end if
end for
result ← join(modified_parts)
return result
end procedure
```

Infinitival verbs

In English, in addition to subject plural marking on the verb, some tenses and aspect are marked on the verb while others are marked via modals. However, there is fairly poor morphology when it comes to marking other aspects of nominal features, for instance, person is not marked on nouns. Some theories of subject dropping predict that it is *consistent* morphology that allows for subject dropping, and not necessarily only rich morphology [68]. So, while we could try to modify the corpus such that English has rich agreement morphology, in which case we would do our best to extract the feature space of the subject and artificially mark the verb with additional person marking, we choose to instead remove all rich morphology on the verb, using SpaCy's POS tagger, which includes identification of word lemmas to convert verbs to their infinitival form.

[68]: Hyams et al. (1993), "On the grammatical basis of null subjects in child language"

Procedure 5: LemmatizeVerbs(`text`)

```
Load spaCy NLP model with POS tagger and lemmatizer
Initialize modified_parts ← an empty list
doc ← process(text, NLP model)
for each token in doc do
  if token.pos_ = 'VERB' then
    append token.lemma_ to modified_parts
  else
    append token.text to modified_parts
  end if
end for
result ← join_with_spaces(modified_parts)
return result
end procedure
```

No Subject Pronominals

Finally, while we have previously attended to fairly indirect evidence for subject-drop, in this case, we specifically target direct evidence for subject-drop, which is the presence of subject pronouns in the dataset in the subject position. In this case, we use SpaCy across sentences to first annotate parts of speech on each token, then we parse it with a dependency parser to determiner basic syntactic roles of words in a sentence. In this case, we remove all pronouns that are acting as nominal subjects. This also creates evidence for the learner that explicit subjects in general are not neccessary.

Procedure 6: RemoveSubjectPronominals(`text`)

```
Load spaCy NLP model with POS tagger and dependency parser
Initialize modified_parts ← an empty list
doc ← process(text, NLP model)
for each token in doc do
    is_subj_pronoun ← token.pos_ = 'PRON' and token.dep_ = 'nsubj'
    if not is_subj_pronoun then
        append token.text_with_ws to modified_parts
    end if
end for
result ← join(modified_parts)
return result
end procedure
```

Planned Experiments

Table 3: Experimental Design for Ablation Studies

Exp.	No Expletives	Poor Determiner Morphology	No Articles	Infinitive Verbal Morphology	No Pronominal Subjects
1	✓	x	x	x	x
2	✓	✓	x	x	x
3	✓	x	✓	x	x
4	✓	x	x	✓	x
5	✓	x	x	x	✓

The selected experiments have been chosen to both, individually test the causal contribution of each piece of evidence as completely as possible while also training an appropriate and viable number of models[2]. In addition to testing these ablative techniques in isolation, the ablative interventions have been chosen to assess specific theories of linguistic theory proposed in the literature, each of which predict specific contributions of each of linguistic evidence of different kinds and combination.

Charles Yang's [76, 77, 85] Variational Learning theory predicts that expletives are the key evidence that children use to acquire obligatory null subjects, and that the small portion that these tokens make up of the linguistic input is the reason why English children take much longer to acquire a strict generalization of obligatory subjects. Experiment 1, in this case tests this theory by only removing expletive subjects and testing the model's ability to acquire such a generalization.

Other primary theories in this field propose that expletive subjects are crucial to the learning of such structures, including

2: Maybe we end up being more thorough with the combinatorial choices later on when there's more time 🤦

[76]: Yang (2003), *Knowledge and learning in natural language*
[77]: Yang (2004), "Universal Grammar, statistics or both?"
[85]: Yang et al. (2011), *Minimalism and Language Acquisition*

Hyam's [68] Triggering theory of parameterization in learning and Duguine's [78] Inverse approach. Duguine's [78] approach specifically implicates a richer determiner system in combination with weak verbal agreement. Whereas Hyam's [68] theory suggests that it is a non-uniform verbal system that primarily contributes this generalization. In Hyam's account, a language like Italian is uniformly marked for gender and person, while Mandarin is uniformly unmarked, both of which contribute to languages that allow for null-subjects, whereas English's inconsistency triggers the opposite generalization. Experiment 2/3 captures a case where only determiner information is insufficient, which should impair learning under Duquine's theory but not Hyam's. Hyam would predict that in Experiment 4, there should be insufficient evidence to acquire the null-subject parameter; whereas Duquine would point to a rich determiner system still present with a weak (but still uniform verbal system) providing sufficient evidence for an English-like generalization.

Likewise, while Duguine's [78] theory points primarily to indirect evidence in acquiring such a rule, Hyam's [68] Theory states directly that the use of an overt pronoun is positive evidence for a non-null-subject language under the assumption that speakers avoid the use of overt pronouns in non-discourse-necessary, non-emphatic contexts. Experiment 5 seeks to assess the role of direct evidence in ablating all pronominal subjects on learning an English-like generalization.

I've summarizes the points made in this paragraph in Table 4.

[68]: Hyams et al. (1993), "On the grammatical basis of null subjects in child language"

[78]: Duguine (2017), "Reversing the approach to null subjects: A perspective from language acquisition"

Table 4: Predictions and Contributions of Ablation Experiments

Exp.	Primary Theory Tested	Predicted Outcome for Learning Obligatory Subjects	Contribution to Understanding
1	Yang	Learning should be significantly impaired.	Tests the causal role of expletives as the key (albeit rare) evidence for the English rule.
2 & 3	Duguine vs. Hyams	**Duguine:** Learning should be impaired, as a key piece of evidence (rich determiner system) is removed. **Hyams:** Learning should be largely unaffected, as the primary trigger (non-uniform verbal morphology) remains intact.	Differentiates theories that prioritize determiner morphology (Duguine/BCC) from those that prioritize verbal morphology (Hyams).
4	Hyams vs. Duguine (Inverse/BCC)	**Hyams:** Learning should be severely impaired or fail, as the main trigger (non-uniform verbs) is removed. **Duguine:** Learning should still succeed, as the crucial evidence (rich determiner system) remains.	Acts as the inverse of Exp. 2 & 3, testing whether verbal morphology (Hyams) or determiner morphology (Duguine/BCC) is the critical input.
5	Hyams (Direct Evidence)	Learning should be impaired, as the direct positive evidence of hearing overt pronouns is ablated.	Assesses the role of direct evidence (hearing overt pronouns) versus the indirect evidence favored by other theories.

Training Procedure

SentencePiece [86] tokenizers are trained on the datasets (base or ablated) and the training sets, making for a total of seven trained tokenizers. Those tokenizers are then used to tokenize the datasets, and the datasets are prepared for training by grouping text into lines of 1000 tokens to maximize training efficiency and consistency across steps. The model parameters are in Table 5.

[86]: Kudo et al. (2018), "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing"

Models are initialized as empty GPT2 transformers [9] with weights randomly initialized based on a random seed (controlled for between experiments). Checkpoints are saved regularly during training: during the first epoch, with a checkpoint saved at the end of each epoch. A final checkpoint is saved at the end of training. The base model, as with all other models were saved and evaluated over the course of 134 total checkpoint steps. 20 checkpoints were saved in the first epoch, and 6 were saved in each subsequent epoch. Each model used AdamW as the learning scheduler and optimizer. Each model was configured to use Flash-Attention 2 to save time and memory while training the models. Further the models used AMP to handle mixed precision during training (a step not usually paired with flash-attention but necessary to prevent catastrophic forgetting effects and instability when using float16 operations — these models used float32 operations). Gradient checkpointing was used to further save memory, and increase the available batch size.

[9]: Radford et al. (2019), "Language Models are Unsupervised Multitask Learners"

Hyperparameter	Value
Layers	12
Embedding size	768
Hidden size	768
Intermediate hidden size	3072
Attention heads	12
Attention head size	?
Activation function	gelu
Vocab size	50004
Max sequence length	1000
Position embedding	?
Batch size	256
Train steps	?
Learning rate decay	?
Warmup steps	?
Learning rate	?
Adam ϵ	?
Adam β_1	?
Adam β_2	?
Dropout	?
Attention dropout	?

Table 5: Language model hyperparameters.

Each model is trained for 20 epochs. Models are analyzed throughout the first epoch and at the end, and across time over the remaining 19 epochs. While this is a study looking at developmentally plausible models, large language models often require training unlike human learners, and so to give the model its best shot at learning the correct generalizations despite any ablative work, we choose to look across all epochs while considering learning from the first epoch separately.

Each model was trained on an NVIDIA RTX A6000 as part of the Psychology Department computing cluster, each model is trained over roughly two GPU hours.

Measures and Analysis

Data and Coding

The outcome encodes preference as a binary response where `correct` = 1 indicates the *null* realization has lower surprisal

than the *overt* realization. Factors include `model` (six experimental conditions), `form_type` (*null*, *overt*), `item_group` (linguistic subfamilies), `form` (processing/manipulation types), and `item_id` (random-effect grouping). Baseline contrasts are enforced by releveling `model` so that the Baseline condition is the reference level. All models use a $\log_{10}$ transformation of training progress, $\log_{10}(\texttt{checkpoint\_num} + 1)$, to capture log-learning dynamics.

Stimuli, Ablations, and Training

Evaluation items are constructed in minimal pairs that differ only in subject realization (*null* vs. *overt*), with lexical and contextual content otherwise held constant. Training corpora are ablated per experiment (e.g., removal of expletives and articles; removal of expletives with verb lemmatization; removal of expletives and subject pronominals). Models are trained under identical optimization settings and checkpointed uniformly. Training progress is analyzed on a $\log_{10}$ scale with ticks at $\{0, 10, 100, 1\text{K}, 10\text{K}\}$ to reflect log-learning dynamics.

Outcome Definition

For each pair at each checkpoint, mean surprisal is computed for the *null* and *overt* realization. A binary response $\text{Y} \in \{0, 1\}$ encodes a preference for the target realization (lower surprisal). Unless stated otherwise, end-state summaries report preference for the *overt* realization on the probability scale; acquisition-time analyses operate on preference for the *null* realization. Item identity is included as a random factor to account for repeated measures.

Learning Curves and Spline Selection

Learning curves are estimated with generalized linear mixed-effects models (GLMMs; logit link):

$$\operatorname{logit} \Pr(\text{Y} = 1) = \beta_0 + \text{ns}(\log_{10}(\text{t} + 1),\ \text{k}) + \text{u}_\text{i}.$$

where $\text{ns}(\cdot)$ is a natural spline over log-checkpoint and $\text{u}_{\texttt{item}} \sim \mathcal{N}(0, \sigma^2)$. Spline complexity is selected per model by AIC over $\text{K} \in \{3, \ldots, 7\}$; the lowest-AIC converged fit is retained for inference and figures. Baseline contrasts are enforced by releveling the `model` factor to set the Baseline as reference.

Acquisition-Time Metrics

Two complementary metrics quantify when *null*-subject behavior emerges.

t_{50} **(chance-level acquisition).** For each model, the fitted probability of *null* preference is evaluated across checkpoints; t_{50} is defined as the *last* crossing of 0.50 following a burn-in at $\texttt{checkpoint} \geq 100$. Crossings are located by linear interpolation between adjacent fitted points. If no crossing occurs, the estimate is treated as right-censored. Uncertainty is quantified via parametric bootstrap of the fitted GLMM (fixed-effects uncertainty; `re.form` = NA), typically with $n = 500$ draws, reporting percentile 95% confidence intervals (CIs). Measure derived from

AoA$_{1/2}$ **(halfway-to-asymptote).** End-state performance p_∞ is estimated from the *last 10%* of training. A dynamic threshold $\theta = (p_\infty + 0.5)/2$ defines *Age of Acquisition* as the first post–burn-in crossing of θ. CIs use the same parametric bootstrap. Between-model differences are summarized as paired-bootstrap ΔAoA$_{1/2}$ relative to Baseline, with empirical p-values given by the proportion of paired draws at or beyond zero in the hypothesized direction.

First-Epoch Analysis Early learning is assessed by summarizing across all checkpoints of the first epoch (operationalized as $\max(\texttt{checkpoint})/20$). For each condition, the mean preference and 95% CI are reported and tested against chance (0.50) using exact binomial tests. When comparing conditions within the first epoch, odds ratios (ORs) with 95% CIs, Wald z, and adjusted p-values are reported.

End-State Analyses (Final 10% of Training) All end-state models use only the last 10% of checkpoints and include a random intercept for item.

Model-level preferences. Estimated marginal means (EMMs) on the probability scale are reported for each model. Pairwise model comparisons are expressed as *odds ratios* (OR) with standard errors (SE), Wald z, and multiplicity-adjusted p-values. Corresponding probabilities with 95% CIs are provided for interpretability.

Item-group effects. Person/number/control/expletive/topic-shift subfamilies are modeled via fixed effects for `item_group` and their interactions with `model`. Within each model, EMMs (probability + 95% CI) are reported alongside all pairwise contrasts as OR (SE, z, adjusted p). In cases of quasi-complete or perfect separation where mixed-effects models cannot converge, inference defaults to Fisher's exact tests with exact p-values.

Processing/structure manipulations. Processing forms (e.g., long NPs, embedded relatives) and negation families (context, target, both) are analyzed using fixed effects for `form` and `form`×`model`. Within-model pairwise contrasts are reported as ORs with 95% CIs and adjusted p-values. Where "no difference from default" is stated, ORs are near 1 with CIs spanning 1.00.

Multiple Comparisons and Reporting

False discovery rate (FDR; Benjamini–Hochberg) is controlled within analysis families (model-vs-baseline, within-model item-group, within-model form, first-epoch). For planned Baseline-vs-treatment contrasts, Holm-adjusted results may additionally be reported. Reporting conventions: probabilities with 95% CIs on the response scale; contrasts as OR (SE, z, adjusted p); acquisition times with bootstrap 95% CIs; exact p-values to three decimals, or "$p<.001$" when smaller.

Diagnostics and Estimation

Models are fit with `bobyqa` (increased iteration limits) and, for spline scans, `nloptwrap` when required. Convergence and boundary fits are monitored; when boundary fits occur, confirmatory fits with simplified random effects are checked. Non-convergence due to separation triggers the exact-test fallback noted above.

Software

Analyses are conducted in R (v4.4) using `lme4` and `lmerTest` (GLMMs and tests), `emmeans` (EMMs and contrasts), `splines` (natural splines), `MASS` (parametric simulation), and the `tidyverse` for data handling.

Experiment 0

In the first experiment the baseline model is trained on the unablated 90M-word training set derived from the BabyLM [80] strict dataset.

Model Evaluation

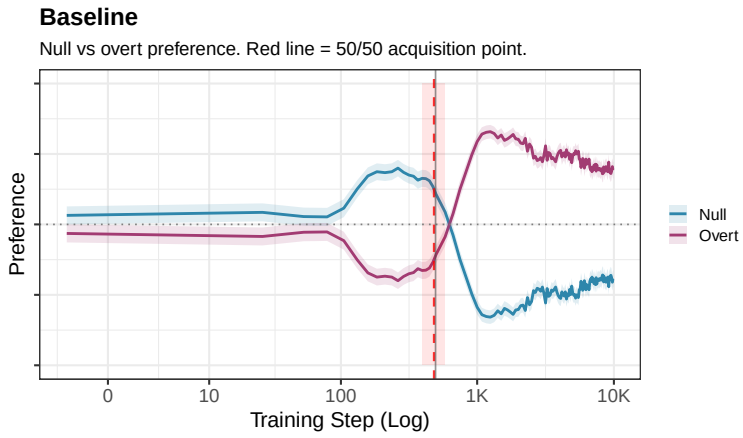

Figure 1: Model preference for null and overt evaluation stimuli over training, training steps transformed to log-scale to reflect model log-learning dynamics for Experiment 0 - Baseline

AIC-based model selection indicated that the baseline model achieved optimal fit with 6 degrees of freedom (AIC = 146242). The model achieved $t50$ at checkpoint 482 (95% CI [396, 576]). Age of Acquisition analysis revealed that baseline achieved AoA at checkpoint 727 (95% CI [664, 791]).

First epoch analysis shows that the Baseline model exhibited a significant preference for null subjects by the end of the first epoch, showing a 63.4% preference for null subjects (95% CI [62.7, 64.1], $p < .001$).

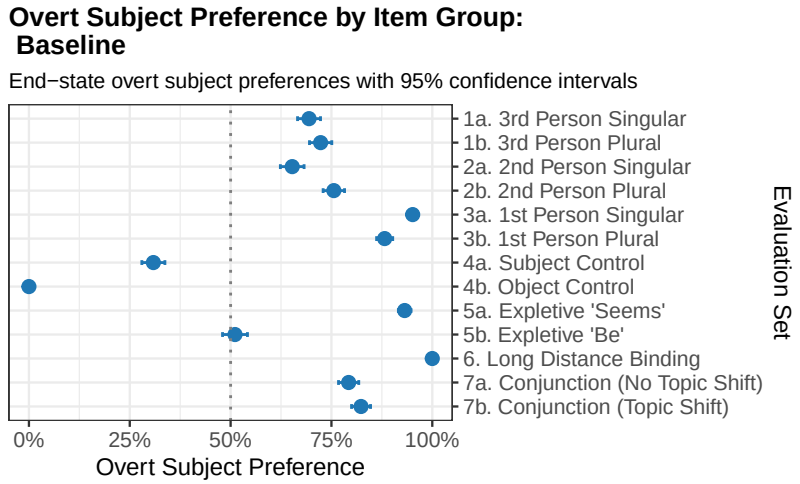

Figure 2: Model preference for overt subjects by evaluation group at final checkpoint for Experiment 0 - Baseline

The end-state analysis showed that the base model strongly preferred overt subjects, with a 69.6% preference over null subjects in the last two epochs of training (95% CI [66.5%, 72.5%], p <.001).

Mixed-effects pairwise comparisons revealed significant person-based differences. First person contexts (93.3%) elicited significantly more overt subjects than both second person (73.1%) (OR = 5.081, 95% CI [4.009, 6.438], $p < .001$) and third person contexts (73.4%) (OR = 5.008, 95% CI [3.952, 6.347], $p < .001$). There was no significant difference between second and third person contexts (OR = 0.986, 95% CI [0.826, 1.177], $p > .05$).

Within-person number contrasts showed opposite patterns across persons. For second person, plural contexts (78.3%) elicited significantly more overt subjects than singular contexts (67.9%) (OR = 0.585, 95% CI [0.473, 0.723], $p < .001$). Conversely, for first person, singular contexts (96.1%) showed significantly higher preference than plural contexts (90.3%) (OR = 2.667, 95% CI [1.857, 3.832], $p < .001$).

Mixed-effect models were unable to converge comparing control contrasts because of Perfect Separation. In these cases the data are instead analyzed with Fisher's exact test. The model showed a complete subject-object control asymmetry, with subject control contexts (30.4%) showing dramatically higher overt preferences than object control contexts (1.6%) ($p < .001$).

Expletive constructions showed differential behavior by verb type. *Seems*-constructions strongly favored overt subjects (91.4%), while *be*-constructions showed no preference over chance (50.3%, $p > .05$). No difference was found between conjoined phrases with (88.3%) and without (89.8%) topic shift.

Table 6: Pairwise comparisons of within Item Group differences

Person	
1st vs 2nd	↑***
1st vs 3rd	↑***
2nd vs 3rd	–
Number	
1st: sg vs pl	↑***
2nd: sg vs pl	↓***
3rd: sg vs pl	–
Control	
Subj vs Obj	↑***
Expletive	
Seems vs Be	↑***
Topic	
No vs Shift	–

Table 7: Overt subject preference by syntactic context at final checkpoint for the baseline model

Item Group	Accuracy	95% CI	vs Chance	p-value
1st Singular	95.1%	[93.5, 96.4]	Above	< .001
1st Plural	88.2%	[86.0, 90.2]	Above	< .001
2nd Singular	65.7%	[62.6, 68.7]	Above	< .001
2nd Plural	75.6%	[72.8, 78.4]	Above	< .001
3rd Singular	69.6%	[66.5, 72.5]	Above	< .001
3rd Plural	72.3%	[69.3, 75.2]	Above	< .001
Subject Control	30.2%	[27.3, 33.3]	Below	< .001
Object Control	0.0%	[0.0, 0.4]	Below	< .001
Seems Expletive	93.2%	[91.4, 94.7]	Above	< .001
Be Expletive	51.0%	[47.7, 54.2]	At	0.578
No Topic Shift	79.5%	[76.8, 82.0]	Above	< .001
Topic Shift	81.4%	[78.8, 83.9]	Above	< .001

All processing manipulations showed preferences for overt subjects (see Table 13). The model preferred overt subjects

Table 8: Pairwise comparisons of within Processing differences

Complex	
Emb vs Long	–
Negation	
Targ vs Cont	↑***
Targ vs Both	–

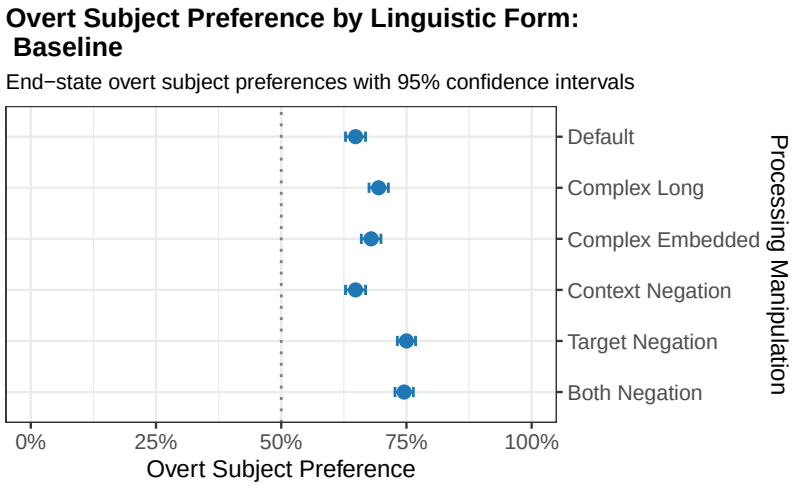

Figure 3: Model preferences for overt subjects by processing manipulation at final checkpoint for Experiment 0 - Baseline

64.7% of the time in default contexts. Complex constructions significantly increased overt preferences: long noun phrases (69.4%) and embedded relatives (68.1%) both exceeded default rates, though the difference between complexity types was not significant (OR = 1.064, 95% CI [0.874, 1.294], $p > .05$).

Context negation showed no difference from default (65.3%, $p > .05$). However, target negation (75.7%) and both-context negation (74.8%) significantly increased overt preferences compared to default ($p < .001$). Target negation significantly exceeded context negation (OR = 0.603, 95% CI [0.493, 0.736], $p < .001$), while target and both-negation conditions did not differ significantly (OR = 1.030, 95% CI [0.836, 1.269], $p = > .05$).

Form	Accuracy	95% CI	vs Chance	p-value
Default	64.7%	[62.6, 66.8]	Above	< .001
Complex Long	69.4%	[67.3, 71.4]	Above	< .001
Complex Emb	68.1%	[66.0, 70.1]	Above	< .001
Context Negation	64.7%	[62.6, 66.8]	Above	< .001
Target Negation	75.0%	[73.0, 76.8]	Above	< .001
Both Negation	74.4%	[72.4, 76.3]	Above	< .001

Table 9: Overt subject preference by processing manipulation at final checkpoint for the baseline model

Experiment 1

The ablation targeted and excised all expletive subjects. After the first round of removal, replacement sentences are placed into the corpus and the process is repeated until satisfied.

Pre-ablation size across sources was N=89,014,604 tokens. A total of 183,431 expletive instances were removed, corresponding to 0.206% of all tokens in the training set. Because the replacement

step occasionally adds content, the net token change was an increase of +27,829 tokens overall.

By source (expletives removed; share of that source; net token change):

- **BNC Spoken**: 24,331 (0.349%); net +957 tokens.
- **CHILDES**: 50,283 (0.194%); net −796 tokens.
- **Gutenberg**: 52,412 (0.221%); net +29,342 tokens.
- **OpenSubtitles**: 37,267 (0.208%); net −1,322 tokens.
- **Simple Wikipedia**: 15,948 (0.121%); net −459 tokens.
- **Switchboard**: 3,190 (0.264%); net +107 tokens.

Taken together, the ablation removes a small, well-defined portion of the corpus (about two expletives per thousand tokens) while leaving overall corpus size essentially unchanged due to the controlled insertions during replacement.

Model Evaluation

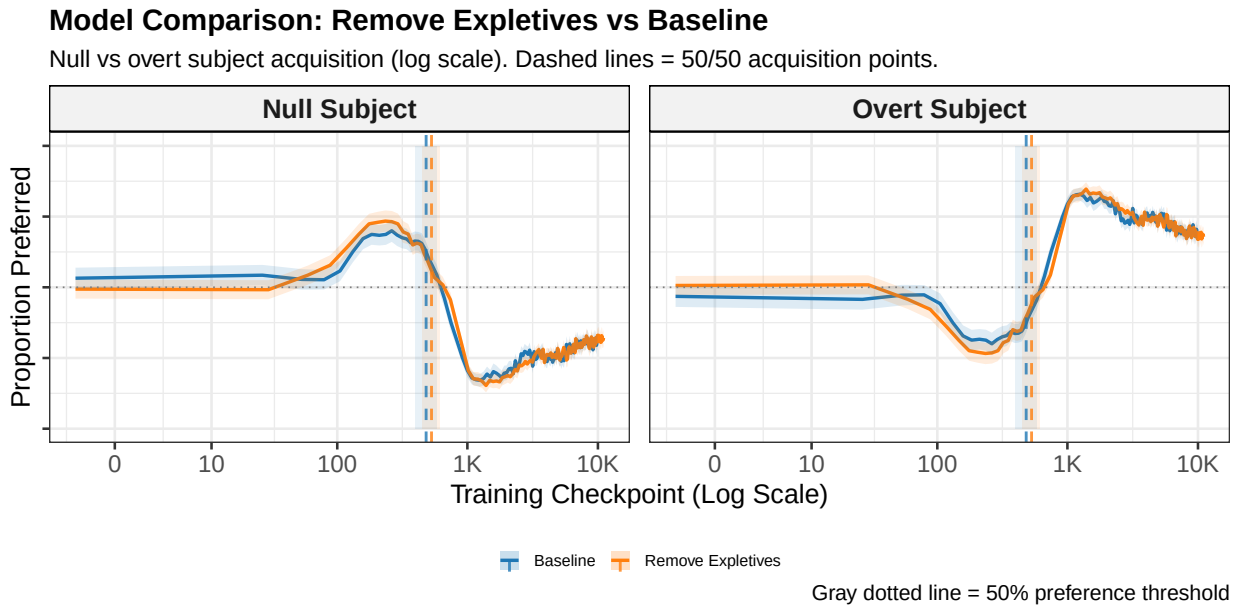

Figure 4: Model preference for null and overt evaluation stimuli over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

AIC-based model selection indicated that the baseline model achieved optimal fit with 6 degrees of freedom (AIC = 146501). The model achieved t50 at checkpoint 531 (95% CI [451, 616]). Age of Acquisition analysis revealed that baseline achieved AoA at checkpoint 767 (95% CI [709, 820]). The ablated model reached acquisition criterion significantly later than the baseline model (ΔAoA = 39.71 epochs, 95% CI [24, 55], $p < .001$).

First epoch analysis shows that the Baseline model exhibited a significant preference for null subjects by the end of the first epoch, showing a 64.1% preference for null subjects (95% CI [63.4, 64.8], $p < .001$). This is not significantly different from the baseline model's first-epoch performance ($p > .05$) when correcting for multiple comparisons.

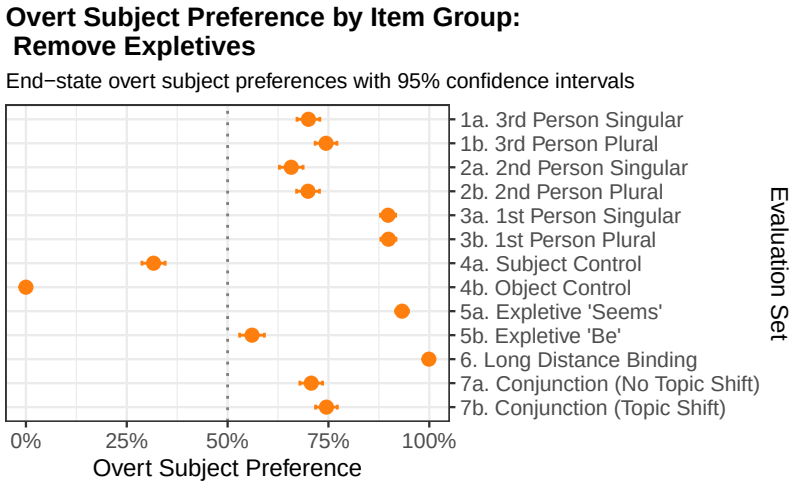

Figure 5: Model preference for overt subjects by evaluation group at final checkpoint

The end-state analysis showed that the base model strongly preferred overt subjects, with a 68.1% preference over null subjects in the last two epochs of training (95% CI [67.2%, 69.1%], $p < .001$). This is not significantly different from the baseline model's performance (69.3%, $p > .05$) when correction for multiple comparisons is applied.

Mixed-effects pairwise comparisons revealed significant person-based differences. First person contexts (91.4%) elicited significantly more overt subjects than both second person (70.4%, OR = 4.494, 95% CI [3.539, 5.707], $p < .001$) and third person contexts (74.2%, OR = 3.704, 95% CI [2.911, 4.713], $p < .001$). There was no significant difference found between 2nd and 3rd person contexts (OR = 0.824, 95% CI [0.682, 0.996], $p = .054$) when correcting for multiple comparisons.

First person singular (91.6%) and plural (91.3%) contexts were not found to be significantly different, (OR = 1.029, 95% CI [0.739, 1.432]), $p > .05$. Second person singular contexts (67.8%) elicited significantly less overt subjects than plural contexts (73%, OR = 1.840, 95% CI [1.469, 2.305], $p < .05$). Conversely, for third person, singular contexts (72.6%) and plural contexts (75.9%) showed no difference (OR = 0.838, 95% CI [0.667, 1.053], $p > .05$) for preference of overt subjects.

Mixed-effect models were unable to converge comparing con-

Table 10: Pairwise comparisons of within Item Group differences

Person	
1st vs 2nd	↑***
1st vs 3rd	↑***
2nd vs 3rd	–
Number	
1st: sg vs pl	–
2nd: sg vs pl	↓***
3rd: sg vs pl	–
Control	
Subj vs Obj	↑***
Expletive	
Seems vs Be	↑***
Topic	
No vs Shift	↓***

trol contrasts because of Perfect Separation. In these cases the data are instead analyzed with Fisher's Exact test. A significant subject-object asymmetry was observed with subject control contexts (31.4%) showing dramatically higher overt preferences than object control contexts (0%) ($p < .001$).

Expletive constructions showed differential behavior by verb type. *Seems*-constructions strongly favored overt subjects (97.1%), than *be*-constructions (56.9%, OR = 0.040, 95% CI [0.026, 0.060], $p > .001$). Counter to the baseline model, *be*-like constructions do in-fact differ from chance in overt preference, see Table 11.

The model shows a significantly higher preference for overt subjects in topic shift (84.3%) than non-topic-shift contexts (79.6%, OR = .726, 95% CI [0.557, 0.946], p < .001)

Item Group	Accuracy	95% CI	vs Chance	p-value
1st Singular	89.9%	[87.6, 91.9]	Above	< .001
1st Plural	89.6%	[87.3, 91.7]	Above	< .001
2nd Singular	65.7%	[62.2, 69.0]	Above	< .001
2nd Plural	70.6%	[67.3, 73.7]	Above	< .001
3rd Singular	70.2%	[66.9, 73.4]	Above	< .001
3rd Plural	73.5%	[70.3, 76.5]	Above	< .001
Subject Control	31.4%	[28.2, 34.8]	Below	< .001
Object Control	0.0%	[0.0, 0.5]	Below	< .001
Seems Expletive	93.6%	[91.6, 95.2]	Above	< .001
Be Expletive	56.1%	[52.5, 59.6]	Above	< .001
No Topic Shift	70.7%	[67.4, 73.9]	Above	< .001
Topic Shift	75.1%	[72.0, 78.1]	Above	< .001

Table 11: Overt subject preference by syntactic context at final checkpoint for the remove expletives model

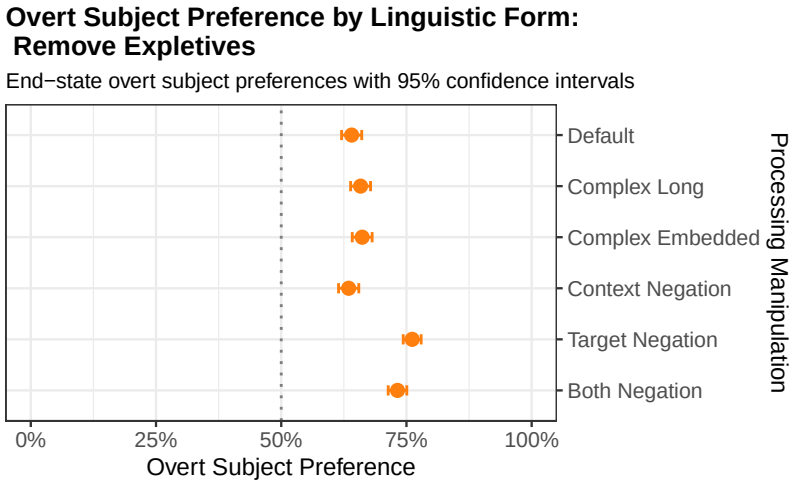

Figure 6: Model preferences for overt subjects by processing manipulation at final checkpoint.

All processing manipulations showed preferences for overt subjects (see Table 13). The model preferred overt subjects 71.2% of the time in default contexts. Complex constructions were significantly more likely to prefer overt complementizers:

comparing long noun phrases (76.4%, OR = 1.311, 95% CI [1.076, 1.596], p = .006) and embedded relatives (75.5%, OR = 1.245, 95% CI [1.024, 1.515], p < .001) and the difference between the two constructions is not significant (OR = 1.053, 95% CI [0.838, 1.322], $p > .05$).

Context negation showed no difference from default (71.2%, OR = 1.072, 95% CI [0.885, 1.300], $p > .05$). Further, overt preference was not significantly higher in target negation (74.5%, OR = 1.184, 95% CI [0.975, 1.439], p > .05). However, overt complementizers were preferred significantly more in stimuli with negation on both target and context sentences (81.4%, OR = 1.773, 95% CI [1.442, 2.179], p < .001) contexts. Target negation significantly exceeded context negation (OR = 0.516, 95% CI [0.414, 0.642], $p < .001$). Contexts with negation on the target alone were significantly lower in overt subject preference than when negation was present on both (OR = 0.668, 95% CI [0.527, 0.847], $p < .001$).

Table 12: Pairwise comparisons of within Processing differences

Complex	
Emb vs Long	–
Negation	
Targ vs Cont	↑***
Targ vs Both	–

Form	Accuracy	95% CI	vs Chance	p-value
Default	64.3%	[62.0, 66.5]	Above	< .001
Complex Long	66.1%	[63.8, 68.4]	Above	< .001
Complex Emb	66.1%	[63.8, 68.3]	Above	< .001
Context Negation	63.1%	[60.7, 65.3]	Above	< .001
Target Negation	76.5%	[74.4, 78.4]	Above	< .001
Both Negation	73.0%	[70.9, 75.1]	Above	< .001

Table 13: Overt subject preference by processing manipulation at final checkpoint for the remove expletives model

Experiment 2

The dataset ablated was the previously ablated dataset from Experiment 1. The ablation systematically impoverishes determiner morphology *in place* (token-preserving substitutions). Unlike removal procedures, this transformation does not alter corpus length.

Pre-ablation size across sources was N=89,042,433 tokens. A total of 6,938,234 determiners were impoverished, corresponding to 7.792% of all tokens in the training set. Because the operation is an in-place substitution, the net token change was 0 overall.

By source (determiners impoverished; share of that source; net token change):

- **BNC Spoken**: 530,987 (7.610%); net 0 tokens.
- **CHILDES**: 1,468,619 (5.652%); net 0 tokens.
- **Gutenberg**: 2,353,331 (9.914%); net 0 tokens.
- **OpenSubtitles**: 1,235,717 (6.886%); net 0 tokens.

- **Simple Wikipedia**: 1,278,332 (9.690%); net 0 tokens.
- **Switchboard**: 71,248 (5.905%); net 0 tokens.

Overall, determiners constitute a substantial, well-delimited portion of the corpus (about one in thirteen tokens), and the impoverishment procedure leaves corpus size unchanged while uniformly modifying the targeted category.

Model Evaluation

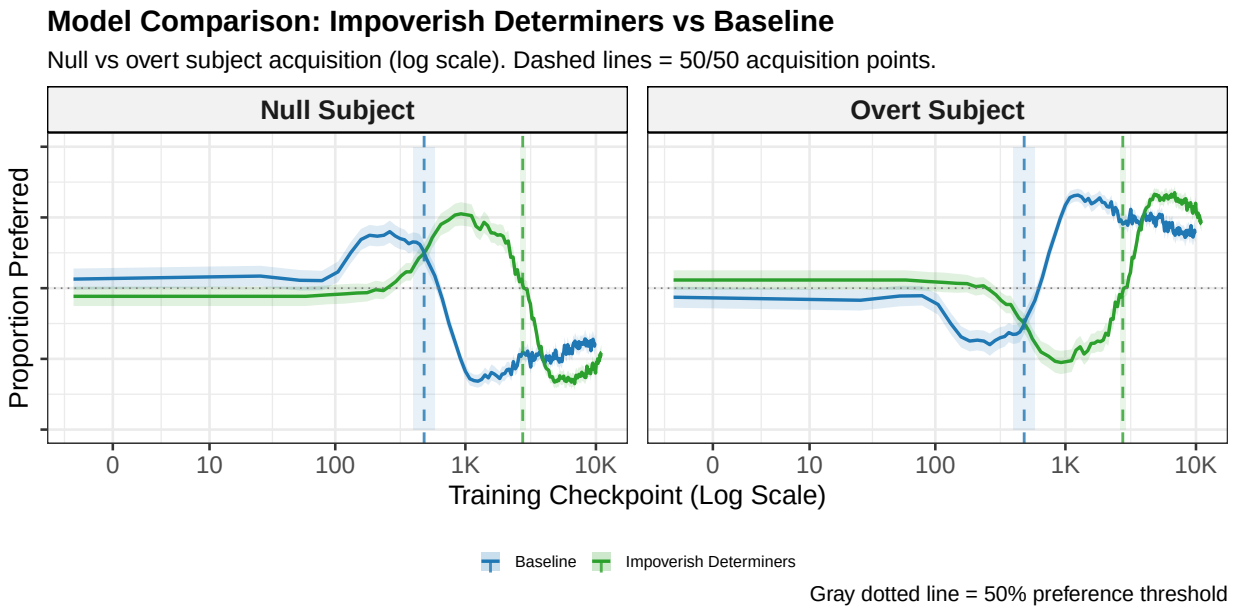

Figure 7: Model preference for null and overt evaluation stimuli over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

AIC-based model selection indicated that the baseline model achieved optimal fit with 7 degrees of freedom (AIC = 139485). The model achieved t50 at checkpoint 2751˜˜ (95% CI [451, 616]). Age of Acquisition analysis revealed that baseline achieved AoA at checkpoint 3400 (95% CI [3306, 3498]). The ablated model reached acquisition criterion significantly later than the baseline model (ΔAoA = 2672 epochs, 95% CI [2620, 2724], $p < .001$).

First epoch analysis shows that the Baseline model exhibited a significant preference for null subjects by the end of the first epoch, showing a 53.2% preference for null subjects (95% CI [52.5,53.9], $p < .001$). This is significantly different from the baseline model's first-epoch performance (63.5%, OR 1.53, 95% CI [1.47, 1.59) $p < .001$).

The end-state analysis showed that the base model strongly preferred overt subjects, with a 74.9% preference over null sub-

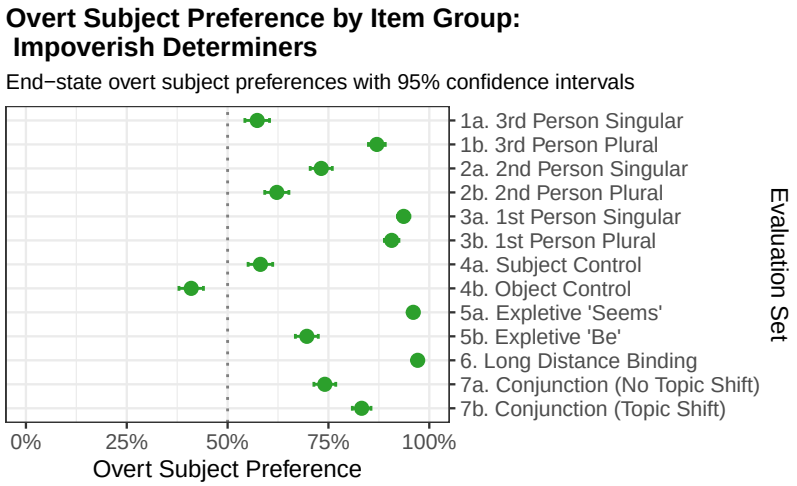

Figure 8: Model preference for overt subjects by evaluation group at final checkpoint

jects in the last two epochs of training (95% CI [74%, 75.7%], p <.001). This is significantly higher than the baseline model's preference (69.3%, (OR = .756, 95% CI [.711, 1.121, p < .001) .

Mixed-effects pairwise comparisons revealed significant person-based differences. First person contexts (93.7%) elicited significantly more overt subjects than both second person (69.7%, OR = 6.462, 95% CI [4.971, 8.400], p < .001) and third person contexts (74.3%, OR = 5.129, 95% CI [3.938, 6.680], p < .001). Further there was a significant difference found between 2nd and 3rd person contexts (OR = 0.794, 95% CI [0.656, 0.960], p < .05).

First person singular (95.2%) and plural (92.5%) contexts were shown to be significantly different, (OR = 1.594, 95% CI [1.477, 2.331], p < .05). Further, second person singular contexts (76.1%) elicited significantly more overt subjects than plural contexts (63.4%) (OR = 1.840, 95% CI [1.469, 2.305], p < .001). Conversely, for third person, singular contexts (58.5%) showed significantly lower preference than plural contexts (89%) (OR = 0.173, 95% CI [0.134, 0.225], p < .001) for overt subjects.

The model showed a strong subject-object control asymmetry, with subject control contexts (57.6%) showing dramatically higher overt preferences than object control contexts (36.5%) (p < .001). It goes without statistical testing that use of overt pronouns in object control conditions increased compared to the baseline.

Expletive constructions showed differential behavior by verb type. *Seems*-constructions strongly favored overt subjects (98.3%), than *be*-constructions (72%, p > .05). Counter to the baseline model, *be*-like constructions do in-fact differ from chance in overt preference, see Table 17.

Table 14: Pairwise comparisons of within Item Group differences

Person	
1st vs 2nd	↑***
1st vs 3rd	↑***
2nd vs 3rd	↑***
Number	
1st: sg vs pl	↑***
2nd: sg vs pl	↑***
3rd: sg vs pl	↓***
Control	
Subj vs Obj	↑***
Expletive	
Seems vs Be	↑***
Topic	
No vs Shift	↓***

The model shows a significantly stronger preference to drop pronouns in non-topic-shift contexts (74.9%) than topic shift contexts (84.3%, OR = .556, 95% CI [0.436, 0.710], p < .001)

Item Group	Accuracy	95% CI	vs Chance	p-value
1st Singular	93.8%	[91.9, 95.4]	Above	< .001
1st Plural	90.7%	[88.4, 92.6]	Above	< .001
2nd Singular	73.5%	[70.3, 76.5]	Above	< .001
2nd Plural	61.6%	[58.1, 65.0]	Above	< .001
3rd Singular	57.2%	[53.7, 60.7]	Above	< .001
3rd Plural	86.7%	[84.2, 89.0]	Above	< .001
Subject Control	57.6%	[54.0, 61.0]	Above	< .001
Object Control	36.5%	[33.1, 40.0]	Below	< .001
Seems Expletive	96.1%	[94.5, 97.3]	Above	< .001
Be Expletive	69.6%	[66.2, 72.8]	Above	< .001
No Topic Shift	71.6%	[68.3, 74.7]	Above	< .001
Topic Shift	81.3%	[78.4, 84.0]	Above	< .001

Table 15: Overt subject preference by syntactic context at final checkpoint for the impoverish determiners model

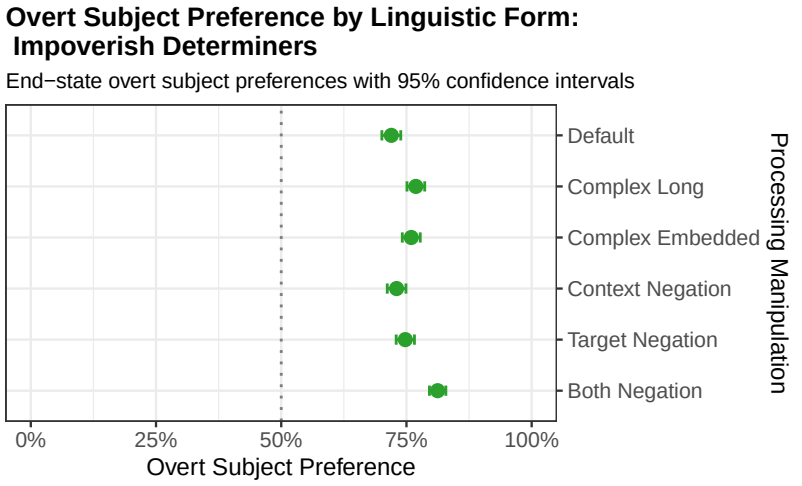

Figure 9: Pairwise comparisons within item groups for null subject preference in the impoverish determiners model

All processing manipulations showed preferences for overt subjects (see Table 13). The model preferred overt subjects 64.3% of the time in default contexts. Complex constructions significantly increased overt preferences: long noun phrases (69.4%) and embedded relatives (68.1%) both exceeded default rates, though the difference between complexity types was not significant (OR = 0.938, 95% CI [0.818, 1.076], p = .362).

Context negation showed no difference from default (65.3%, p > .05). However, target negation (75.7%) and both-context negation (74.8%) significantly increased overt preferences compared to default (p < .001). Target negation significantly exceeded context negation (OR = 0.605, 95% CI [0.527, 0.694], p < .001), while target and both-negation conditions did not differ significantly (OR = 0.971, 95% CI [0.842, 1.120], p = .689).

Table 16: Pairwise comparisons of within Processing differences

Complex	
Emb vs Long	–
Negation	
Targ vs Cont	–
Targ vs Both	↑***

Form	Accuracy	95% CI	vs Chance	p-value
Default	70.8%	[68.6, 72.9]	Above	< .001
Complex Long	76.0%	[73.9, 78.0]	Above	< .001
Complex Emb	75.1%	[72.9, 77.1]	Above	< .001
Context Negation	72.2%	[70.0, 74.3]	Above	< .001
Target Negation	74.1%	[72.0, 76.2]	Above	< .001
Both Negation	81.0%	[79.1, 82.8]	Above	< .001

Table 17: Overt subject preference by processing manipulation at final checkpoint for the impoverish determiners model

Experiment 3

The dataset ablated was the previously ablated dataset from Experiment 1. The ablation excises all articles. After the first pass, replacement sentences are reinserted to preserve the corpus size, and the procedure is iterated; negative values in the token–change summaries therefore indicate a net addition of tokens.

Pre-ablation size across sources was N=89,042,433 tokens. In total, 6,255,688 article tokens were removed, corresponding to 7.026% of all tokens in the training set. Because the replacement step can add or remove material, the net token change was a *decrease* of 323,169 tokens overall.

By source (articles removed; share of that source; net token change):

- **BNC Spoken**: 449,971 (6.449%); net −41,025 tokens.
- **CHILDES**: 1,195,160 (4.600%); net −254,272 tokens.
- **Gutenberg**: 2,198,527 (9.262%); net +34,627 tokens.
- **OpenSubtitles**: 1,017,403 (5.669%); net −150,074 tokens.
- **Simple Wikipedia**: 1,334,027 (10.113%); net +84,118 tokens.
- **Switchboard**: 60,600 (5.022%); net +3,457 tokens.

Overall, the ablation removes a substantial, well-delimited category (roughly one article per fourteen tokens) while keeping corpus size broadly comparable despite localized insertions during replacement.

Model Evaluation

AIC-based model selection indicated that the baseline model achieved optimal fit with 7 degrees of freedom (AIC = 143567). The model achieved t50 at checkpoint 595 (95% CI [532, 660]). Age of Acquisition analysis revealed that baseline achieved AoA at checkpoint 807 (95% CI [758, 861]). The ablated model reached acquisition criterion significantly later than the baseline model (ΔAoA = 80 epochs, 95% CI [81, 108], $p < .001$).

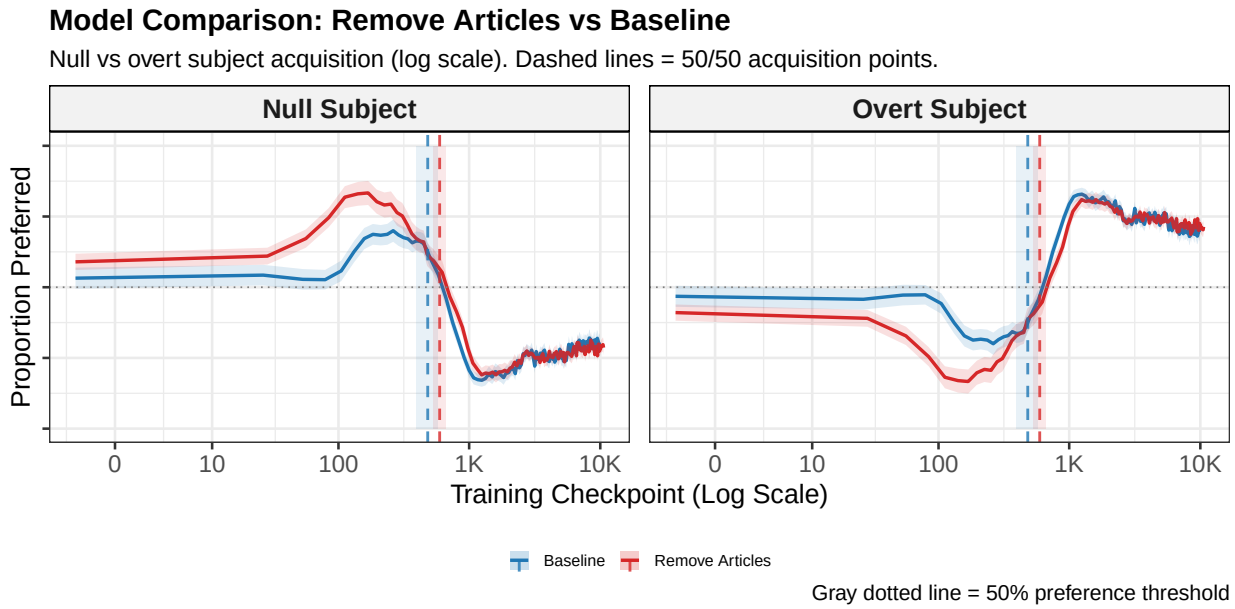

Figure 10: Model preference for null and overt evaluation stimuli over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

First epoch analysis shows that the Baseline model exhibited a significant preference for null subjects by the end of the first epoch, showing a 71.7% preference for null subjects (95% CI [71.1,72.4], $p < .001$). This is significantly different from the baseline model's first-epoch performance (63.5%, OR = .68, 95% CI [.65, .71], p < .001).

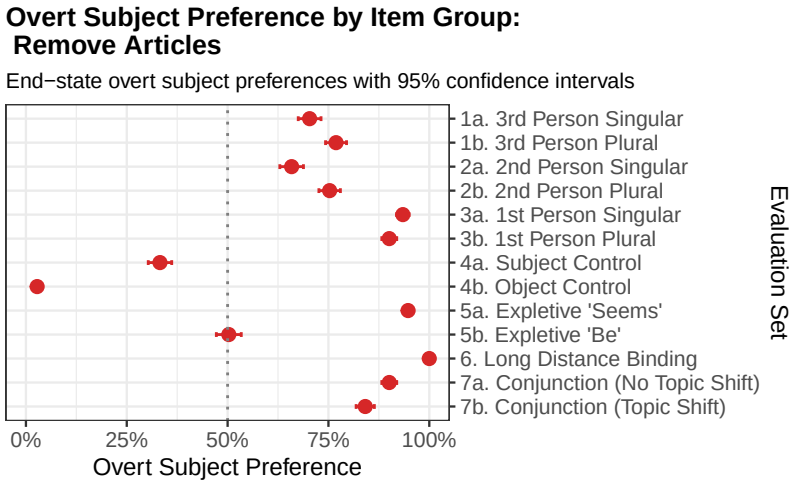

Figure 11: Model preference for overt subjects by evaluation group at final checkpoint

The end-state analysis showed that the base model strongly preferred overt subjects, with a 68.2% preference over null subjects in the last two epochs of training (95% CI [74%, 75.7%], $p < .001$). This is significantly lower than the baseline model's

preference (69.3%, (OR = .92, 95% CI [.87, .975, p < .001) after correction for multiple comparisons.

Mixed-effects pairwise comparisons revealed significant person-based differences. First person contexts (92.8%) elicited significantly more overt subjects than both second person (72.1%, OR = 4.987, 95% CI [3.91, 6.36], $p < .001$) and third person contexts (75%, OR = 4.302, 95% CI [3.367, 5.497], $p < .001$). There was no significant difference found between 2nd and 3rd person contexts (OR = .863, 95% CI [0.718, 1.037], $p > .05$).

First person singular (94.6%) and plural (90.9%) contexts were shown to be significantly different (OR = 1.761, 95% CI [1.237, 2.508], $p = .003$). Further, second person singular contexts (67.5%) elicited significantly less overt subjects than plural contexts (76.6%, OR = .633, 95% CI [.510, .785], $p < .001$). Finally, third person singular contexts (72.1%) showed significantly lower preference for overt subjects than plural contexts (77.8%, OR = .737, 95% CI [.591, .920], $p < .01$).

Mixed-effect models were unable to converge comparing control contrasts because of Perfect Separation. In these cases the data are instead analyzed with Fisher's Exact test. A significant subject-object asymmetry was observed with subject control contexts (31.9%) showing dramatically higher overt preferences than object control contexts (2.7%, $p < .001$).

Expletive constructions showed differential behavior by verb type. *Seems*-constructions strongly favored overt subjects (98.3%), than *be*-constructions (50.6%, OR = .031, 95% CI [.021, .045], $p > .05$). Like the baseline model, *be*-like constructions do not differ from chance in overt preference, see Table **??**.

The model shows a significant difference between topic-shift contexts, with a stronger preference for non-topic shift (96.6%) than topic shift contexts (93.7%, OR = 1.95, 95% CI [1.433, 2.655], p < .001).

All processing manipulations showed preferences for overt subjects (see Table 21). The model preferred overt subjects 67.6% of the time in default contexts. There was no significant difference comparing the default with forms including long noun phrases (68.5%, OR 1.043, 95% CI [.955,1.368], p > .05) and embedded relatives (70.4%, OR 1.143, 95%). Further, the difference between complexity types was not significant (OR = .912, 95% CI [.745, 1.118], $p > .05$).

Context negation showed no difference from default (66%, OR .988, CI [827, 1.179,$p > .05$). However, there was a significantly increased overt preferences compared to the default for target negation contexts (78.4%, OR = 1.738, 95% CI [1.439, 2.100],

Table 18: Pairwise comparisons of within Item Group differences

Person	
1st vs 2nd	↑***
1st vs 3rd	↑***
2nd vs 3rd	–
Number	
1st: sg vs pl	↑***
2nd: sg vs pl	↓***
3rd: sg vs pl	↓***
Control	
Subj vs Obj	↑***
Expletive	
Seems vs Be	↑***
Topic	
No vs Shift	↑***

Table 20: Pairwise comparisons of within Processing differences

Complex	
Emb vs Long	–
Negation	
Targ vs Cont	↑***
Targ vs Both	–

Item Group	Accuracy	95% CI	vs Chance	p-value
1st Singular	93.6%	[91.8, 95.2]	Above	< .001
1st Plural	89.5%	[87.2, 91.4]	Above	< .001
2nd Singular	65.9%	[62.6, 69.0]	Above	< .001
2nd Plural	74.7%	[71.6, 77.5]	Above	< .001
3rd Singular	70.3%	[67.1, 73.3]	Above	< .001
3rd Plural	75.8%	[72.8, 78.6]	Above	< .001
Subject Control	31.9%	[28.8, 35.2]	Below	< .001
Object Control	2.7%	[1.7, 4.0]	Below	< .001
Seems Expletive	94.9%	[93.2, 96.3]	Above	< .001
Be Expletive	50.6%	[47.2, 54.0]	At	0.759
No Topic Shift	90.0%	[87.9, 92.0]	Above	< .001
Topic Shift	83.7%	[81.0, 86.1]	Above	< .001

Table 19: Overt subject preference by syntactic context at final checkpoint for the remove articles model

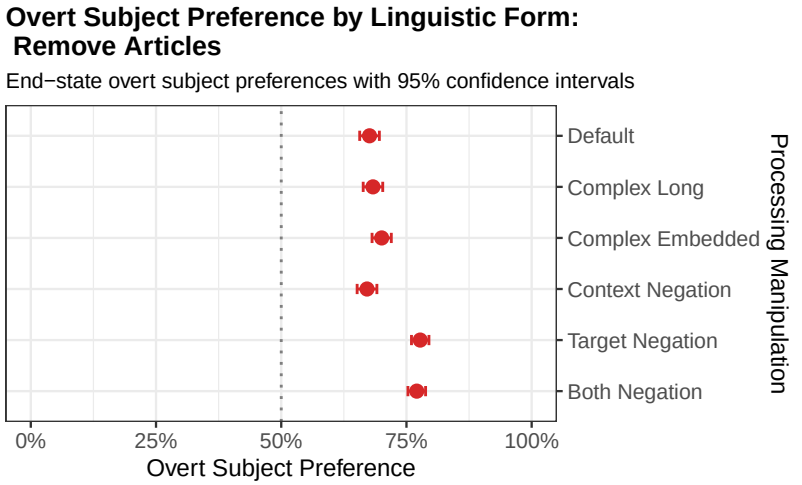

Figure 12: Model preferences for overt subjects by processing manipulation at final checkpoint.

$p < .001$) and constructions with negation in both the target and context (77.4%, OR = 1.644, 95% CI [1.363, 1.983). Target negation significantly exceeded context negation (OR = .568, 95% CI [0.456, 0.703], $p < .001$), while target and both-negation conditions did not differ significantly (OR = 1.057, 95% CI [0.845, 1.323], $p > .05$).

Form	Accuracy	95% CI	vs Chance	p-value
Default	67.0%	[64.9, 69.2]	Above	< .001
Complex Long	67.9%	[65.8, 70.1]	Above	< .001
Complex Emb	69.9%	[67.7, 71.9]	Above	< .001
Context Negation	66.8%	[64.6, 68.9]	Above	< .001
Target Negation	77.8%	[75.8, 79.6]	Above	< .001
Both Negation	76.8%	[74.8, 78.7]	Above	< .001

Table 21: Overt subject preference by processing manipulation at final checkpoint for the remove articles model

Experiment 4

The dataset ablated was the previously ablated corpus from Experiment 1. The ablation lemmatizes all verbal tokens to the infinitival form *in place* to remove inflectional variability while preserving tokenization; as a token-preserving substitution, it does not change corpus length.

Pre-ablation size across sources was N=89,042,433 tokens. In total, 11,924,417 verb tokens were lemmatized, corresponding to 13.392% of all tokens in the training set. Because the operation is in-place, the net token change was 0 overall.

By source (verbs lemmatized; share of that source; net token change):

- **BNC Spoken**: 964,088 (13.816%); net 0 tokens.
- **CHILDES**: 3,537,141 (13.613%); net 0 tokens.
- **Gutenberg**: 3,358,317 (14.148%); net 0 tokens.
- **OpenSubtitles**: 2,763,574 (15.400%); net 0 tokens.
- **Simple Wikipedia**: 1,158,213 (8.780%); net 0 tokens.
- **Switchboard**: 143,084 (11.859%); net 0 tokens.

Overall, the manipulation uniformly targets a large and well-delimited category (about one verb per seven to eight tokens) while leaving total corpus size unchanged.

Model Evaluation

AIC-based model selection indicated that the baseline model achieved optimal fit with 6 degrees of freedom (AIC = 155563). The model achieved t50 at checkpoint 551 (95% CI [484, 623]). Age of Acquisition analysis revealed that baseline achieved AoA at checkpoint 705 (95% CI [660, 748]). The ablated model reached acquisition criterion significantly earlier than the baseline model (ΔAoA = -22 epochs, 95% CI [-43, -1.65], p = .034).

First epoch analysis shows that the Baseline model exhibited a significant preference for null subjects by the end of the first epoch, showing a 65.2% preference for null subjects (95% CI 64.5,65.9], $p < .001$). This is significantly different from the baseline model's first-epoch performance (63.5%, OR = .926 95% CI [886, 967], p = .003).

The end-state analysis showed that the base model strongly preferred overt subjects, with a 61.4% preference for overt subjects in the last two epochs of training (95% CI [60.6%, 62.3%], $p <.001$). This is significantly lower than the baseline model's preference (69.3%, (OR = 1.432, 95% CI [1.36, .975], p < 1.51).

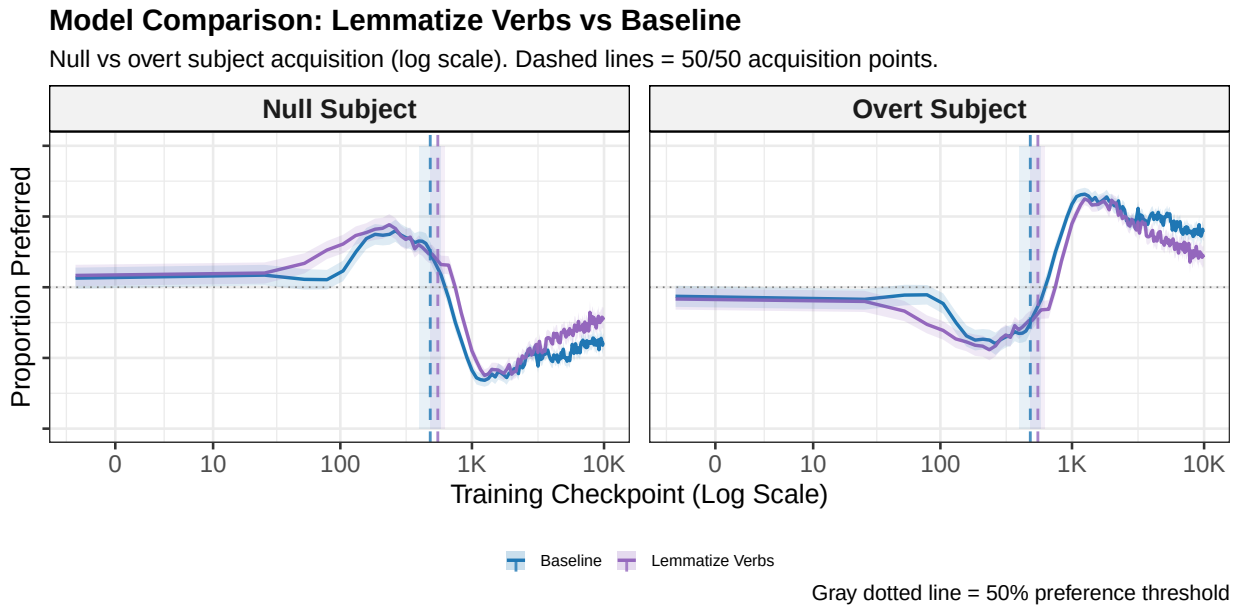

Figure 13: Model preference for null and overt evaluation stimuli over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

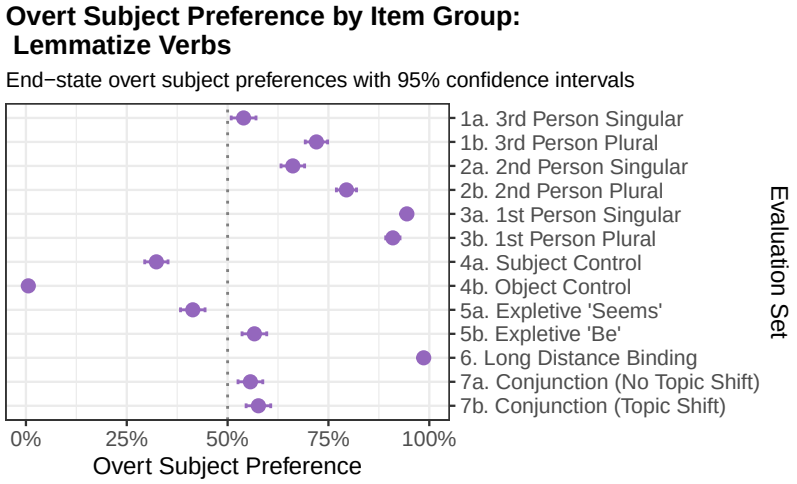

Figure 14: Model preference for overt subjects by evaluation group at final checkpoint

Mixed-effects pairwise comparisons revealed significant person-based differences. First person contexts (94.1%) elicited significantly more overt subjects than both second person (75.1%, OR = 5.315, 95% CI [4.147, 6.810], $p < .001$) and third person contexts (65%, OR = 4.6734, 95% CI [6.734, 10.989], $p < .001$). A significant difference was found between 2nd and 3rd person contexts (OR = 1.619, 95% CI [1.360, 1.926], $p > .05$). ˇ First person singular (95.7%) and plural (92.7%) contexts were shown to be significantly different (OR = 1.738, 95% CI [1.206, 2.504], $p = .004$). Further, second person singular contexts (68.2%) elicited

significantly less overt subjects than plural contexts (81.9%, OR = .473, 95% CI [.381, .588], $p < .001$). Finally, third person singular contexts (55.5%) showed significantly lower preference for overt subjects than plural contexts (74.4%, OR = .430, 95% CI [.351, .527], $p < .001$).

Mixed-effect models were unable to converge comparing control contrasts because of Perfect Separation. In these cases the data are instead analyzed with Fisher's Exact test. A significant subject-object asymmetry was observed with subject control contexts (31.6%) showing dramatically higher overt preferences than object control contexts (0.3%, $p < .001$).

Expletive constructions showed differential behavior by verb type. *Seems*-constructions strongly favored overt subjects (40.3%) more than *be*-constructions (40.3%, OR = 1.969, 95% CI [1.622, 2.390], $p < .001$). Unlike the baseline model, *be*-like constructions differ from chance in overt preference, see Table **??**.

The model shows no significant difference between topic-shift contexts, with no difference in the preference for non-topic shift (56.6%) than topic shift contexts (58.5%, OR = .924, 95% CI [.761, 1.122], p > .05).

Table 22: Pairwise comparisons of within Item Group differences

Person	
1st vs 2nd	↑***
1st vs 3rd	↑***
2nd vs 3rd	↑***
Number	
1st: sg vs pl	↑***
2nd: sg vs pl	↓***
3rd: sg vs pl	↓***
Control	
Subj vs Obj	↑***
Expletive	
Seems vs Be	↑***
Topic	
No vs Shift	–

Item Group	Accuracy	95% CI	vs Chance	p-value
1st Singular	94.6%	[92.9, 95.9]	Above	< .001
1st Plural	91.0%	[89.0, 92.8]	Above	< .001
2nd Singular	65.9%	[62.8, 69.0]	Above	< .001
2nd Plural	79.3%	[76.5, 81.8]	Above	< .001
3rd Singular	54.5%	[51.2, 57.7]	Above	0.007
3rd Plural	71.8%	[68.8, 74.7]	Above	< .001
Subject Control	31.6%	[28.7, 34.7]	Below	< .001
Object Control	0.3%	[0.1, 0.9]	Below	< .001
Seems Expletive	41.3%	[38.2, 44.6]	Below	< .001
Be Expletive	56.5%	[53.3, 59.7]	Above	< .001
No Topic Shift	56.0%	[52.7, 59.2]	Above	< .001
Topic Shift	57.7%	[54.5, 60.9]	Above	< .001

Table 23: Overt subject preference by syntactic context at final checkpoint for the lemmatize verbs model

All processing manipulations showed preferences for overt subjects (see Table 21). The model preferred overt subjects 53.2% of the time in default contexts. There was no significant difference comparing the default with forms including long noun phrases (55.1%, OR = 1.078, 95% CI [+29, 1.267], p > .05) and embedded relatives (55.5%, OR = 1.098, 95% CI [.935, 1.290], p > .05). Further, the difference between complexity types was not significant (OR = .982, 95% CI [.818, 1.178], $p > .05$).

Context negation showed no difference from default (53%, OR = .992, CI [.845, 1.165,$p > .05$). However, there was a significantly increased overt preference compared to the default for target

Table 24: Pairwise comparisons of within Processing differences

Complex	
Emb vs Long	–
Negation	
Targ vs Cont	↑***
Targ vs Both	↑***

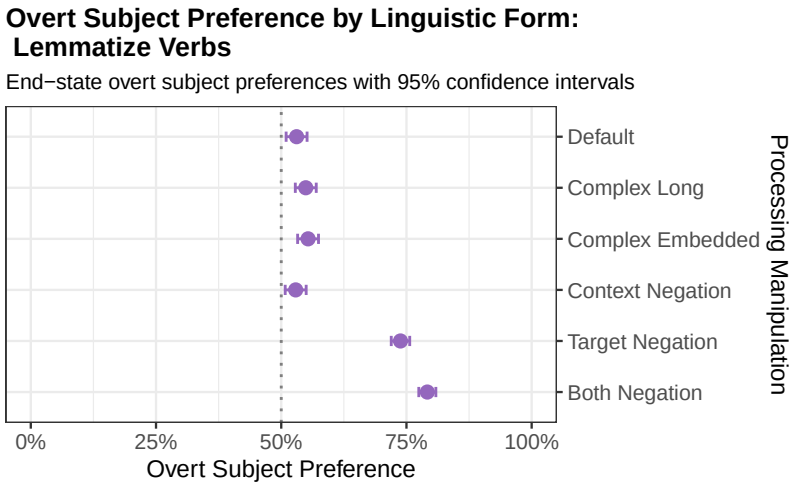

Figure 15: Model preferences for overt subjects by processing manipulation at final checkpoint.

negation contexts (74.3%, OR = 2.546, 95% CI [2.144, 3.023], p < .001) and constructions with negation in both the target and context (79.7%, OR = 3.453, 95% CI [2.885, 4.133, p < .001). Target negation significantly exceeded context negation (OR = .390, 95% CI [0.321, 0.473], $p < .001$), while target and both-negation conditions did not differ significantly (OR = 1.057, 95% CI [0.845, 1.323], $p > .05$).

Form	Accuracy	95% CI	vs Chance	p-value
Default	53.0%	[50.8, 55.2]	Above	0.007
Complex Long	54.8%	[52.6, 57.0]	Above	< .001
Complex Emb	55.3%	[53.1, 57.5]	Above	< .001
Context Negation	52.8%	[50.6, 55.0]	Above	0.012
Target Negation	73.8%	[71.8, 75.7]	Above	< .001
Both Negation	79.1%	[77.3, 80.9]	Above	< .001

Table 25: Overt subject preference by processing manipulation at final checkpoint for the lemmatize verbs model

Experiment 5

The dataset ablated was the previously ablated corpus from Experiment 1. The ablation excises all *subject pronominals*. After the first pass, replacement sentences are reinserted to preserve well-formedness, and the procedure is iterated; negative values in the token–change summaries indicate a net addition of tokens.

Pre-ablation size across sources was N=89,042,433 tokens. In total, 8,184,685 subject pronouns were removed, corresponding to 9.192% of all tokens (roughly one in eleven). Because replacement can add or remove material, the net token change was an *increase* of 227,736 tokens overall.

By source (subject pronouns removed; share of that source;

net token change):

- **BNC Spoken**: 892,344 (12.788%); net +65,143 tokens.
- **CHILDES**: 2,663,055 (10.249%); net −272,576 tokens.
- **Gutenberg**: 1,773,765 (7.473%); net +354,581 tokens.
- **OpenSubtitles**: 2,361,473 (13.159%); net −34,351 tokens.
- **Simple Wikipedia**: 338,283 (2.564%); net +128,945 tokens.
- **Switchboard**: 155,765 (12.910%); net −14,006 tokens.

Overall, the ablation targets a broad, well-delimited category while keeping corpus size broadly comparable, owing to controlled insertions during replacement.

Model Evaluation

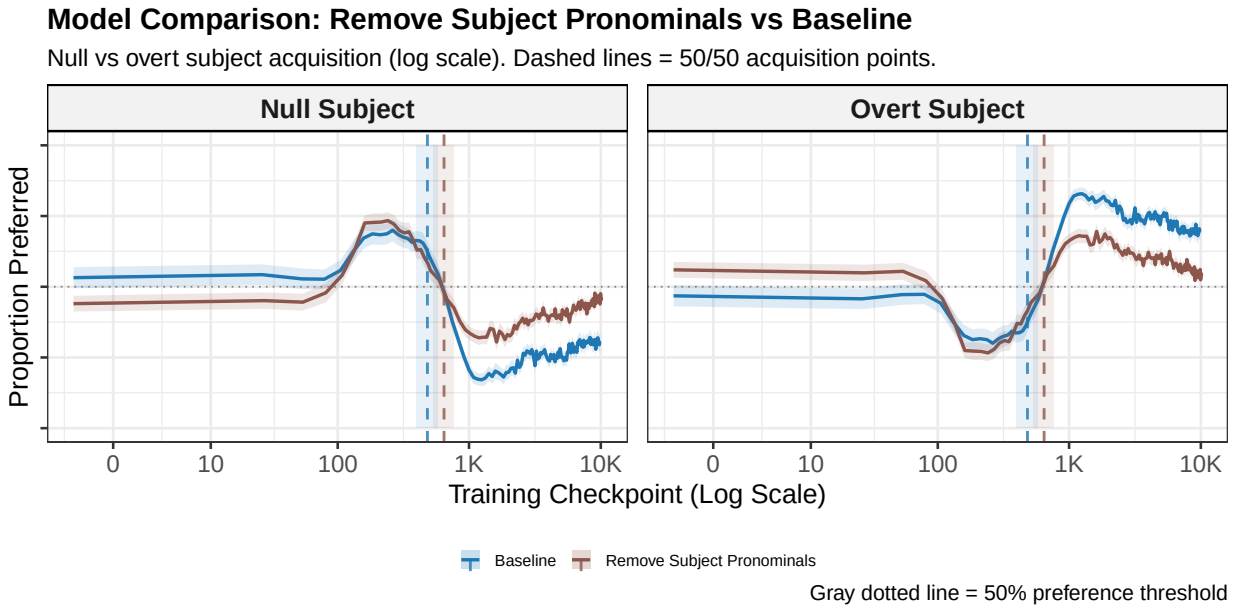

Figure 16: Model preference for null and overt evaluation stimuli over training, training steps transformed to log-scale.

AIC-based model selection indicated that the baseline model achieved optimal fit with 6 degrees of freedom (AIC = 166269). The model achieved t50 at checkpoint 646 (95% CI [536, 660]). Age of Acquisition analysis revealed that baseline achieved AoA at checkpoint 774 (95% CI [706, 9733[3]]). The ablated model reached acquisition criterion significantly earlier than the baseline model (ΔAoA = 80 epochs, 95% CI [35.9, 8974], p < .001).

First epoch analysis shows that the Baseline model exhibited a significant preference for null subjects by the end of the first epoch, showing a 62.3% preference for null subjects (95% CI 61.5,62.9], $p < .001$). This is not significantly different from the

3: Something about this condition made it hard for the function to estimate the right edge of the curve.

baseline model's first-epoch performance (63.5%, OR = 1.06 95% CI [1.009, 1.1], p = .085) when corrections for multiple comparisons is applied.

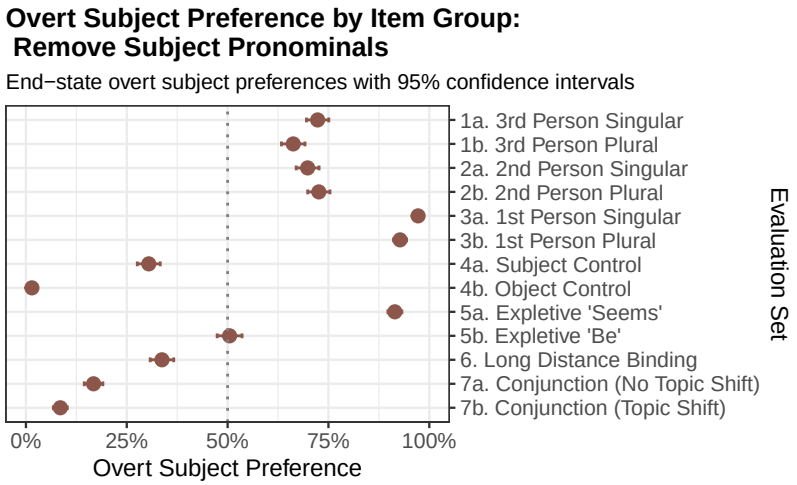

Figure 17: Model preference for overt subjects by evaluation group at final checkpoint

The end-state analysis showed that the base model strongly preferred overt subjects, with a 54.4% preference for overt subjects in the last two epochs of training (95% CI [53.4%, 55.3%], p <.001). This is significantly lower than the baseline model's preference (69.3%, (OR = 1.93, 95% CI [1.82, 2.03], p < .001).

Mixed-effects pairwise comparisons revealed significant person-based differences. First person contexts (96.4%) elicited significantly more overt subjects than both second person (74.8%, OR = 9.037, 95% CI [6.698, 12.193], p < .001) and third person contexts (72.9%, OR = 9.999, 95% CI [7.416, 13.482], p < .001). There was no significant difference found between 2nd and 3rd person contexts (OR = 1.106, 95% CI [.918, 1.333], p > .05). ˇ First person singular (98%) and plural (94.8%) contexts were shown to be significantly different (OR = 2.686, 95% CI [1.648, 4.379], p < .001). Second person singular contexts (73.3%) elicited no significant difference than plural contexts (76.4%, OR = .851, 95% CI [.682, 1.063], p > .05). Finally, third person singular contexts (76%) showed significantly higher preference for overt subjects than plural contexts (69.7%, OR = 1.375, 95% CI [1.104, 1.712], p < .01).

Mixed-effect models were unable to converge comparing control contrasts because of Perfect Separation. In these cases the data are instead analyzed with Fisher's Exact test. A significant subject-object asymmetry was observed with subject control contexts (30.4%) showing dramatically higher overt preferences than object control contexts (1.6%, p < .001).

Expletive constructions showed differential behavior by verb type. *Seems*-constructions strongly favored overt subjects (95.5%), than *be*-constructions (49.6%, OR = .046, 95% CI [.033, .065], $p > .001$). Like the baseline model, *be*-like constructions do not differ from chance in overt preference, see Table **??**.

The model shows a significant difference between topic-shift contexts, with a weaker preference for non-topic shift (15.9%) than topic shift contexts (7.8%, OR = 2.218, 95% CI [1.643, 2.995], $p < .001$).

Item Group	Accuracy	95% CI	vs Chance	p-value
1st Singular	97.2%	[95.9, 98.2]	Above	< .001
1st Plural	93.1%	[91.2, 94.7]	Above	< .001
2nd Singular	70.1%	[67.0, 73.2]	Above	< .001
2nd Plural	73.0%	[69.9, 76.0]	Above	< .001
3rd Singular	72.7%	[69.6, 75.6]	Above	< .001
3rd Plural	66.8%	[63.5, 69.9]	Above	< .001
Subject Control	30.4%	[27.4, 33.6]	Below	< .001
Object Control	1.6%	[0.9, 2.7]	Below	< .001
Seems Expletive	91.4%	[89.4, 93.2]	Above	< .001
Be Expletive	50.3%	[47.0, 53.7]	At	0.865
No Topic Shift	16.8%	[14.3, 19.4]	Below	< .001
Topic Shift	8.4%	[6.7, 10.5]	Below	< .001

Table 26: Overt subject preference by syntactic context at final checkpoint for the remove subject pronominals model

Table 27: Pairwise comparisons of within Item Group differences

Person	
1st vs 2nd	↑***
1st vs 3rd	↑***
2nd vs 3rd	–
Number	
1st: sg vs pl	↑***
2nd: sg vs pl	–
3rd: sg vs pl	↑***
Control	
Subj vs Obj	↑***
Expletive	
Seems vs Be	↑***
Topic	
No vs Shift	↑***

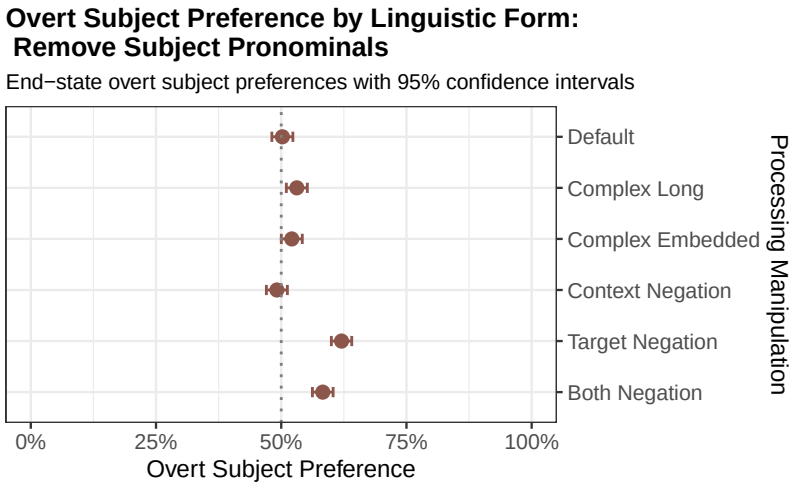

Figure 18: Model preferences for overt subjects by processing manipulation at final checkpoint.

All processing manipulations showed preferences for overt subjects (see Table 21). The model preferred overt subjects 67.6% of the time in default contexts. There was no significant difference comparing the default with forms including long noun phrases (68.5%, OR 1.043, 95% CI [.955,1.368], $p > .05$) and embedded relatives (70.4%, OR 1.143, 95%). Further, the difference between complexity types was not significant (OR = .912, 95% CI [.745, 1.118], $p > .05$).

Context negation showed no difference from default (66%, OR .988, CI [827, 1.179,$p > .05$). However, there was a significantly increased overt preferences compared to the default for target negation contexts (78.4%, OR = 1.738, 95% CI [1.439, 2.100], $p < .001$) and constructions with negation in both the target and context (77.4%, OR = 1.644, 95% CI [1.363, 1.983). Target negation significantly exceeded context negation (OR = .568, 95% CI [0.456, 0.703], $p < .001$), while target and both-negation conditions did not differ significantly (OR = 1.057, 95% CI [0.845, 1.323], $p > .05$).

Table 28: Pairwise comparisons of within Processing differences

Complex	
Emb vs Long	–
Negation	
Targ vs Cont	↑***
Targ vs Both	–

Form	Accuracy	95% CI	vs Chance	p-value
Default	50.3%	[48.0, 52.6]	At	0.799
Complex Long	53.5%	[51.2, 55.8]	Above	0.003
Complex Emb	52.5%	[50.2, 54.7]	Above	0.035
Context Negation	49.4%	[47.1, 51.7]	At	0.627
Target Negation	62.1%	[59.8, 64.3]	Above	< .001
Both Negation	58.5%	[56.2, 60.7]	Above	< .001

Table 29: Overt subject preference by processing manipulation at final checkpoint for the remove subject pronominals model

Discussion

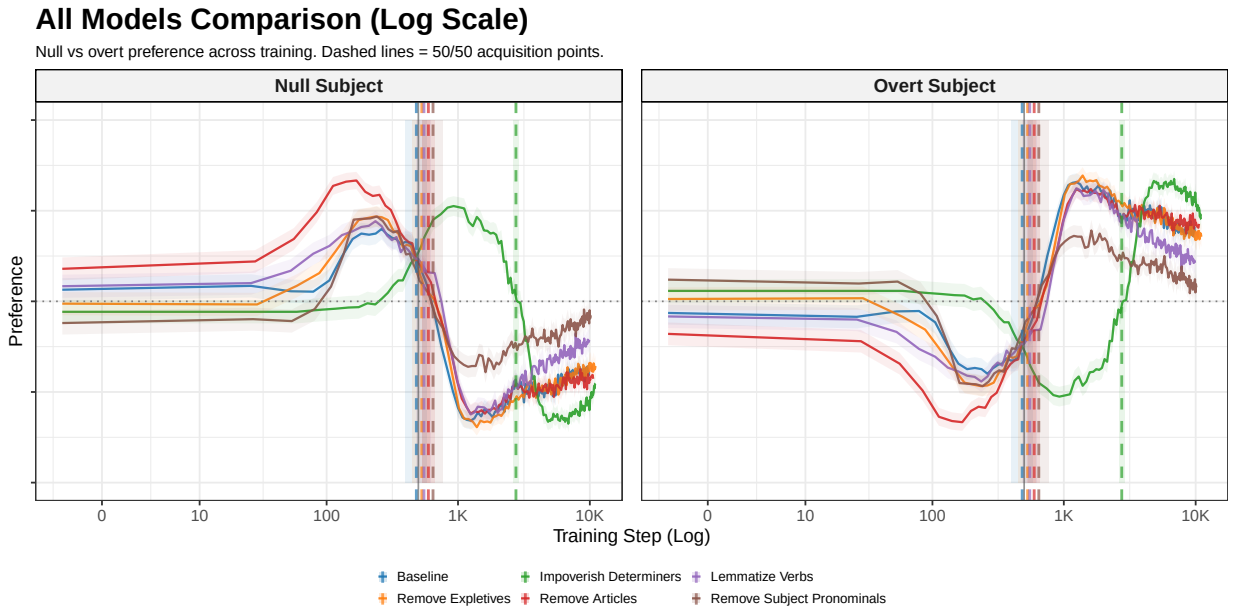

Figure 19: Cross-model comparison of null subject acquisition trajectories (log scale)

In this study, I used small Transformer models trained on differently ablated linguistic corpuses to investigate the causal

role that different sources of linguistic evidence play in a model forming syntactic preferences. In addition, I investigated the role that complexity within a production context would influence the performance of the model across different experiments. The goal was to investigate theories of evidence and performance on English children's mis-production of null subjects while learning an overt-subject language.

After training the baseline model, we found that models fail to learn the overt subject constraint within the first 90 million words of training presented – the amount of words that a 9 - 12 year old roughly has seen. By that age, English children uniformly produce overt subjects in adult-like ways. The model eventually did become more like an English speaker in its preference for overt subjects roughly after 1.5 epochs, maintaining a preference for overt subjects. The preference decreased (becoming more flexible) somewhat over log-time.

The baseline model, and all ablated models, demonstrate a stage before learning the English generalization, with a preference for null subjects to overt subjects. It is unclear what is the source, or evidence, for null subjects in English that the model is sensitive to in the early steps of training, or whether this is an artifact of a bias within the model. That is to say, whether the model defaults to a state that is available to null subjects and only learns otherwise with sufficient evidence later in training, or whether there is input within the English corpora that when presented to the model guides it towards a preference for null subjects. Our ablative experiments failed to eliminate this stage, at times increasing it. It is possible that insufficient lexical evidence could lead to an illusion of a null-subject preference in the sense that the model dis-prefers the presence of a pronoun due to unfamiliarity and not due to a linguistic generalization.

Such a stage would be consistent with null-first accounts of overt subject learning, some of these theories attribute such a preference to innate learning biases (a kind of parameter setting) [67, 73], while others may attribute it to something like an artifact of an easy-first or good-enough bias [38, 55, 62, 87]. This is however evidence counter to Bloom's [62] prediction that English children default to overt subjects, despite processing effects.

[73]: Hyams (1986), *Language acquisition and the theory of parameters*
[67]: Hyams (1989), "The null subject parameter in language acquisition"

there may be biases present within transformers that would bias them early towards null subjects. It's also possible there is information available early on that leads them to the wrong preference. It's not necessarily the case that the model learns in such a way, and thus we should think that children learn such a way; rather, the model demonstrates that there is either evidence

sufficient for English children to acquire an early null subject bias available in the environment (as such a thing as a Transformer can learn), or some other bias could be responsible for the effect. The former question can be asked further with more ablative work, looking for the causal source that leads to this initial bias, whereas the latter question is better served by manipulating model architecture and developing interpretability experiments to better understand the model's performance.

Model	AOA	CI
Lemmatize Verbs	705.00	[661, 749]
Baseline	727.00	[665, 792]
Rmv. Expletives	767.00	[709, 821]
Rmv. Subject Pronominals	775.00	[707, >5000]
Rmv. Articles	808.00	[759, 861]
Impvr. Detrmn.	3400.00	[3307, 3499]

Table 30: Age of acquisition for null subject preference across experimental models

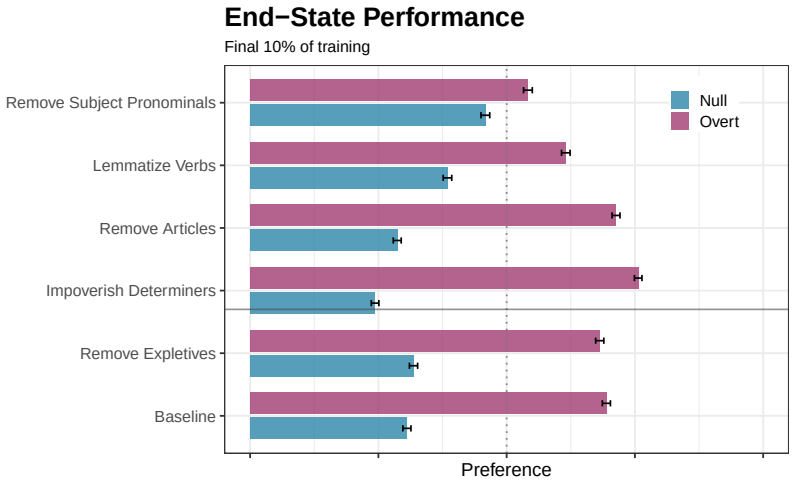

Figure 20: Final checkpoint performance comparison across all experimental models

Despite the fairly wide-reaching interventions we performed on the training data, all models successfully acquired an english-like generalization fairly early within training, although most models' preferences differed from each other by the end of training. The first manipulation, to remove expletives, influenced the model's behavior the least. While the ablated model took longer to acquire the generalization, by 40 training steps, it otherwise did not differ in terms of changing the state of the model's training at the end of the first checkpoint and the last. This is unlike the other ablated models. Nonetheless, this makes an account like Yang's [76, 77] variational learning theory of null subjects unlikely to explain the model's behavior. However, it serves to question whether this

[76]: Yang (2003), *Knowledge and learning in natural language*
[77]: Yang (2004), "Universal Grammar, statistics or both?"

result is adequate given that the model which lacks evidence of unbound expletives, still shows preference for their presence in seems-like and be-like constructions (even to a greater extent). Either the ablative intervention did not sufficiently remove all expletives (which would still not explain the additive behavior), or other sources of evidence are contributing to the model's performance in these cases. This is all to say, even with fairly robust perturbation, there seems to be sufficient evidence present within the BabyLM dataset to acquire such a preference.

The second experiment demonstrated the most dramatic learning effect of an ablation: despite the fact that experiment 2 did not remove any words, only impoverishing forms, learning was delayed 5 full epochs, at least. Despite this fact, the model acquires the strongest preference for overt subjects than any other model. What makes this manipulation different from the others, in its outsized effect?

The third experiment, was predicted to be even more extreme than the second, removing all articles, but this did not show nearly the same effect; instead showing the strongest initial preference for null subjects between models within the first epoch. This manipulation did in-fact slow the acquisition of the generalization more than any other experiment besides the second. However, it simply doesn't compare in magnitude or character.

This is problematic for Hyam's [67, 68, 73] accounts which do not differentiate between impoverishment due to uniformity or absence, but also that do not consider articles to be primary evidence for the acquisition of this structure. Instead, Duguine [78] proposed that a rich determiner system, and the absence of a rich verbal paradigm would lead to the generalization. That kind of account would seem to corroborate with model performance [79]. Duguine treats evidence of a rich determiner system as a kind of 'blocking' rule, which allows for the blocking of the null-subject parameter. It could be that in the absence of evidence for a kind of 'blocking' rule, you are forced to search for a less shallow source of evidence. This could extend the time to learn a construction, but being forced to form such a generalization in a less shallow may also encourage a stronger generalization with more acceptable dropping of both subject and object pronouns in allowed contexts.

[79]: Bertolino (2024), "The setting of the null subject parameters across (non-)null-subject languages"

That is not to say that articles alone are important to learning this generalization (as it succeeds anyways), as Experiment 4 demonstrates, that among the targeted sources of indirect evidence (exps. 1-4), the absence of verbal morphology leads to the lowest overt subject preference. Note that in this ablation,

the status of determiners or pronouns are all maintained, and in the absence of verbal agreement or tense information, the likelihood of overt subjects decreases significantly. Future work will look at the effect on learning when both verbal and determiner information is impoverished. Under a topology of non-null-subject languages, impoverished verbal information should lead to more Mandarin-like generalizations.

Despite the outside role of indirect evidence in this discussion, Experiment 5 directly tested the role that direct pronominal evidence has on learning. Hyams [67, 73] proposed that children can learn from the presence of a pronoun that they may be learning a non-null-subject language if they assume that there is some pressure for speakers to otherwise not speak a pronoun, and that the presence of a pronoun is evidence of its necessity, whether this constraint is pragmatic or syntactic/semantic in nature. In fact, Experiment 5 while demonstrating a slight but significant bias for overt subjects, in-fact shows no bias whatsoever in non-processing related contexts, that is you can see that the default forms without any manipulation do not significantly differ from chance. So, you could say in this case that the model does in fact fail to acquire the generalization in the case that direct evidence is not available. If the model is able to utilize the presence of pronouns as such evidence, it may indicate that children themselves are able to utilize the presence or absence of a pronoun as positive evidence towards either generalization, as Hyams predicted. Such questions should still be addressed by experimentally investigating children's learning, but these models offer a tool to test such questions with methods that are unavailable in human work.

Table 31: Syntactic forms showing significant deviation from default performance by experimental model

Form	Baseline	Rmv. Expletives	Impvr. Detrmn.	Rmv. Articles	Lemmatize Verbs	Rmv. Subject Pronominals
Complex Long	✓		✓			
Complex Emb			✓			
Context Negation						
Target Negation	✓	✓		✓	✓	✓
Both Negation	✓	✓	✓	✓	✓	✓

In addition to testing the model's performance across training, we also compared model performance on specific stimuli designed to address questions about the role that processing load (or specifically the complexity of a processing environment)

Table 32: Pairwise comparisons between processing manipulations across experiment model

Test	Baseline	Rmv. Expletives	Impvr. Detrmn.	Rmv. Articles	Lemmatize Verbs	Rmv. Subject Pronominals
Complex						
Emb vs Long	–	–	–	–	–	–
Negation						
Targ vs Cont	↑***	↑***	–	↑***	↑***	↑***
Targ vs Both	–	–	↑***	–	↑***	–

has on preferences for overt subjects. We found that, counter to processing accounts of null-subject use in children [29, 62, 63, 88], the model when faced with extra complexity instead seems to prefer overt pronouns more, to the extent that the positive result for Experiment 4 is from the negated forms increasing the overall rate of pronoun use in the model. Across the board, we do not see any models that are influenced by the presence of negation in the context sentence, whereas the presence of negation in the target sentence in almost all cases results in increased overt subject use.

Of course, it could be said that whatever manifests as processing difficulty within such a model does not reflect how processing difficulty is managed by speakers. Whereas surprisal has been shown to capture much about comprehender's difficulties reading, less work has been done as to how it captures language production[1]. Some work, however, for instance, has shown that the dynamics of Large Language Models, when trained for such a task, can predict hierarchical aspects of language planning [89], so it is not strictly out of the question that production effects could be captured in such a way. It could be in this case that parallel work needs to be done with adult human participants to determine whether this effect reflects human-like processing artifacts, or this is instead an artifact of the model's processing alone.

Similarly, preferences across processing stimuli seem to follow similar patterns of acquisition across all stimuli, which would suggest the model is forming a fairly general heuristic which is robust to differences across contexts. One place of note is the presence of negation in the target sentence seems to eliminate the initial null-bias across its learning curve. This is still counter to the predictions made by the processing account, as this bias disappears as a preference, or lack of bias towards null for overt subjects. This can be seen in Figure 21

1: Except in the realm of aphasiology where there has been much work done, although not exactly relevant to the current discussion

[89]: Mingfang et al. (2025), "From thought to action: How a hierarchy of neural dynamics supports language production"

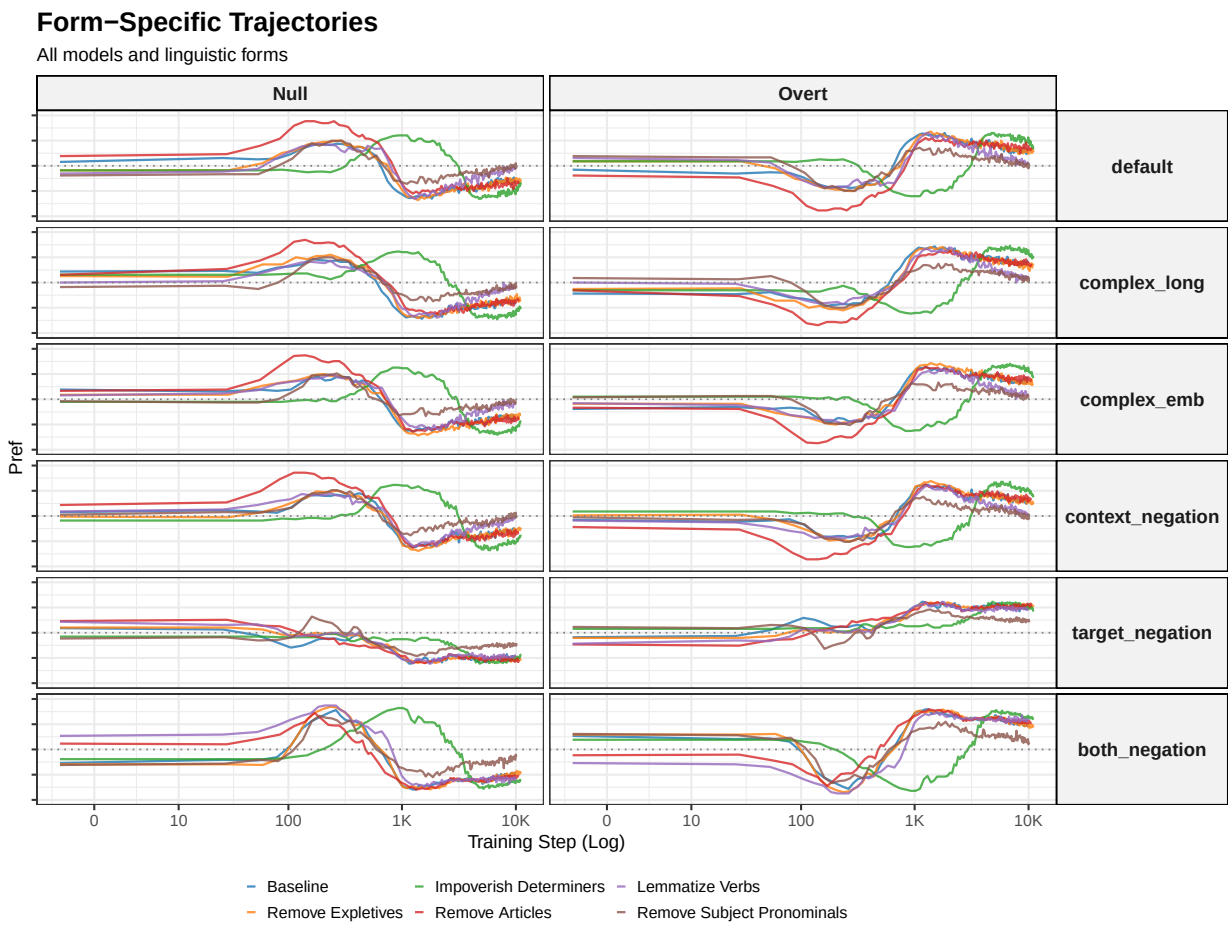

Figure 21: Developmental trajectories of null subject preferences by processing manipulation across all models

We can see fairly robust person asymmetries in the models. In general, models prefer overtly producing 1st person pronouns over 2nd and 3rd person pronouns. It is possible that this is the result of the kind of evidence available, vs specific characteristics as to how the model processes different pronouns. That is, in general you would expect to see more 2nd and 3rd person referents in imperative contexts. Although, this would not necessarily capture the whole picture, as diary drop '__didn't go to work today' involves only 1st person pronouns (in the case of declarative statements, unlike 'didn't go to work today?'). Only in cases where we impoverished determiners, or lemmatized the verbs did we see a full person hierarchy, with a further difference between 2nd person and 3rd person. Regardless, this demonstrates, that even with the model having learned the English preference fairly robustly across items, there is still a graded difference between syntactic and lexical contexts that would influence this behavior. This pattern is fairly coherent with languages that allow for

Table 33: Cross-model pairwise comparisons of null subject preferences by syntactic context

Test	Baseline	Rmv. Expletives	Impvr. Detrmn.	Rmv. Articles	Lemmatize Verbs	Rmv. Subject Pronominals
Person						
1st vs 2nd	↑***	↑***	↑***	↑***	↑***	↑***
1st vs 3rd	↑***	↑***	↑***	↑***	↑***	↑***
2nd vs 3rd	–	–	↑***	–	↑***	–
Number						
1st: sg vs pl	↑***	–	↑***	↑***	↑***	↑***
2nd: sg vs pl	↓***	↓***	↑***	↓***	↓***	–
3rd: sg vs pl	–	–	↓***	↓***	↓***	↑***
Control						
Subj vs Obj	↑***	↑***	↑***	↑***	↑***	↑***
Expletive						
Seems vs Be	↑***	↑***	↑***	↑***	↑***	↑***
Topic						
No vs Shift	–	↓***	↓***	↑***	–	↑***

subject drop, where there are proposed person, and number hierarchies which determine how arguments are connected with their verbs. The study at present did not specifically investigate the topic of verbal agreement (compared to verbal morphology independently), but in languages that allow for robust person-based verbal agreement, such as Spanish, there are theories that differentiate the three persons in a pattern that follows from the results of this study [90, 91] . Some even argue that 3rd singular pronouns in many languages are not pronouns at all, but instead determiners ('el' → 'el', 'la' → 'ella') [92]. Among the biggest changes present in the within Itemgroup differences are in Experiment 2 and 4, impoverishing determiners and lemmatizing verbs effect preferences for 3rd person singular pronouns more than its plural counterpart towards a bias for null subjects. In addition, these two conditions also introduce a difference between 2nd and 3rd person pronouns in their production[2]. To better elucidate this effect, it may be interesting to perform a representational analysis on 3rd singular pronouns relative to other pronouns, and likewise to do such an analysis across models to see how representations change from ablation. We would predict, for example, that following a kind of analysis like Bonet's [92], that the more Italian-like a model becomes, the 3rd singular should become more like a determiner.

[90]: Harley et al. (2002), "Person and number in pronouns: A feature-geometric analysis"

[91]: Carminati (2005), "Processing reflexes of the Feature Hierarchy (Person > Number > Gender) and implications for linguistic theory"

[92]: Bonet (1991), "Morphology After Syntax: Pronominal Clitics in Romance"

2: This question, of the status of the 3rd person pronoun was one I was quite interested in when I used to study syntax. I had some independent evidence from Basque that syntactically, the 3rd singular object pronoun had such a status. Maybe there is also a subject-object asymmetry there as well as noted in the Carminati [91] paper

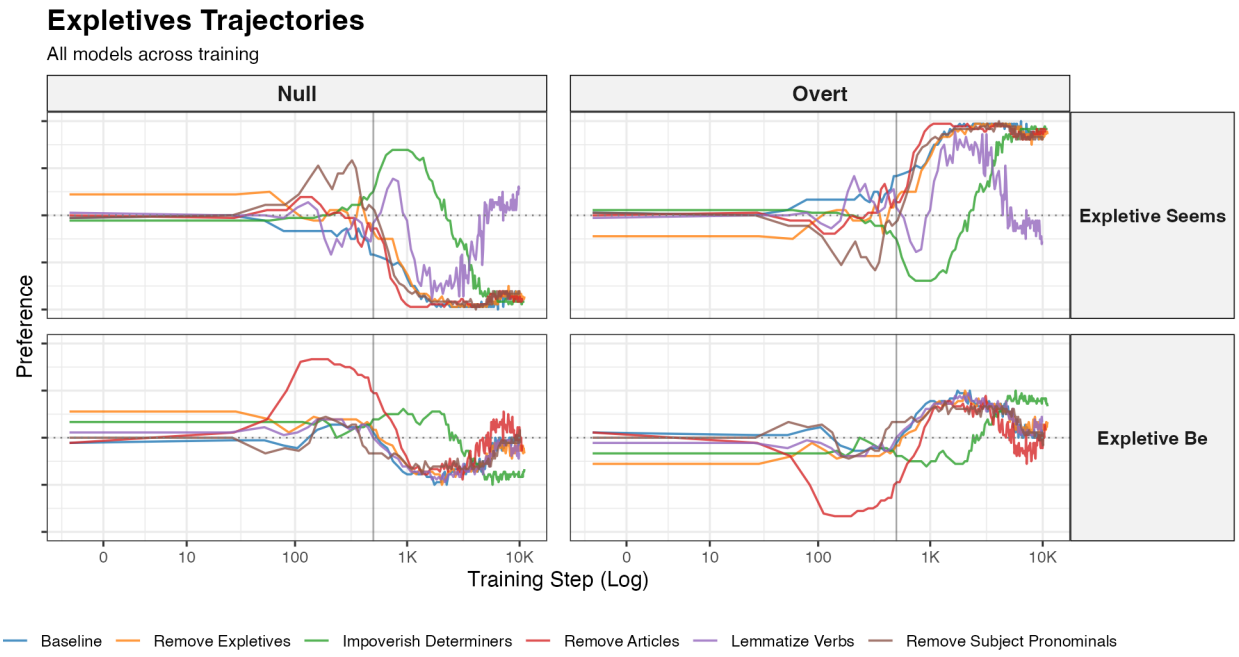

Figure 22: Learning trajectories for expletive constructions across experimental models

All models demonstrated the human-like asymmetry between subject and object control verbs, preferring to drop object controlled pronouns far more than subject controlled pronouns. It is of note that, unlike the other models: Experiment 2, while maintaining the same asymmetry, incorrectly prefers overt pronouns in subject control contexts.

The models seem, for the most part, capable of developing a fairly strict preference for expletives in *seems*-type verbs[3]. In the case of Experiment 4, lemmatize verbs, the model at some point reverses a preference for overt expletives towards one for null expletives — although this is reflected as a slight-overt bias when averaged over the last two epochs. On the other hand, all models, including the baseline model, struggle to form a strong preference for *be*-like constructions. In contrast, when impoverishing determiners the model forms a later generalization for the expletive be, but its final impression is both correct, and stronger than other models. Such an effect could support the account that this model, when faced with the ablation developed a less shallow heuristic despite a longer acquisition time. Likewise for *be*-like constructions, only Experiment 2 and 3 seem to demonstrate an early null-bias, and in the case of Experiment 3, this bias is very large. The lack of such an effect in the *seems*-like contexts, could indicate that the development of English-like performance in *be*-like contexts is more dependent upon evidence

3: I say seems-*type* here, but indeed every verb in the set is 'seem'

from articles, whereas the presence of verbal evidence is more important for *seems*-type verbs, as demonstrated by Experiment 4.

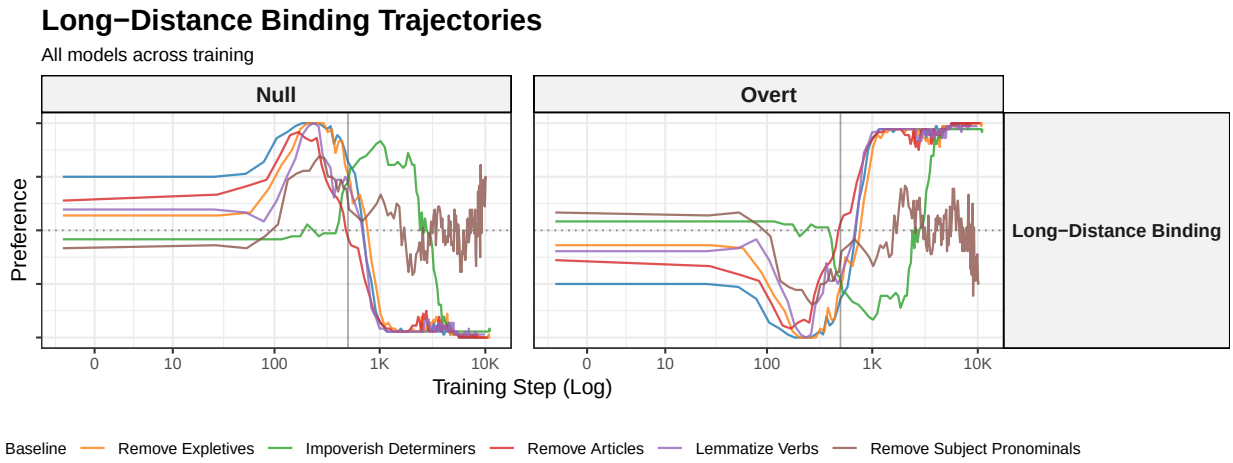

Figure 23: Learning trajectories for long-distance binding constructions across experimental models

All models except for the model with ablated subject pronouns successfully acquired a strong preference for overt pronouns in long-distance binding contexts (where such binding is elicit and otherwise required to drop a pronoun). It seems then, that most models are able to acquire this constraint, which is largely bound by syntactic dependencies. However, the question is, why is there such variability in the performance of the model trained without subject pronouns? The most obvious answer to this, is of course, that there is simply much less exposure to pronouns in that model, especially in the subject position which these stimuli were targeting. Still, there should be plenty of independent evidence that should allow the model to acquire similar binding constraints, an heuristic that shouldn't necessarily be lexically-bound. Perhaps in this area we should cross-reference performance on this particular binding task and performance on other binding tasks, such as those found in the BLiMP evaluation setw.

In the case of conjunction without topic shift and conjunction with topic shift, models, except for Experiment 5, correctly preference overt pronouns, but the baseline model fails to acquire a difference between topic shift contexts. However, four out of five ablated models acquire a distinction, Experiments 1 and 2 acquire the correct distinction, whereas Experiments 3 and 5 acquire the incorrect one. Experiment 1 prefers overt subjects 4.7% (OR = .726, 95% CI [0.557, 0.946], $p < .001$) more in topic shift contexts, whereas Experiment 2 shows a much stronger

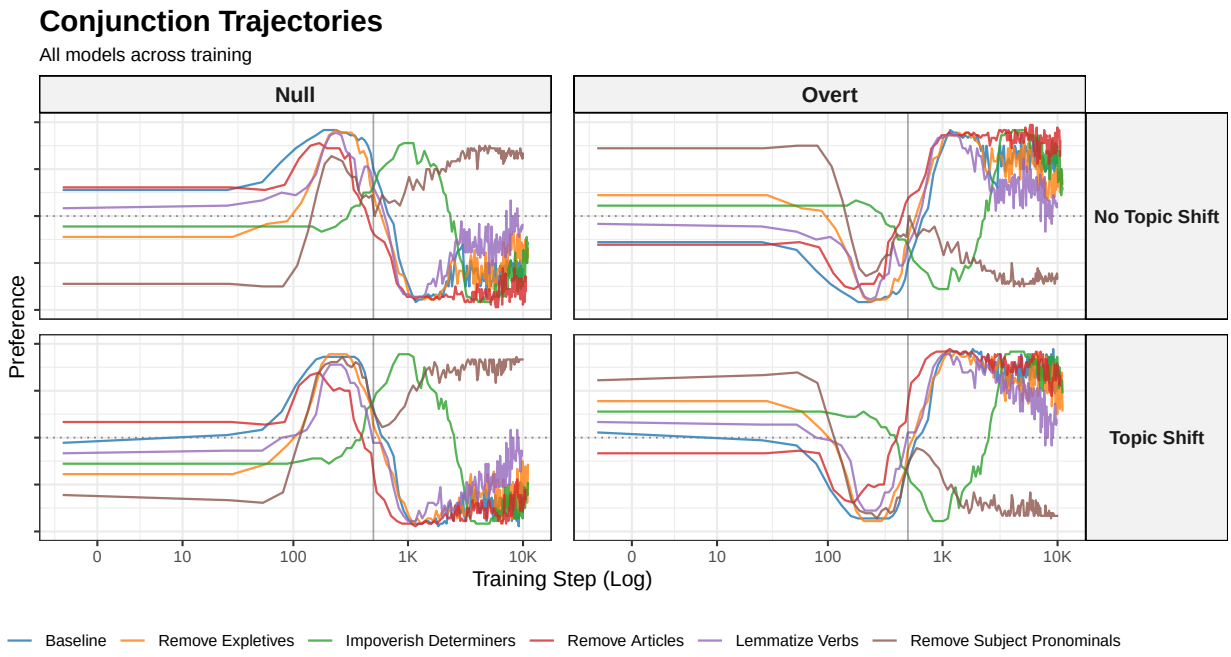

Figure 24: Learning trajectories for person-number agreement constructions across experimental models

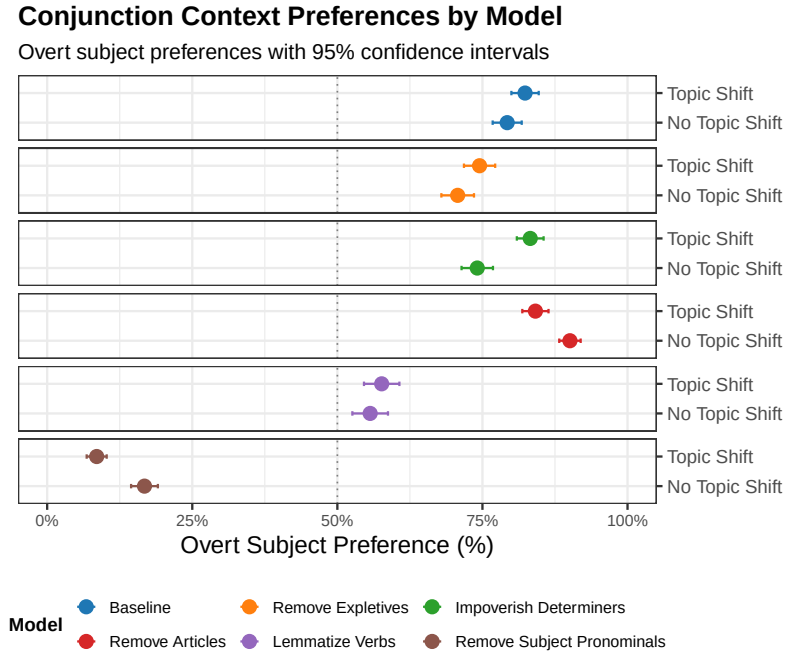

Figure 25: Overt subject preference in conjunction contexts by model

contrast of 9.4% (OR = .556, 95% CI [0.436, 0.710], $p < .001$). It's possible the effect that we're seeing when determiners are impoverished, is part of the same phenomenon that we have been tracking with its progress throughout this paper. That is, perhaps, in the course of its long acquisition of these constructions,

it is actually establishing some deeper heuristics that in-fact are stronger, or more human-like in their content. Perhaps it has deeper competence that could be examined in BLiMP. We propose this as a possible explanation for the performance of the model when impoverishing determiners.

The idea that a model, when certain short-cut features are removed, will learn something slower but more robustly is not a new idea in the training of Large Language Models. This idea has been captured in a literature about a phenomena called 'Grokking' [93]. Essentially, models which have long since overfit on the training data available can experience sudden and extreme improvements in their performance on the validation data, through an emergent process called grokking. They find also that as their training set gets smaller, the amount of training steps to generalize to the validation data takes longer. While in this case, grokking is about a model which reaches perfect performance on a test case, we do not expect the model to reach such performance in this complex linguistic context. Some work suggests that what models are doing in these scenarios is transitioning from memorization-like states to generalization-like states in its internal state [94]. They show that if you re-introduce new sources of data to the model, it can revert back from generalization-like states, towards a memorization-like state, losing its performance gains. Perhaps we would see such a case comparing these ablated models in bilingual contexts, when they are later exposed to a new language.

[93]: Power et al. (2022), "Grokking: Generalization beyond overfitting on small algorithmic datasets"

[94]: Varma et al. (2023), "Explaining grokking through circuit efficiency"

It's possible that while determiner richness is a very available piece of evidence for the model, it introduces the model to a kind of short-cut learning [95], which in the long-term could introduce shallow heuristics that harm out of distribution performance. On the other hand, not removing, but instead un-enriching the determiner morphology removes this as an obvious and available source of evidence for models, leading it to richer more robust generalizations. Of course, this process may not lead to a stronger model overall, this model was the only one that preferred overt pronouns in control conditions. Perhaps acquiring a stronger generalization for overt subjects, like the one in English could cause the model to incorrectly apply this rule in a control context. In this case, it may be considered that this model is in the middle of a U-like curve, first memorizing, then forming generalizations which mis-apply to incorrect contexts, and perhaps if training continued it may have corrected such mistakes.

[95]: Roman (2024), *Shortcut learning of large language models in natural language understanding*

In fact, looking at the learning patterns for control contexts from Experiment 2, one can see that this result is an artifact of

averaging over the last two epochs of training. You can observe, that the model after learning that overt subjects are obligatory soon rapidly corrects this rule in control contexts, making a U-shaped curve towards English-like performance (it's important to note that other models showed such a u-shaped curve for this stimuli, just at earlier points).

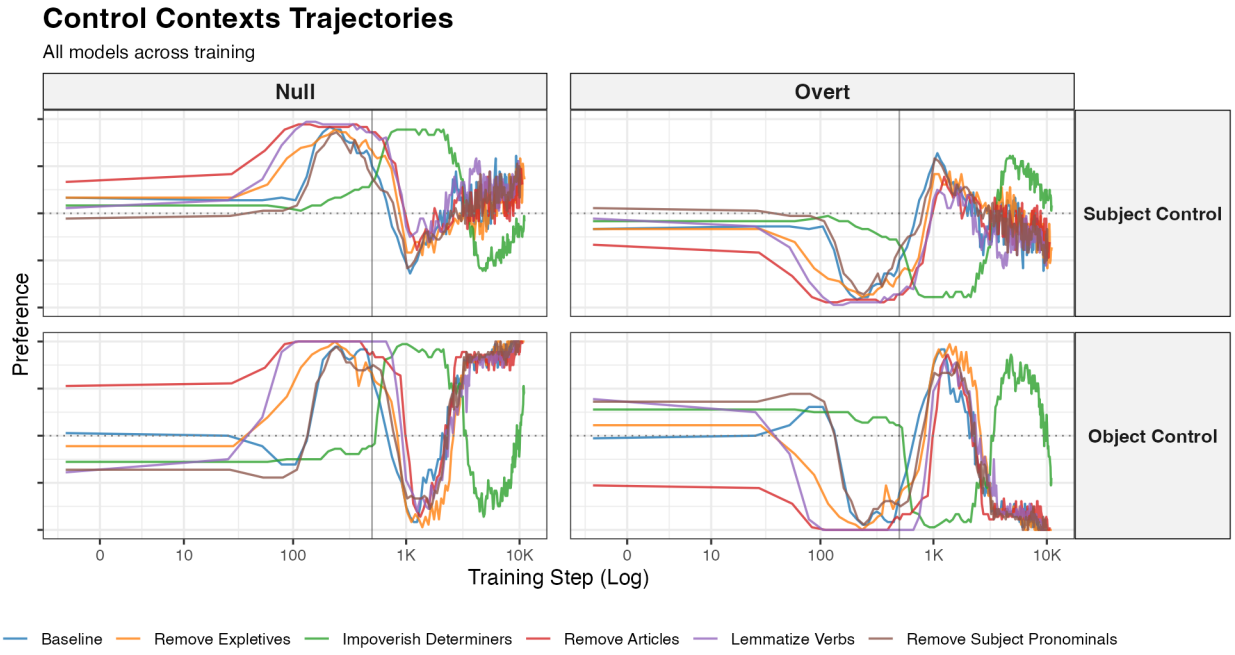

Figure 26: Learning trajectories for control constructions across experimental models

Future Directions

This study is by no means perfect, and is a first try at attempting a study of this kind for me. Some things that are notably missing by the time this proposal is sent out.

1. More cross-wise experiments investigating the role of more than one source of linguistics evidence, manipulating multiple kinds of ablations.
2. Simple BLiMP or test perplexity analysis (I have all the stuff there, just couldn't get to it)
3. Looking closer at the distributional changes that appear as part of the ablative task, designing evaluation stimuli to test whether the ablation was correctly target, or certain linguistic information was lost

4. Discussion human-size LLMs, etc. Looking at performance within context to see how the model processed it's own training data.
5. look at the baseline, and other models across random seeds, learning rates, other hyperparameters etc.
6. a.o. this was getting to be quite long and cumbersome.

These are my future plans:

1. Perform representational analysis on these models to track how changes in representations occur over time, and whether this tracks with how/when learning occurs over time.
2. Look at whether within-language priming effects occur for the presence of absence of a subject.
3. Run bilingual language studies investigating the role the generalizations formed in one language influence the learning of another language, for instance, would freezing a model at the point where it prefers null subjects in one Language transfer an advantage in learning another language that allows for null subjects
4. Can we ablate a null-subject language in such a way that it behaves like an overt subject language?
5. We've talked about data ablations, but can we introduce rearing that allows for data manipulation? What if we artificially insert a rich agreement system into a language that otherwise lack one. Perhaps Mandarin is a good investigative domain for this.

Fit any one of these into a Chapter 2 and 3, I currently do not have the energy to do so myself. I will flesh two of these out at least on the proposal date. I also presented several ideas within the discussion itself, and this does not mention further investigations that I will put specifically towards the behavior of the model in Experiment 2.

Conclusion

This proposal sought to investigate the role that different sources of causal evidence plays in the learning of the null-subject constraint. It also sought to assess theories of processing difficulty leading to non-english like behavior in English-learning children. We found that LLMs were unable to learn the English overt subject constraint given the same amount of linguistic stimuli available to children after processing it once through. Instead,

the model demonstrated the same non-english like behavior that young children do, preferring null subjects in positions where null subjects are elicit. Despite that, within the second epoch most models were able to develop an English-like preference for overt subjects. We found that the presence of different kinds of linguistic information significantly impacts the learning of this generalization, and that specifically the presence of a rich determiner system is important to quickly guide the model towards English-like subject rules. In addition, we found that processing-based manipulations of the data do not make the model behave in ways that English children do, instead under increased contextual complexity Transformer models omit subjects less. Either this is a difference between models and children, in how a model and child responds to complexity, or it suggests that we should look further at the causes of children's performance in these cases. The results of the learning study suggest that maybe removing some sources of information, like rich determiner morphology can herd a model towards processing more data over longer period of time to develop a stronger and more robust generalization. In this case, this could be an instance of adaptive Grokking as a result of linguistic deprivation of the model.

Bibliography

[1] Jeffrey L Elman. “Finding structure in time”. en. In: *Cogn. Sci.* 14.2 (Mar. 1990), pp. 179–211.

[2] David E Rumelhart, James L McClelland, and AU. *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations*. en. The MIT Press, July 1986.

[3] Willem J M Levelt, Ardi Roelofs, and Antje S Meyer. “A theory of lexical access in speech production”. In: *Behav. Brain Sci.* 22.01 (Feb. 1999), pp. 1–75.

[4] Franklin Chang. “Symbolically speaking: a connectionist model of sentence production”. en. In: *Cogn. Sci.* 26.5 (Sept. 2002), pp. 609–651.

[5] Franklin Chang, Gary S Dell, and Kathryn Bock. “Becoming syntactic”. en. In: *Psychol. Rev.* 113.2 (Apr. 2006), pp. 234–272.

[6] Sida I Wang and Christopher D Manning. “Baselines and bigrams: Simple, good sentiment and topic classification”. In: *Annu Meet Assoc Comput Linguistics* (July 2012). Ed. by Haizhou Li et al., pp. 90–94.

[7] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. “A simple but tough-to-beat baseline for sentence embeddings”. In: 2017.

[8] Ian Tenney, Dipanjan Das, and Ellie Pavlick. “BERT rediscovers the classical NLP pipeline”. In: *arXiv [cs.CL]* (May 2019).

[9] Alec Radford et al. “Language Models are Unsupervised Multitask Learners”. In: *OpenAI* (2019).

[10] Christopher D Manning et al. “Emergent linguistic structure in artificial neural networks trained by self-supervision”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 117.48 (Dec. 2020), pp. 30046–30054.

[11] Christopher Potts. *Finding linguistic structure in large language models*. LSA Annual Meeting. Jan. 2025.

[12] S T Piantadosi. “Modern language models refute Chomsky’s approach to language”. en. In: *lingbuzz.net* (July 2024).

[13] Michael C Frank and Noah D Goodman. “Cognitive modeling using artificial intelligence”. In: *PsyArXiv. Retrieved from osf. io/preprints/psyarxiv/wv7mg v1 doi* 10 (2025).

[14] Richard Futrell and Kyle Mahowald. “How linguistics learned to stop worrying and love the language models”. In: *arXiv [cs.CL]* (Jan. 2025).

[15] Joe Pater. “Generative linguistics and neural networks at 60: Foundation, friction, and fusion”. en. In: *Language (Baltim.)* 95.1 (2019), e41–e74.

[16] Alex Warstadt and Samuel R Bowman. “Can neural networks acquire a structural bias from raw linguistic data?” In: *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.* 2020.

[17] Sam Whitman McGrath et al. “How can deep neural networks inform theory in psychological science?” en. In: *Curr. Dir. Psychol. Sci.* 33.5 (Oct. 2024), pp. 325–333.

[18] Eva Portelance and Masoud Jasbi. “The roles of neural networks in language acquisition”. en. In: *Lang. Linguist. Compass* 18.6 (Nov. 2024).

[19] Raphaël Millière and Cameron Buckner. “A philosophical introduction to language models - part II: The way forward”. In: *arXiv [cs.CL]* (May 2024).

[20] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. “Assessing the ability of LSTMs to learn syntax-sensitive dependencies”. In: *arXiv [cs.CL]* (Nov. 2016).

[21] Kristina Gulordava et al. “Colorless green recurrent networks dream hierarchically”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1195–1205.

[22] Rebecca Marvin and Tal Linzen. “Targeted syntactic evaluation of language models”. In: *arXiv [cs.CL]* (Aug. 2018).

[23] Jennifer Hu et al. “A systematic assessment of syntactic generalization in neural language models”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 1725–1744.

[24] Ethan Wilcox et al. “What do RNN language models learn about filler–gap dependencies?” In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Tal Linzen, Grzegorz Chrupała, and Afra Alishahi. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 211–221.

[25] Ethan Gotlieb Wilcox et al. “Testing the predictions of Surprisal Theory in 11 languages”. In: *arXiv [cs.CL]* (July 2023).

[26] Tom M Mitchell. *The Need for Biases in Learning Generalizations*. Tech. rep. CBM-TR-117. Rutgers University, 1980.

[27] James K Baker. “Stochastic modeling as a means of automatic speech recognition”. PhD thesis. Pittsburg, PA: Carnegie Melon University, 1975.

[28] A Newell and H Simon. “The logic theory machine–A complex information processing system”. en. In: *IEEE Trans. Inf. Theory* 2.3 (Sept. 1956), pp. 61–79.

[29] Guillermo Valle-Pérez, Chico Q Camargo, and Ard A Louis. “Deep learning generalizes because the parameter-function map is biased towards simple functions”. In: *arXiv [stat.ML]* (May 2018).

[30] Mikhail Belkin et al. “Reconciling modern machine-learning practice and the classical bias-variance trade-off”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 116.32 (Aug. 2019), pp. 15849–15854.

[31] Chiyuan Zhang et al. “Understanding deep learning (still) requires rethinking generalization”. en. In: *Commun. ACM* 64.3 (Mar. 2021), pp. 107–115.

[32] Tom Henighan et al. *Superposition, Memorization, and Double Descent*. en. Tech. rep. 6:24. Anthropic, 2023.

[33] Idan Attias et al. “Information complexity of stochastic convex optimization: Applications to generalization and memorization”. In: *arXiv [cs.LG]* (Feb. 2024).

[34] Alexander Maloney, Daniel A Roberts, and James Sully. “A solvable model of neural scaling laws”. In: *arXiv [cs.LG]* (Oct. 2022).

[35] Anirudh Goyal and Yoshua Bengio. “Inductive biases for deep learning of higher-level cognition”. en. In: *Proc. Math. Phys. Eng. Sci.* 478.2266 (Oct. 2022).

[36] Noam Chomsky. *The minimalist program*. Cambridge, Massachusetts: MIT Press, 1995, p. 420.

[37] Adele Goldberg and Laura Suttle. “Construction grammar”. en. In: *Wiley Interdiscip. Rev. Cogn. Sci.* 1.4 (July 2010), pp. 468–477.

[38] Maryellen C Macdonald. “How language production shapes language form and comprehension”. en. In: *Front. Psychol.* 4 (Apr. 2013), p. 226.

[39] Anne S Hsu, Nick Chater, and Paul M B Vitányi. “The probabilistic analysis of language acquisition: theoretical, computational, and experimental analysis”. en. In: *Cognition* 120.3 (Sept. 2011), pp. 380–390.

[40] Amy Perfors, Joshua B Tenenbaum, and Terry Regier. “The learnability of abstract syntactic principles”. en. In: *Cognition* 118.3 (Mar. 2011), pp. 306–338.

[41] Ezer Rasin et al. “Approaching explanatory adequacy in phonology using Minimum Description Length”. In: *J. Lang. Model.* 9.1 (Oct. 2021), pp. 17–66.

[42] Elena Voita and Ivan Titov. “Information-theoretic probing with minimum description length”. In: *arXiv [cs.CL]* (Mar. 2020).

[43] Minyoung Huh et al. “The platonic representation hypothesis”. In: *arXiv [cs.LG]* (May 2024).

[44] Atticus Geiger et al. “Causal abstraction: A theoretical foundation for mechanistic interpretability”. In: *arXiv [cs.AI]* (Jan. 2023).

[45] Shauli Ravfogel et al. “Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction”. In: *arXiv [cs.CL]* (May 2021).

[46] Catherine Arnett et al. “On the acquisition of shared grammatical representations in bilingual language models”. In: *arXiv [cs.CL]* (Mar. 2025).

[47] James Michaelov et al. “Structural priming demonstrates abstract grammatical representations in multilingual language models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, Dec. 2023, pp. 3703–3720.

[48] Richard Futrell et al. “Neural language models as psycholinguistic subjects: Representations of syntactic state”. In: *Proceedings of the 2019 Conference of the North*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Stroudsburg, PA, USA: Association for Computational Linguistics, June 2019, pp. 32–42.

[49] Alex Warstadt et al. “BLiMP: The Benchmark of Linguistic Minimal Pairs for English”. en. In: *Trans. Assoc. Comput. Linguist.* 8 (Dec. 2020). Ed. by Mark Johnson, Brian Roark, and Ani Nenkova, pp. 377–392.

[50] Jaap Jumelet et al. “Language models use monotonicity to assess NPI licensing”. In: *arXiv [cs.CL]* (May 2021).

[51] Kabir Ahuja et al. “Learning syntax without planting trees: Understanding hierarchical generalization in transformers”. In: *arXiv [cs.CL]* (Apr. 2024).

[52] Abhinav Patil et al. “Filtered Corpus Training (FiCT) shows that language models can generalize from indirect evidence”. In: *arXiv [cs.CL]* (May 2024).

[53] Kanishka Misra and Kyle Mahowald. “Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs”. In: *arXiv [cs.CL]* (Mar. 2024).

[54] Steven Y Feng, Noah D Goodman, and Michael C Frank. “Is child-directed speech effective training data for language models?” In: *arXiv [cs.CL]* (Aug. 2024).

[55] Qing Yao et al. “Both direct and indirect evidence contribute to dative alternation preferences in language models”. In: *arXiv [cs.CL]* (Mar. 2025).

[56] Noam Chomsky, Ian Roberts, and Jeffrey Watumull. “Noam Chomsky: The False Promise of ChatGPT”. en. In: *The New York Times* (Mar. 2023).

[57] Julie Kallini et al. “Mission: Impossible Language Models”. In: *arXiv [cs.CL]* (Jan. 2024).

[58] Michael Hahn and Mark Rofin. “Why are Sensitive Functions Hard for Transformers?” In: *arXiv [cs.LG]* (Feb. 2024).

[59] William Merrill, Ashish Sabharwal, and Noah A Smith. “Saturated Transformers are Constant-Depth Threshold Circuits”. In: *arXiv [cs.CL]* (June 2021).

[60] William Merrill, Jackson Petty, and Ashish Sabharwal. “The illusion of state in state-space models”. In: *arXiv [cs.LG]* (Apr. 2024).

[61] Lena Strobl et al. “What formal languages can transformers express? A survey”. In: *arXiv [cs.LG]* (Oct. 2023).

[62] Lois Bloom. *Language Development*. en. The MIT Press, Massachusetts Institute of Technology, 1970.

[63] Lois Bloom et al. “Structure and Variation in Child Language”. In: *Monogr. Soc. Res. Child Dev.* 40.2 (May 1975), p. 1.

[64] V Valian. “Syntactic subjects in the early speech of American and Italian children”. en. In: *Cognition* 40.1-2 (Aug. 1991), pp. 21–81.

[65] P Bloom. “Subjectlees sentences in child language”. In: *Linguistic Inquiry* 21.4 (1990), pp. 491–504.

[66] Kathryn Bock. *Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation*. 1982.

[67] Nina Hyams. “The null subject parameter in language acquisition”. en. In: *The Null Subject Parameter*. Dordrecht: Springer Netherlands, 1989, pp. 215–238.

[68] Nina Hyams and Kenneth Wexler. “On the grammatical basis of null subjects in child language”. In: *Linguist. Inq.* 24.3 (1993), pp. 421–459.

[69] Shota Momma, L Robert Slevc, and Colin Phillips. “Unaccusativity in sentence production”. en. In: *Linguist. Inq.* 49.1 (Jan. 2018), pp. 181–194.

[70] H Schriefers, E Teruel, and R M Meinshausen. “Producing simple sentences: Results from picture–word interference experiments”. en. In: *J. Mem. Lang.* 39.4 (Nov. 1998), pp. 609–632.

[71] H Schriefers. “Phonological facilitation in the production of two-word utterances”. en. In: *Eur. J. Cogn. Psychol.* 11.1 (Mar. 1999), pp. 17–50.

[72] Dana McDaniel, Cecile McKee, and Merrill F Garrett. “Children’s sentence planning: syntactic correlates of fluency variations”. en. In: *J. Child Lang.* 37.1 (Jan. 2010), pp. 59–94.

[73] Nina Hyams. *Language acquisition and the theory of parameters*. en. Studies in Theoretical Psycholinguistics. Dordrecht, Netherlands: Kluwer Academic, Aug. 1986.

[74] Howard Lasnik and Terje Lohndal. “Government-binding/principles and parameters theory: Government-Binding/Principles and Parameters Theory”. en. In: *Wiley Interdiscip. Rev. Cogn. Sci.* 1.1 (Jan. 2010), pp. 40–50.

[75] Frederick J Newmeyer. “Against a parameter-setting approach to typological variation”. en. In: *Linguist. Var. Yearb.* 4 (Dec. 2004), pp. 181–234.

[76] Charles D Yang. *Knowledge and learning in natural language*. en. London, England: Oxford University Press, Feb. 2003.

[77] Charles D Yang. “Universal Grammar, statistics or both?” en. In: *Trends Cogn. Sci.* 8.10 (Oct. 2004), pp. 451–456.

[78] Maia Duguine. “Reversing the approach to null subjects: A perspective from language acquisition”. en. In: *Front. Psychol.* 8 (Feb. 2017), p. 27.

[79] Karina Bertolino. “The setting of the null subject parameters across (non-)null-subject languages”. In: *Languages* (Aug. 2024).

[80] Alex Warstadt et al. “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*. Ed. by Alex Warstadt et al. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1–34.

[81] Alex Warstadt. “Artificial neural networks as models of human language acquisition”. PhD thesis. New York University, 2022.

[82] James A Michaelov and Benjamin K Bergen. “Do language models make human-like predictions about the coreferents of Italian anaphoric zero pronouns?” In: *arXiv [cs.CL]* (Aug. 2022).

[83] Cara Su-Yi Leong and Tal Linzen. “Language models can learn exceptions to syntactic rules”. In: *arXiv [cs.CL]* (June 2023).

[84] Matthew Honnibal et al. “spaCy: Industrial-strength Natural Language Processing in Python”. In: *Zenodo* (2020). doi: 10.5281/zenodo.1212303.

[85] Charles Yang and Tom Roeper. *Minimalism and Language Acquisition*. Oxford University Press, Mar. 2011.

[86] Taku Kudo and John Richardson. “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2018, pp. 66–71.

[87] Chigusa Kurumada and T Florian Jaeger. “Communicative efficiency in language production: Optional case-marking in Japanese”. In: *J. Mem. Lang.* 83 (Aug. 2015), pp. 152–178.

[88] Paul Bloom. “Why Do Children Omit Subjects?” en. In: (Apr. 1989).

[89] Mingfang et al. “From thought to action: How a hierarchy of neural dynamics supports language production”. In: *arXiv [q-bio.NC]* (Feb. 2025).

[90] H Harley and E Ritter. “Person and number in pronouns: A feature-geometric analysis”. In: *Language* 78.3 (Sept. 2002), pp. 482–526.

[91] Maria Nella Carminati. “Processing reflexes of the Feature Hierarchy (Person > Number > Gender) and implications for linguistic theory”. en. In: *Lingua* 115.3 (Mar. 2005), pp. 259–285.

[92] M. Eulalia Bonet. “Morphology After Syntax: Pronominal Clitics in Romance”. PhD thesis. Massachusetts Institute of Technology, 1991.

[93] Alethea Power et al. “Grokking: Generalization beyond overfitting on small algorithmic datasets”. In: *arXiv [cs.LG]* (Jan. 2022).

[94] Vikrant Varma et al. “Explaining grokking through circuit efficiency”. In: *arXiv [cs.LG]* (Sept. 2023).

[95] David Roman. *Shortcut learning of large language models in natural language understanding*. en. https://cacm.acm.org/research/shortcut-learning-of-large-language-models-in-natural-language-understanding/?utm_source=chatgpt.com. Accessed: 2025-8-15. Jan. 2024.

Appendix

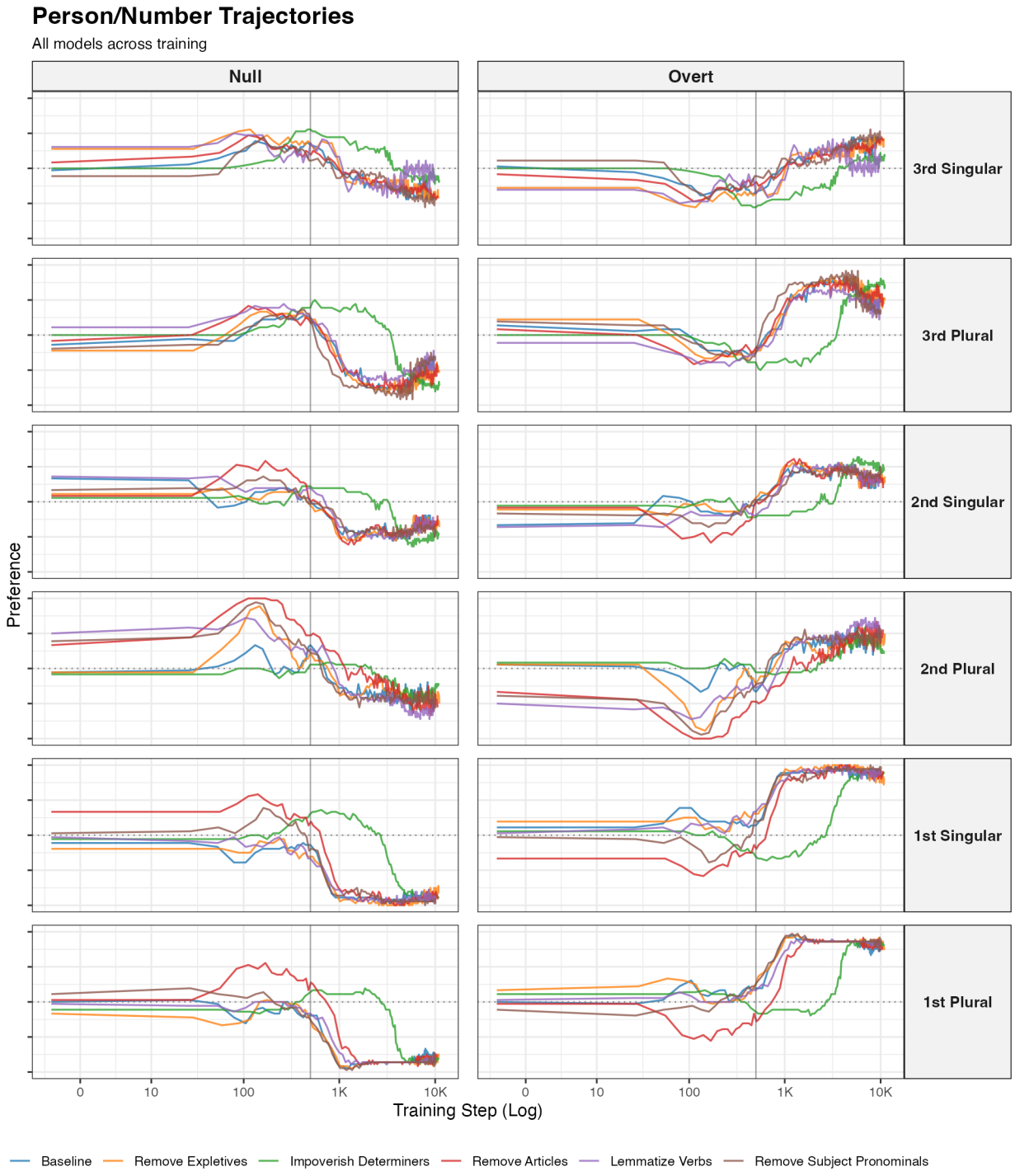

Figure 27: Learning trajectories for person/number item groups across experimental models

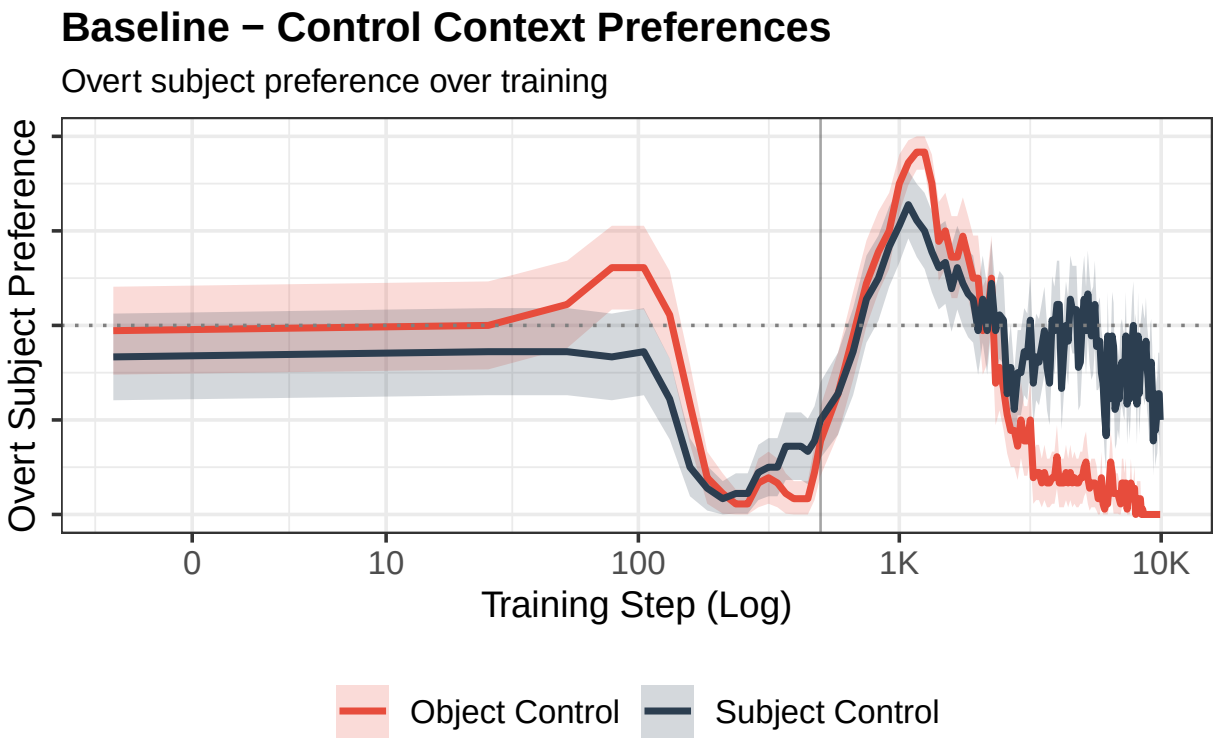

Figure 28: Learning trajectories for control constructions in the baseline Model

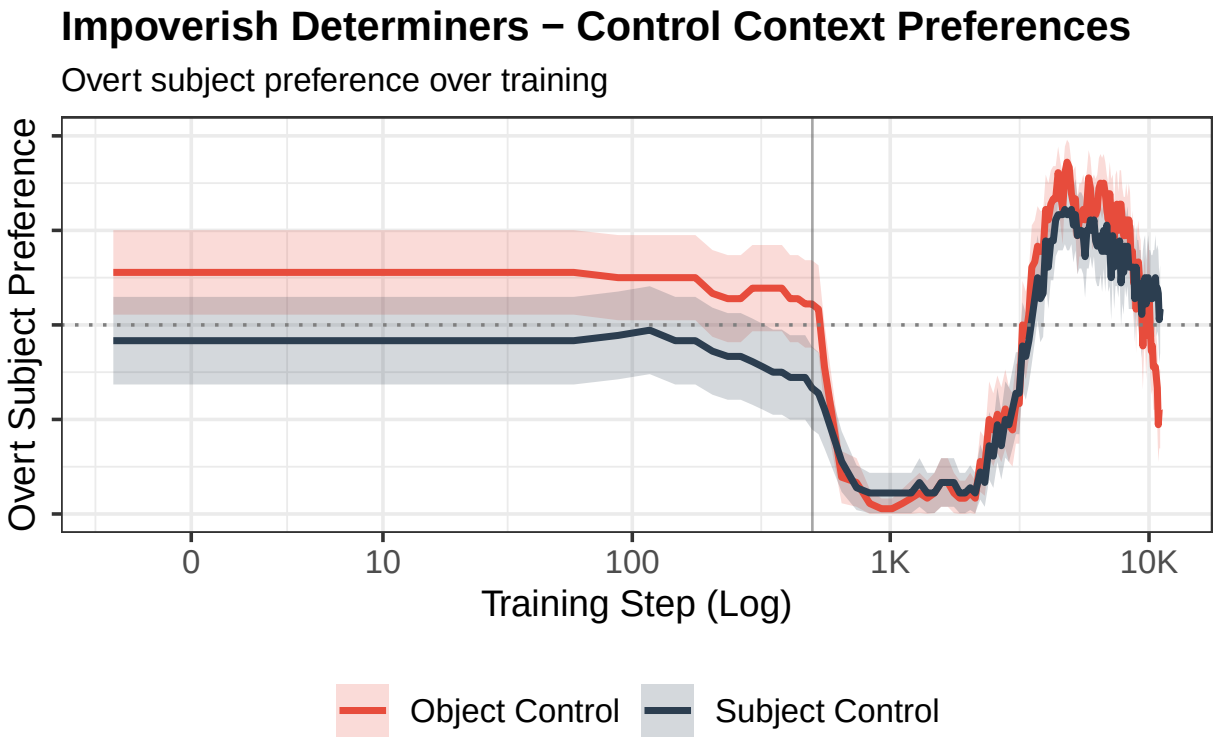

Figure 29: Learning trajectories for control constructions in the impoverish determiners modle

Stimuli creation prompt

```
### **Prompt for Generating Remaining Stimuli Sets**

**Objective**: Generate evaluation sets for the remaining Italian
syntactic phenomena, ensuring **isolation from previously generated
stimuli** to avoid bias. Each set must include:
1. **Context sentences** (Italian + English) to establish reference.
2. **Target sentences** (grammatical vs. ungrammatical minimal pairs).
3. **Glosses + translations**.
4. **Hotspots** (critical words for surprisal measurement).

---

### **Instructions**
#### **1. Target Phenomena**
Generate **12 minimal pairs** (12 grammatical, 12 ungrammatical) for each of:
- **3rd/2nd/1st person pronoun agreement** (separate sg/pl).
- **Expletive constructions** (*ø* vs. *ci*).
- **Distant antecedents in embedded clauses** (pronoun vs. *ø*).
- **Coordinate structures with topic shift** (pronoun vs. *ø*).

#### **2. Requirements**
- **Novel lexical items**: Avoid verbs/nouns used in previous sets
(*pensare*, *convincere*, etc.).
- **Natural contexts**: Must pragmatically justify the target structure.
- **Balanced design**: 50% grammatical, 50% ungrammatical per set.
- **Hotspots**: Mark finite verbs, pronouns, or auxiliaries for surprisal.

#### **3. Template**
For each pair:
```
Context: [Italian sentence]. / "[English translation]."
Target (G): [Grammatical sentence].
  "[Gloss]." → *"[Translation]."*
  Hotspot: [critical word]
Target (U): [Ungrammatical sentence].
  "[Gloss]." → *"[Translation]."*
  Hotspot: [critical word]
```

#### **4. Steps**
1. **Generate contexts** that logically precede the target sentence.
   - Example for pronoun agreement:
```

```
     *"Luca è stanco dopo il lavoro."* / *"Luca is tired after work."* →
     Targets: *Lui/ø vuole dormire*.
2. **Create minimal pairs**: Alter *only* the critical pronoun/verb.
3. **Verify ungrammaticality** with Italian syntax rules
(e.g., *ø* required in control, *ci* banned in expletives).
4. **Randomize order** of grammatical/ungrammatical items.

#### **5. Output Format**
Provide each set as a separate table (like the subject/object control sets), with:
- **Phenomenon label** (e.g., "Expletives").
- **12 numbered pairs**.
- **No overlap** with existing stimuli (check against previous lists).

---

### **Example (Expletives)**
**Phenomenon**: Expletive *ø* vs. *ci*
(GLOSS TABLE]

---

### **Final Checks**
- **No lexical overlap** with previous sets.
- **All ungrammatical versions** violate Italian syntax.
- **Hotspots** consistently marked.

Proceed iteratively: **Complete one phenomenon at a time**,
then confirm before moving to the next.
---
```

Processing manipulation prompt

```
'complex_long': 'Rewrite the CONTEXT sentence to include more descriptive,
longer noun phrases (NPs). For example, "the dog" could become
"the large brown dog with the red collar". Do not change the TARGET sentence.',
    'complex_emb': 'Rewrite the CONTEXT sentence by adding an embedded
    relative clause.
    For example, "the dog barked" could become "the dog that lived down
    the street barked".
    Do not change the TARGET sentence.',
    'target_negation': "Rewrite the TARGET sentence to be negative.
    For example, \"She thinks the ending is perfect\" becomes \"She doesn't
    think the ending is perfect\".
    Do not change the CONTEXT sentence.",
```

```
    'context_negation': "Rewrite the CONTEXT sentence to be negative.
    For example, \"Anna finished the book\" becomes
    \"Anna didn't finish the book\".
    Do not change the TARGET sentence.",
    'both_negation': 'Rewrite BOTH the CONTEXT and
    TARGET sentences to be negative.',

    You are a linguistics research assistant.
    Your task is to manipulate sentences according to
    specific instructions and extract key linguistic features.

**Instructions:**
1. You will be given a context sentence and a target sentence.
2. You will be given a manipulation instruction.
3. Apply the manipulation ONLY to the specified sentence(s) (context or target).
4. After manipulation, identify the "hotspots" in the MODIFIED target sentence.
5. Return the result as a single JSON object with the specified schema.
Do not add any extra text, explanations, or markdown formatting.

**Input Sentences:**
- Context: "{context}"
- Target: "{target}"

**Hotspots to identify in the MODIFIED target sentence:**
- subject: The subject of the main clause.
If the subject is omitted (a "subject drop"), return "Ø".
- verb: The main verb immediately following the subject.
- object: The direct or indirect object pronoun, if one exists.
Otherwise, return null.
- spillover1: The first word immediately following the main verb.
Return null if not present.
- spillover2: The second word immediately following the main verb.
Return null if not present.


Return only a JSON object with this structure:
{{
    "manipulated_context": "the manipulated context sentence",
    "manipulated_target": "the manipulated target sentence",
    "hotspots": {{
        "subject": "subject or Ø",
        "verb": "main verb",
        "object": "object or null",
        "spillover1": "first spillover or null",
```

```
        "spillover2": "second spillover or null"
    }}
}}
```