

**The Development of Abstract Syntax in Large Language Models**

Thomas G. Morton

University of California, San Diego

### The Development of Abstract Syntax in Large Language Models

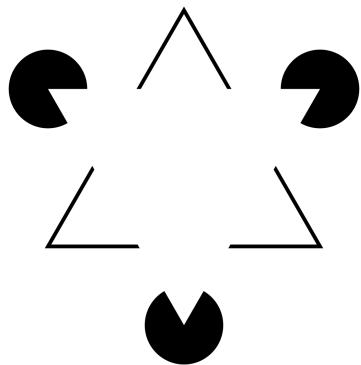
When humans produce language, the primary task is to translate one's intentions into a structured ordering of words. For any one idea we intend to communicate with language, we can express it in many different ways. For instance, describing a scene that includes a transitive action, we can choose to describe it in active voice or passive voice: 'the sailor kicked the clown,' or 'the clown was kicked by a sailor.' The choices we make when describing that scene are influenced by our own preferences to these alterations: people more frequently produce active sentences than passive ones. However, our recent experience also influences the choices that we make describing that scene: if we had just heard someone else describe a wholly different scene using passive voice, we are more likely then to use passive voice in describing our scene; or, if we had previously described a different scene with passive voice we tend towards using it again. This effect is called structural priming and it occurs even if the two scenes elicit entirely different words (Bock, 1986; Dell & Ferreira, 2016). The fact that this occurs in a lexically-independent manner suggests that what is being primed is abstract grammatical representations. Due to the breadth of work over the last four decades on structural priming, and its robust appearance across groups and languages, it has gone beyond merely a psycholinguistic effect but is regarded as a "reliable measure that permits insights into... abstract structure" and the structural representations of the mind (Dell & Ferreira, 2016, pg. 3).

Recent work suggests that large language models (LLMs) show effects of structural priming, suggesting that these models could be forming abstract representations of grammar (Sinclair et al., 2022). This proposal puts forward three studies that seek to use structural priming as a measure of the internal syntactic representations of large language models over the course of training examining their performance and competence to ask questions about learnability in language. Specifically I propose studies investigating large language models' representation of unspoken complementizers in English, unspoken subject pronouns in Italian, and a playground study investigating structural priming as a learning mechanism.

Human perception relies upon our ability to make inferences about perceptual reality even when our sensory experience provides sparse evidence. At times we are gluttonous in our perception of things that aren't there: inadvertently *perceiving*, that is forming an internal representation, of something that exists in the negative. For example, in the case of a Kanizsa contour illusion as in Figure 1, we can't help but form positive representations of occluded and incomplete triangles and circles (Kanizsa, 1987).

**Figure 1**

*Kanizsa triangle contour illusion*



Our representation of something that is simply not there may not be relevant only to visual perception. Many mainstream generative theories of syntax propose that there must be some phonologically empty, unspoken units of grammar that are functionally important for composing grammatical structures (Chomsky, 1995; Halle & Marantz, 1994). Two examples of these null elements can be found in the English null complementizer ('that') or in dropping of subject pronouns in romance languages like Italian. In both of these cases, in the absence of either the complementizer or pronoun, it is suggested that people's grammars contain a phonologically empty, but syntactically compositional ' $\emptyset$ ' item that maintains grammatical relationships. A syntactic derivation of the following examples can be found in Figure 2:

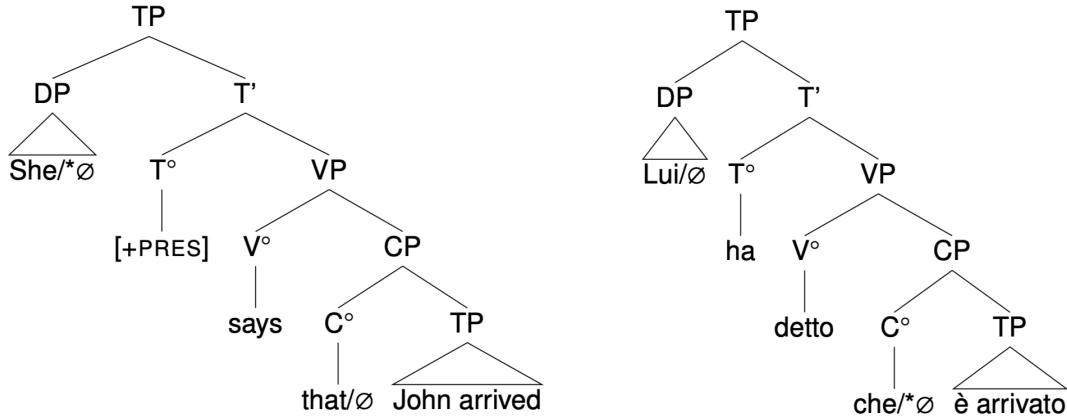
- (1) a. She/\* $\emptyset$  says that/ $\emptyset$  John arrived

3.SG.F say.PRS CMP John arrive.PST

b.	Lui/∅	ha	detto	che/*∅	è	arrivato
	3.SG.F/3.SG	3.SG.have.PRS	say	CMP	3.SG.BE.PRS	arrived.PTCP.M.SG
'she/(he/she) said that he arrived'						

**Figure 2**

Syntactic trees of ‘says that’ sentences in English and Italian



There are, of course, counter arguments against the presence of these null items in our grammatical representations. Some theories are closer to a ‘What You See is What You Get’ grammar, proposing that minimal, observable words are alone represented. These theories include constructionist approaches (Goldberg & Suttle, 2010) and modern dependency grammar theories (de Marneffe & Nivre, 2019).

Recent psycholinguistic work on structural priming has provided evidence for the psychological reality of these null elements. Momma et. al. (2025) asked participants to produce target sentences containing an overt and null ‘that’ given different prime conditions. They show that speakers are more likely to produce a null ‘that’ when primed with a construction that also contains a null ‘that.’ These results occur within and across syntactic structures, that is, null complementizers from embedded and relative clauses equally prime the use of a null complementizer in an embedded clause. This effect does not occur with the presence or

absence of a demonstrative ‘that’ like ‘that dog,’ indicating that this priming is sensitive to functional categories. Further, this priming occurs in contexts where theories of generative syntax predict the presence of a null complementizer even if an overt complementizer cannot surface in that position: those marked by ‘whether’ or ‘who’ show positive priming effects for the production of a null ‘that,’ but not those marked by ‘if.’ This suggests the speakers positively represent null complementizers specific to their functional position and in positions where overt complementizers otherwise would not surface. In this case, structural priming is used as a measure to investigate the psychological reality of speakers’ grammatical representations.

The existence of these unspoken grammatical units is important to questions about how humans are able to learn language. Specifically, accounts of the acquisition of these items suggest that these null items are core to questions about the poverty of stimulus (Hyams & Wexler, 1993; Rizzi, 1994; C. D. Yang, 2004; C. Yang & Roeper, 2011). Specifically, the necessary evidence for children to acquire these representations would require negative evidence (the absence of a thing), something that children do not utilize during language learning. This learnability problem seemed problematic for many distributional accounts at the time; however, large language models and their overwhelming success at learning language without supervision has offered the most clear and obvious opponent to generative approaches to learnability (Piantadosi, 2024).

Transformer-style large language models have shown incredible performance on language tasks that persists in improvement, and even shows emergent abilities in areas like pragmatics and problem solving after being trained on huge amounts of data (Vaswani et al., 2017; Wei et al., 2022). Despite the problems that these models create for many learnability accounts of language, these models are trained on amounts of data far beyond what children require to learn language effectively (Frank & Goodman, 2025; Wilcox et al., 2025). And further, Large Language Models have become essential tools as models of psycholinguistic subjects

and as candidate models of language learners (Warstadt, 2022; Warstadt et al., 2023; Warstadt & Bowman, 2022).

Following from recent work investigating the linguistic representations of large language models using structural priming (Arnett et al., 2025; Michaelov et al., 2023; Sinclair et al., 2022), the first study in this dissertation investigates the acquisition and representation of null complementizers in large language models.

**Chapter 1: Did you learn Nothing at school today? The Representation and Acquisition of Null Complementizers**

Children's learning of abstract structure has been richly explored in the areas of structural priming in adults (Dell & Ferreira, 2016) and in children (Kumarage et al., 2024). However, as researchers, we are limited in knowing the input available to children during learning, nor can we ethically introduce any long-term interventions into their learning to provide further insight. These problems are not present researching large language models. This study proposes investigating the development of abstract structure in large language models over time by training models on developmentally-plausible corpora and frequently sampling the structural priming effect for targeted constructions. The results of the human behavioral experiments and the evaluation stimuli can be found in Table 1 and Examples (1)-(8) in the Appendix. I will measure structural priming for a construction using Sinclair et. al.'s (2022) method. To measure the structural priming effect for a construction, for example a passive construction, I collect a passive target stimulus ( $t_p$ ) and an active and passive prime stimulus ( $p_a$  and  $p_p$ , respectively) and I create two strings, prepending the active and passive prime to the passive target, making  $[ (t_p | p_a), (t_p | p_p) ]$ , then I run the large language model over those strings, collecting the calculated log probability of each word. Then, I mask the probabilities from the prime and calculate the cumulative probability of the target given the prime, this gives me two measures,  $-\text{logProb}(t_p | p_a)$  and  $-\text{logProb}(t_p | p_p)$ , the probability of the target given each prime. Then to calculate the prime effect I find the difference,  $-\text{logProb}(t_p | p_p) - -\text{logProb}(t_p | p_a)$ .

), which returns a positive or negative number corresponding to how positively or negative the prime influenced the model's probability of the target. This is done for each target stimulus (25 of each kind) in the null complementizer evaluation set. While both a positive and negative priming effect are indicative of representational strength, we will not consider priming effects in the wrong direction as being human-like.

Our primary research questions are as follows:

- **RQ1:** Are developmentally-plausible large language models capable of acquiring adult-like representations of unspoken complementizers?
- **RQ2:** What role does model architecture have on the learning of these structures?
- **RQ3:** Can interventions on training text corpora introduce perceptual evidence that is helpful in learning these structures?

In order to address these questions I will systematically produce models by pre-training them (that is, starting from zero) using the BabyLM corpus (Choshen et al., 2024). The BabyLM corpus is a developmentally-plausible collection of diverse corpora containing a mixture of, among others, child-directed speech, dialogue, and literary/informative text. The total size of the corpus is approximately 100M words. Children encounter roughly between 7 million to 14 million words a year, making the corpus an analogical equivalent of a child's input at approximately 10 years old. No work has been performed to date demonstrating structural priming on models of this small size across pre-training.

While transformer models are the most popular commercial models in use today, many different types of neural networks have been used in psycholinguistic research and machine learning research. To this end, I intend to investigate different model architectures, so that in the case of close comparison I can compare them as candidate models of language learning. In addition to testing a basic GPT-style transformer, I will test RNN, LSTM, and simple 5-gram models to determine how much architecture alone contributes to the ability to form abstract generalizations. In addition to this, within the transformers family I will be testing StructFormer

models and Mixture of Expert models. The former is tested to consider the role that explicit dependency relationships and hierarchical relationships play in developing these structures while the latter is evaluated to determine how models trained to compartmentalize and modularize transformational functions form representations. Each of these models will be tested across multiple random seeds to determine the role of random initialization of weights and the presentation of data (which is presented to the model in random order during training). They will also be tested across multiple sizes to determine whether models with many or few parameters show differential behavior. Finally, models will be tested across different learning rates, to determine whether fast or slow traversal through the learning landscape supports or harms structural development. I believe that this manipulation alone is a worthwhile contribution, as no work has been done performing comparisons of structural priming on models across different architectures and sizes over time. Using structural priming as a probe for structural learning without having to assess the actual performance of the models themselves allows for a rich profile of the processes of learning in these models. I seek to find models that learn successfully where humans learn successfully and fail to learn where humans fail to learn as well, the goal is not to find a model that hits above or below these goal-posts.

Before, I made the argument that, to some extent, representing null complementizers requires forming a positive representation of something that is not perceptually present. This is not something that transformers, necessarily, are capable of doing: their ‘perceptual’ experience is dependent upon the ordering of data as presented in training and the way that data has been tokenized (the way words have been turned into operable vectors in the model). I propose performing interventions on the corpus data to determine how modifying the perceptual experience of the model impacts the model’s ability to learn. In this case, I am proposing using annotation large language models to insert a perceptually realized null character ‘[0]’ in positions where a null complementizer could be inferred to exist. There are three methods that I propose for performing this task. The first of these is a counterfactual intervention: an annotation

model inserts a '[0]' in a position where it thinks an overt complementizer *could* be, but otherwise isn't. The second is a lossy-context intervention. This treats the annotator like a comprehender with limited memory, where the perception of an unspoken complementizer is the result of the meeting of lossy-memory and an *a priori* satiation of lossy-context. Essentially, the training data is cut up into process-able chunks, and as the annotator works linearly, word-by-word, through those chunks; a random noise factor is introduced on the previous context biased against function words and towards content words, meaning that one is more likely to forget function words than content words. As context is erased, an auto-regressive model runs through the noisy context and seeks to satiate the loss with its own *a priori* distribution, in this case you may see something like hallucination, where a model predicts a 'that' being forgotten that otherwise was not there. If a 'that' is inserted as a result of this lossy process that was not there before, it is finally instantiated into the final dataset as a '[0]'. The final intervention is to run an unsupervised syntactic parser on the training set. After parsing the training set, a number of algorithmic rules are run within the context of the parsed data, inserting null complementizers in locations that explicitly should be found due to the syntactic, not lexical, context. After this is run the parsed structure is flattened and all that remains are the '[0]' items which were derived from these syntactic guidelines. The transformer models will then be pre-trained on these annotated data-sets and their performance across annotated and non-annotated models will be compared.

This work seeks to explore how large language models are able to form abstract representations of concepts that are not directly observable from the environment. The hope is that this direction of research will help to build a foundation of observing aspects of linguistic knowledge in models that may not be easily observable in a model's behavior. The next study builds on robust work in children's acquisition of the parameters required to drop subjects in their language, with large language models as candidate monolingual and multilingual learners.

## **Chapter 2: Will you ever learn to just drop the subject? The case of Italian Pro-Drop**

While there is little to no research on how children learn to represent null complementizers, there has been extensive research on the problem of how children acquire null subjects in languages that allow them (Hyams & Wexler, 1993; Rizzi, 1994; C. D. Yang, 2004; C. Yang & Roeper, 2011). This topic became a linchpin during debates on the poverty of stimulus, with children's data supporting accounts of parameterization in language learning. Arguments of this nature propose that as children learn language they are using positive evidence to bias themselves against default parameterizations of linguistic features (Hyams & Wexler, 1993). Evidence for this comes from the fact that children often omit subjects and objects in their early speech even if their language does not allow for it. However, there is much dispute about whether this is an effect of performance or competence (C. D. Yang, 2004). One theory is that children know that they should produce subjects, but that processing constraints restrict their ability to produce such constructions (Lutken et al., 2020; Valian & Aubry, 2005). For instance, work has shown that children who speak English that drop their subjects show reduced rates of subject dropping when asked to do a sentence repetition task multiple times (Valian & Aubry, 2005). In this study I propose two experiments designed to test large language models' learned expectations and representations of null subjects in Italian across several different developmental milestones. In the first experiment, I train a large language model on an Italian corpus of a developmentally appropriate size. In the second experiment, I build sequential bilingual models trained in either order on Italian and English of different amounts to investigate how L1 evidence facilitates or inhibits L2 learning.

### **Experiment 1**

Our research questions are as follows:

- **RQ1:** Can developmentally-plausible large language models form expectations about null subjects in Italian?

- **RQ2:** How much evidence is required for large language models to form expectations about null subjects in Italian?
- **RQ3:** Do large language models acquire expectations about null subjects that are constrained by syntactic and discourse factors at the similar times?

The training corpus is constructed to contain as diverse as possible a dataset, containing input from the Italian CHILDES corpora (MacWhinney, 1995), the Italian School-Age Children Corpus (ISAAC; (Brunato & Dell'Orletta, 2016), and the Italian L1 Learners Essay Corpus (ILECI (Barbagli et al., 2016); the bulk of the corpus consists of a diverse collection of clean monolingual text data acting as adult indirect speech and formal instruction, these are the Leipzig Italian Public Web Corpus (Goldhahn et al., 2012), the Parallel Italian European Parliament Corpus (Koehn, 2005), the Parallel Corpus of Complex-Simple Sentences for ITalian (PaCCSS-IT; (Brunato et al., 2016), and, finally, Italian texts from the Standardized Project Gutenberg Corpus (Gerlach & Font-Clos, 2018). Specific datasets, word counts, and equivalent ages can be found in Table 2 in the appendix. The adult indirect speech corpora will decrease as the corpus gets smaller, but the smaller, more child-plausible corpora will stay the same size to simulate more complex input being processed later in development (and to maintain higher proportions of the child-directed speech in smaller training contexts).

The strict training size consists of 10 million words and is roughly equivalent to the BabyLM corpus of the same size. This represents a young enough learner that would likely struggle in subject drop contexts that adults would not. The early training size is 25 million words and represents a child that should show semi-mature patterns of null subject use. The 50 million word dataset represents a late learner, one that should have received necessary exposure to show mature patterns of pronominal use. Finally, the 100 million corpus represents a mature learner and offers a size equivalent to the BabyLM dataset. 22.4 million words are held-out as an evaluation set for training.

Evaluation stimuli will be created to capture the model's expectations of the presence or absence of a subject across the Italian verbal agreement paradigm. The acquisition of null subjects and the verbal expression of gender are fairly close, as it is theorized that the additional syntactic information present on the verb is what allows for the dropping of the subject (as is the case in many languages that allow for broad subject dropping) (C. D. Yang, 2004). In each of these stimuli, a null or overt subject is followed by a verb. Surprisal of the verb will be measured to represent the model's expectations in the presence or absence of an overt subject. See Examples (9)-(11) in the appendix. Additionally, further evaluation stimuli are designed to capture the model's expectations and preferences of null pronouns in syntactically bound and discourse bound contexts. These are lower-frequency contrasts that will be used to determine whether large language models have acquired more subtle distinctions within syntactically-bound and discourse-bound contexts. See Examples (12)-(15) in the Appendix.

Strict, Early, Late, and Mature learner models will be pre-trained for a single epoch on the respectively sized BebèLM corpus. During pre-training regular checkpoints will be saved and evaluated on the held-out test set and the Italian evaluation stimuli. At the end of the first epoch, each model will be evaluated for their final performance on the evaluation stimuli. I plan to preregister that if a model is unable to complete the tasks in a human-like way, I will run it through additional epochs until performance improves or the model fails to continue learning. This is under the assumption that, in the most likely case, models and humans learn in different ways, and even if they fail to acquire human behaviors with similar amounts of data, their ability to re-process that data may allow such behavior to emerge from limited data. Additionally, models of different random initializations will be trained to determine the role that random initialization has on learning, further the role that different learning rates have on learning will be examined.

## Experiment 2

Our research questions are as follows:

- **RQ1:** Are large language models capable of maintaining competence when choosing to produce or omit subjects when trained multilingually
- **RQ2:** Does the order of language presentation impact the learnability of subject drop?
- **RQ3:** Do large language models form cross-linguistic abstract representations of the subject status as reflected by structural priming?

In this experiment I split up the BabyLM corpus such that it includes 10, 25, 50, and 100 million word increments, to follow the Italian corpus. Models are trained on either Italian or English first, and then are presented the other language as an L2 after being fully trained on the L1. Simply, I manipulate which language is first learned as the L1, and how much of that language is presented in words before the next language is learned in sequential order. This procedure is inspired by the multilingual training procedure detailed in Arnett et. al. (2025). Specifically, models are always trained as unbalanced early-learner models (L1: 10 M / L2: 25 M), balanced models (L1: 25 M / L2: 25 M), or unbalanced late-learner models (L1: 50 M / L2: 25). During training models are evaluated on the Italian evaluation stimuli, as well as English equivalent stimuli sets that display different grammatical preference, see examples (16)-(22) in the Appendix. Evaluation performance will only be compared once the L2 has been introduced.

Research suggests that the language you learn first has an impact on your ability to learn to drop subjects later on. Specifically speakers from languages that do not drop subjects routinely persist in speaking subjects even when it is otherwise dispreferred by native speakers (Jin, 1994). I predict then, that a model learning Italian first will have more success or an earlier onset at learning English competence whereas a model that first learned English will have a harder time learning Italian competence. Further I anticipate that late-learner English models will have a harder time forming Italian competence than early-learner English models, but that this same asymmetry will not be present for Italian models.

So far, the evaluation stimuli that I have used has tested the performance of these models but does not formally examine these models' linguistic competence of these

constructions. In order to do this, I propose probing for abstract cross-linguistic representations of the ‘absence’ of a subject in multilingual models while learning their second language. This follows from recent work suggesting that models form very robust cross-linguistic priming representations, indicating that the abstract representations that are formed within large language models are not necessarily Language specific (Arnett et al., 2025; Lindsey et al., 2025; Michaelov et al., 2023). In this case, I will use the same Italian and English evaluation materials that were used for evaluation of overt model performance. Stimuli will be coded for whether they contain an overt or null subject, and for each language I will examine surprisal on the target verb (in the same way that I did before) given primes with and without a subject.

Following from Chapter 1, I compare the surprisal of the target given both primes:  $-\text{logProb}(w_{\text{verb\_null}} | w_{n-1} \dots w_{n-x\_null}) - -\text{logProb}(w_{\text{verb\_null}} | w_{n-1} \dots w_{n-x\_overt})$ . It has been shown in previous work that surprisal, which is a measure of expectation, when presented as the probe measure for structural priming captures many of the same effects when targeting single words as whole sentence probabilities in large language models (Momma et al., 2025; Sinclair et al., 2022) and in humans, as observed via pupillometry (Kumarage et al., 2025). The goal then, is to turn a measure of behavioral performance, as in surprisal, into a measure of linguistic competence by manipulating the presence of a prime that shares nothing in common with the target either in words or in origin language. I do not perform this evaluation in the monolingual example, as it seems impossible to disentangle the priming of a particular pronoun for a particular pronoun, or likewise its absence (c.f. an abstract structure), however a cross-linguistic comparison does not share this same confound. I hope to find that the presence of a subject positively primes the preference for a subject across languages and that likewise the absence of a subject positively primes the absence of a subject across languages. These priming representations may appear despite the fact that the model otherwise doesn’t demonstrate adequate performance in the evaluation tasks, in which case priming allows for the probing of representations that may not be directly observable from behavior. If the presence of such structural priming is found, the

formation of an abstract representation of an absent subject could support the learning of either structure, and serve potentially as a mechanism of learning without requiring the use of negative evidence (Momma et al., 2025).

It may seem somewhat of a departure from our previous discussion to frame syntactic priming in terms of measures like surprisal, and further as a mechanism of language learning; however, there is a model of language production that proposes a connection between surprisal and structural priming as a central mechanism of language learning. The following chapter will examine this relationship by using large language models as candidate learners in a theory of language learning that connects comprehension, production, and language acquisition using surprisal and syntactic priming.

### **Chapter 3: Where did you learn to talk like that? Playground Learning and the P-Chain**

#### **with a Multi-Agent Conversational Exercise**

In this proposal so far, we have connected the ideas of language learning and structural priming, one in terms of acquiring linguistic representation and the other in terms of probing for it. What role does structural priming itself have in language learning? A series of seminal works in the area of language production suggest that structural priming itself is the mechanism by which linguistic structures are learned. Dell & Chang, (2014) and the literature that preceded it (Chang, 2002; Chang et al., 2006, a.o.) proposes a model of language learning, the P-Chain, that not only seeks to use structural priming as a primary mechanism for learning, but it also suggests that the mechanism that allows comprehenders to predict is the same one that allows them to produce (as large language models do). The P-Chain is essentially this: when someone is talking to us, we make predictions about what they are going to say next; the way that we make those predictions is by engaging our production system (the same one responsible for choosing the next word that we say). If our prediction is wrong, we update our system of production to be less wrong for future predictions. This update is learning, and it is also priming.

Prediction is production, which makes error, error leads to priming which is implicit learning. Let me explain this in two examples.

First, structural priming undergoes a phenomenon called the inverse frequency effect. When speakers are primed with the less frequent structure, they experience a stronger priming effect than if they were primed with the more frequent structure. This can be explained in terms of surprisal, if surprisal is a measure of prediction error. Structural priming can occur from language comprehension, in this case imagine a participant is hearing a sentence spoken out loud. While listening, the participant is predicting the next word in sequence, encountering the less-frequent construction (e.g. a passive); the participant experiences increased surprisal because the construction was less likely. The participant updates their expectations, which, under a P-Chain model, also updates their own language distribution, making them more likely to produce a passive in the short-term and the long-term. When prompted to describe a scene, that change in their language distribution, in the short-term, displays as a priming effect.

Now, let's discuss a more grounded example I call 'playground learning'. This can be thought of as one way of understanding how linguistic convergence emerged from a context like Nicaraguan Sign Language. In this scenario you have children who have all learned aspects of a language from signers with sparse, incorrect representations of whatever language they wanted to teach. Children go out into the playground and in interacting with each other satiate this sparse input by simply talking to each other. We can think of it like this. As the children gather in the playground, one child signs to the other children. The other children initially share very little language input with this child. So, they make predictions given their own *a priori* linguistic distribution and when they're wrong they experience strong error signals which they use to change their expectations, and in a way are primed by that first child. The next child does the same: signs towards the rest, they all predict, and likewise are primed. If this continues in a circle, what eventually happens is that these children should all align with each other in expectation and production. Eventually the children all speak the same 'language': a new

language as credit to their uneducated teachers. Now, it might be reasonable to think that such a process would just lead children to catastrophically interfere with each other, that production from insufficient sources should lead to learning that is insufficient and that all the children would not necessarily emerge with a more complete language from this process. In this study I propose testing the efficacy of the P-Chain for learning grammar in a multi-LLM ‘playground learning’ study investigating how large language models learn from each other when restricted to modeling each other’s outputs. Simply, I pre-train models on separate, or otherwise incomplete data sources; then, I ask the models to freely produce speech to each other given prompts. Each model ‘produces’ to the group and the other models are tasked to perform next-token prediction for the model whose turn it is to produce; at the end of production the models learn from that error and the next model who’s turn it is to produce begins. This proceeds in turns and I measure each model’s alignment to the other’s model both in knowledge of each other’s learned data but also in terms of alignment of their syntactic representations.

This study will consist of two experiments: one evaluating the P-Chain in practice and the other evaluating the P-Chain in principle. For Experiment 1, each model will receive a similarly sized english text corpus containing different content; for Experiment 2, each model will receive an incomplete portion of a complete context-free Dyck Language (Schützenberger, 1963) and are assessed on their acquisition of the complete Language from each other’s inputs. Essentially, I wish to test the models in a naturalistic setting to assess the capability of the P-Chain in those circumstances, and I wish to test the models on formal Language to assess the P-Chain’s feasibility as a mechanism of Language learning more broadly. I chose a Dyck Language as it has been theoretically and experimentally proven that transformer models with four layers are capable of learning these Languages to completion (Yao et al., 2021).

## Experiment 1

Our research questions are as follows:

- **RQ1:** Are models capable of learning out-of-distribution data from another model in the P-Chain procedure?
- **RQ2:** Do models that converse with each other experience change towards more similar structural representations over time?

Four transformer models will be trained on 20 million word selections of the WikiText 103 million word corpus. A final 20 million word selection will be held-out as an evaluation data-set. After pre-training of the models, the conversational fine-tuning procedure will be performed in rounds. The producer model and the listener models are decided in round-robin fashion. The producer is tasked to generate from an empty token, essentially producing freely from its underlying distribution. This continues on for 10,231 tokens, at the end of the model's context length, the context is wiped and the model continues producing from the last word of the previous context — the number is calculated as the context length of 1024 times 10 minus 9 (for the carry over token). The purpose of this is to allow the model to express a diverse enough sampling of its data set. The other three listener models receive the 10,231 new input tokens for fine-tuning, where they perform next-token prediction and propagate loss by gradient descent. Then, the next model, with a now-altered distribution begins the same procedure. After all models have been the producer once, a turn has passed and each model is sent for evaluation. At evaluation all models are tested on the same out-of-distribution test corpus. In addition, each model is tested on a sample of every other model's data-set. So, Model 1, would measure perplexity on samples of Model 2, 3, and 4's training sets. In addition to this, after each turn every model is evaluated on traditional priming measures. The conversational procedure will proceed until the models sufficiently converge or collapse.

I will track each model's improvement on each other's training sets and the overall evaluation set. In addition, I will track the distance between different model's priming measures, in order to determine whether models are getting closer or further away in representation. It is entirely possible that I find that such a procedure is not healthy for the models, and performance

on each other's data sets and the overall data set decrease from this procedure. However, our hope is that other models are able to become more like each other simply through the act of alignment at scale. I hope to confirm that priming is a central mechanism to this process by looking not only at performance at these models but also probing their internal representations. It is possible that I do not see any improvements in the overall prediction of the datasets but I still see model alignment at the level of representation that is still evidence of learning.

## Experiment 2

Our research questions are as follows:

- **RQ1:** Is the mechanism of the P-Chain in principle capable of transferring knowledge of a formal Language.
- **RQ2:** What happens to learning when models encounter production from models trained on impossible grammars in its Language.

We are not able to test for completeness in learning natural languages because we ourselves do not know its corners or bounds. However, the foundation of linguistics as a formal study has involved the definition and formalization of what can and cannot be a learnable language. Dyck Languages (Schützenberger, 1963) are context-free grammars that largely involve (but do not have to be made up of) brackets nested and concatenated within each other using production rules. For instance,  $S \rightarrow ("[" S "]")^*$  would allow for the generation of unlimited nested '[[[..]]]' brackets. A complex example could allow for embedding and concatenation: "[({})()]"." You can make up such a string with a set of production rules like the following:

$$(2) SA \rightarrow \epsilon$$

$$S \rightarrow ( S )$$

$$S \rightarrow [ S ]$$

$$S \rightarrow \{ S \}$$

$$S \rightarrow a S$$

You can then train models with small datasets generated from portions of these production rules such that you have models with imperfect selections of the language that they have learned. For instance: Model 1:  $S \rightarrow ( S )$ ,  $S \rightarrow a S = ()((())(((())(())..)$ , Model 2:  $S \rightarrow [ S ]$ ,  $S \rightarrow \{ S \} = [[([...])]]$ , Model 3:  $S \rightarrow ( S )$ ,  $S \rightarrow \{ S \} = \{\{\{((...))\}\}\}$ , Model 4,  $S \rightarrow [ S ]$ ,  $S \rightarrow \{ S \}$ ,  $S \rightarrow a S = [\{\{\}\}\{\}]$ . The question is, whether these models when placed under a round-robin procedure as is done with models in the naturalistic context are still able to converge onto the final Language from evidence as outputted by other models. In this case, models will not be asked to produce long sequences, instead they will produce even sequences between four and twelve tokens of length, which models will then fine-tune on. Because the model will require only three tokens, and four layers, there is ample area to perform very small scale interpretability experiments on the functionality of these models ([Yao et al. 2021](#)).

### Conclusion

The three studies in this dissertation seek to look at the acquisition and representation of abstract syntactic knowledge in large language models. First, I propose examining whether and how large language models are able to acquire representations of unspoken complementizers, and whether manipulating the architecture of models, and the content of their perceptual experience can give us insight into the mechanisms of language learning underlying these processes. Then, I propose two studies examining how large language models form expectations and representations of null subject languages. The first experiment examines the performance of monolingual language models in their expectations of null subjects in syntactic and discourse constrained contexts. The second experiment examines how multilingual learners are able to acquire those expectations and whether probing for multilingual representations can offer insight into the mechanisms of learning available to these models. Finally, the third study looks at structural priming as a mechanism for language learning, examining the P-Chain procedure in naturalistic and formal contexts for language learning.

## References

- Arnett, C., Chang, T. A., Michaelov, J. A., & Bergen, B. K. (2025). On the acquisition of shared grammatical representations in bilingual language models. In *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/2503.03962>
- Barbagli, A., Lucisano, P., Dellorletta, F., Montemagni, S., & Venturi, G. (2016). CltA: an L1 Italian Learners Corpus to Study the Development of Writing Competence. In *Proceedings of 10th Edition of International Conference on Language Resources and Evaluation* (pp. 23–28).
- Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18(3), 355–387.
- Brunato, D., Cimino, A., Dellorletta, F., & Venturi, G. (2016). PaCCSS-IT: A Parallel Corpus of Complex-Simple Sentences for Automatic Text Simplification“. In *Proceedings of Conference on Empirical Methods in Natural Language Processing* (pp. 351–361).
- Brunato, D., & Dell'Orletta, F. (2016). ISACCO: a corpus for investigating spoken and written language development in Italian school-age children. *IJCoL*, 2(1), 63–76.
- Chang, F. (2002). Symbolically speaking: a connectionist model of sentence production. *Cognitive Science*, 26(5), 609–651.
- Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, 113(2), 234–272.
- Chomsky, N. (1995). *The minimalist program* (p. 420). MIT Press.
- Choshen, L., Cotterell, R., Hu, M. Y., Linzen, T., Mueller, A., Ross, C., Warstadt, A., Wilcox, E., Williams, A., & Zhuang, C. (2024). [Call for Papers] The 2nd BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. In *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/2404.06214>
- Dell, G. S., & Chang, F. (2014). The P-chain: relating sentence production and its disorders to

- comprehension and acquisition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 369(1634), 20120394.
- Dell, G. S., & Ferreira, V. S. (2016). Thirty years of structural priming: An introduction to the special issue. *Journal of Memory and Language*, 91, 1–4.
- de Marneffe, M.-C., & Nivre, J. (2019). Dependency grammar. *Annual Review of Linguistics*, 5(1), 197–218.
- Frank, M. C., & Goodman, N. D. (2025). Cognitive modeling using artificial intelligence. *PsyArXiv*. Retrieved from [Osf. io/preprints/psyarxiv/wv7mg/](https://osf.io/preprints/psyarxiv/wv7mg/) v1 Doi, 10. [https://osf.io/preprints/psyarxiv/wv7mg\\_v1](https://osf.io/preprints/psyarxiv/wv7mg_v1)
- Gerlach, M., & Font-Clos, F. (2018). A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. In *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/1812.08092>
- Goldberg, A., & Suttle, L. (2010). Construction grammar. *Wiley Interdisciplinary Reviews. Cognitive Science*, 1(4), 468–477.
- Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. *International Conference on Language Resources and Evaluation*, 759–765.
- Halle, M., & Marantz, A. (1994). Some Key Features of Distributed Morphology. *MIT Working Papers in Linguistics*, 21.
- Hyams, N., & Wexler, K. (1993). On the grammatical basis of null subjects in child language. *Linguistic Inquiry*, 24(3), 421–459.
- Jin, H. (1994). Topic-prominence and subject-prominence in L2 acquisition: Evidence of English-to-Chinese typological transfer. *Language Learning*, 44(1), 101–122.
- Kanizsa, G. (1987). Quasi-perceptual margins in homogeneously stimulated fields. In *The Perception of Illusory Contours* (pp. 40–49). Springer New York.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings*

- of Machine Translation Summit X: Papers*, 79–86.
- Kumarage, S., Donnelly, S., & Kidd, E. (2024). A meta-analysis of syntactic priming experiments in children. *Journal of Memory and Language*, 138(104532), 104532.
- Kumarage, S., Malko, A., & Kidd, E. (2025). Indexing prediction error during syntactic priming via pupillometry. *Language, Cognition and Neuroscience*, 1–21.
- Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., ... Batson, J. (2025). On the biology of a large language model. *Transformer Circuits Thread*.  
[https://scholar.google.com/citations?view\\_op=view\\_citation&hl=en&citation\\_for\\_view=5sxXSfwAAAAJ:aqlVkm33-oC](https://scholar.google.com/citations?view_op=view_citation&hl=en&citation_for_view=5sxXSfwAAAAJ:aqlVkm33-oC)
- Lutken, C. J., Legendre, G., & Omaki, A. (2020). Syntactic creativity errors in children's wh-questions. *Cognitive Science*, 44(7), e12849.
- MacWhinney, B. (1995). *The CHILDES Project: Tools for Analyzing Talk* (2nd ed.). Erlbaum.
- Michaelov, J., Arnett, C., Chang, T., & Bergen, B. (2023). Structural priming demonstrates abstract grammatical representations in multilingual language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3703–3720.
- Momma, S., Richards, N., & Ferreira, V. (2025). *Speakers encode silent structures: evidence from complementizer priming in English*. Human Sentence Processing Conference 2025, University of Maryland, College Park. <https://hsp2025.github.io/abstracts/152.pdf>
- Piantadosi, S. T. (2024). Modern language models refute Chomsky's approach to language. In *lingbuzz.net*. UC Berkeley. <https://lingbuzz.net/lingbuzz/007180>
- Rizzi, L. (1994). Early null subjects and root null subjects. In B. Lust (Ed.), *Language Acquisition Studies in Generative Grammar* (Vol. 2, p. 151). John Benjamins Publishing Company.
- Schützenberger, M. P. (1963). On context-free languages and push-down automata. *Information and Control*, 6(3), 246–264.

- Sinclair, A., Jumelet, J., Zuidema, W., & Fernández, R. (2022). Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10, 1031–1050.
- Valian, V., & Aubry, S. (2005). When opportunity knocks twice: two-year-olds' repetition of sentence subjects. *Journal of Child Language*, 32(3), 617–641.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *arXiv [cs.CL]*. arXiv.  
<http://arxiv.org/abs/1706.03762>
- Warstadt, A. (2022). *Artificial neural networks as models of human language acquisition*. New York University.
- Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In S. Lappin & J.-P. Bernardy (Eds.), *Algebraic Structures in Natural Language* (pp. 17–60). CRC Press.
- Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, & R. Cotterell (Eds.), *Proceedings of the BabyLM challenge at the 27th conference on computational natural language learning* (pp. 1–34). Association for Computational Linguistics.
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. In *arXiv [cs.CL]*. arXiv.  
<http://arxiv.org/abs/2206.07682>
- Wilcox, E. G., Hu, M. Y., Mueller, A., Warstadt, A., Choshen, L., Zhuang, C., Williams, A., Cotterell, R., & Linzen, T. (2025). Bigger is not always better: The importance of

- human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144(104650), 104650.
- Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, 8(10), 451–456.
- Yang, C., & Roeper, T. (2011). *Minimalism and Language Acquisition*. Oxford University Press.
- Yao, S., Peng, B., Papadimitriou, C., & Narasimhan, K. (2021). Self-attention networks can process bounded hierarchical languages. In *arXiv [cs.CL]*. arXiv.  
<http://arxiv.org/abs/2105.11115>

## Appendix

### Chapter 1

**Table 1**

*Priming conditions and results of Momma et. al. (2025)*

Exp.	Prime Condition 1	Prime Condition 2	Effect on Null Complementizer
1	VP-complement clause	Relative clause	Both conditions with a null complementizer primed the use of a null complementizer at similar rates.
2	Null complementizer in a relative clause	Absence of the word 'that' (demonstrative)	The null complementizer was more effective at priming another null than the mere absence of the word 'that'.
3	'Who' relative clause	Neutral baseline	Clauses with 'who' primed the use of the null complementizer.
4	'Who' relative clause	Neutral baseline	Replicated the finding that clauses with 'who' primed the null complementizer.
5	'Who' relative clause	Complex neutral prime	Clauses with 'who' primed the null complementizer, which was not due to the prime's complexity.
6	'Whether' clause	'If' clause	Clauses with 'whether' primed the null complementizer, while clauses with 'if' had no effect.

### **Evaluation Stimuli Examples**

- (3) Experiment 1: Relative Clause → Verb Clause (Momma et al., 2025):
  - a. Prime\_that: The professor appreciated the thoughts that the students expressed during class.
  - b. Prime\_null: The professor appreciated the thoughts the students expressed during the class.
  - c. Target\_that: The director announced that the actor would be in the new movie.
  - d. Target\_null: The directory announced the actors would be in the new movie.
- (4) Experiment 1: Verb Clause → Verb Clause (Momma et al., 2025):
  - a. Prime\_that: The professor thought that the students appreciated the idea during class

- b. Prime\_null: The professor thought the students appreciated the idea during class
- c. Target\_that: The director announced that the actor would be in the new movie.
- d. Target\_null: The directory announced the actors would be in the new movie.

(5) Experiment 2: Determiner → Verb Clause (Momma et al., 2025):

- a. Prime\_that: The professor appreciated the thoughts of that student.
- b. Prime\_null: The professor appreciated the thoughts of the student.
- c. Target\_that: The director announced that the actor would be in the new movie.
- d. Target\_null: The directory announced the actors would be in the new movie.

(6) Experiment 3: Who-Clause → Verb Clause (Momma et al., 2025):

- a. Prime\_that<sup>1</sup>: They appreciated the professor that was really lenient
- b. Prime\_null: They appreciated the professor who was really lenient
- c. Target\_that: The director announced that the actor would be in the new movie.
- d. Target\_null: The directory announced the actors would be in the new movie.

(7) Experiment 6: If-Clause → Verb Clause (Momma et al., 2025):

- a. Prime\_that: The scientist is unsure if the theory is accurate
- b. Prime\_cntrl<sup>2</sup>: The scientist is unsure about the accuracy of the theory.
- c. Target\_that: The director announced that the actor would be in the new movie.
- d. Target\_null: The directory announced the actors would be in the new movie.

(8) Experiment 6: Whether-Clause → Verb Clause (Momma et al., 2025):

- a. Prime\_that: The scientist is unsure if the theory is accurate
- b. Prime\_null: The scientist is unsure whether the theory is accurate.
- c. Target\_that: The director announced that the actor would be in the new movie.
- d. Target\_null: The directory announced the actors would be in the new movie.

---

<sup>1</sup> In Shota's original stimuli which he used to assess some large language model's performance on priming this data, he had the primes labeled the wrong way, so unfortunately his results for 'who' are uninterpretable given this caveat. We've fixed this for our stimuli.

<sup>2</sup> Shota's original design where he looks at these using large language models includes a control stimuli with 'about', this isn't functionally helpful to the priming stimuli and necessarily makes it that this is primarily looking at the effect of 'if' on either target vs the effect of an unrelated stimuli. We may change this in the future depending on what we're looking for.

## Chapter 2

**Table 2**

*Word Count of Corpora across learner-dataset sizes including final proportion and approximate age given dataset-size for Italian BébéLM (calculated with N\_words/7M)*

Dataset	Strict	Early	Late	Mature	Test
CHILDES (MacWhinney 2000)	589K (5.89%)	589K	589K	589K (0.59%)	
Italian School-Age Children Corpus (ISAAC)	42K (0.42%)	42K	42K	42K (0.04%)	
Italian L1 Learners Essay Corpus (ILEC)	215K (2.15%)	215K	215K	215K (0.21%)	
Leipzig Web Public Corpus	1.45M (14.67%)	3.87M	7.88M	15.89M (15.89%)	3.59M
Italian Corpus of Complex-Simple sentence pairs	80K (0.8%)	212K	432K	871K (0.87%)	197K
QCRI Educational Domain Corpus	765K (7.68%)	2.01M	4.11M	7.28(8.28%)	1.78M
Standardized Project Gutenberg Corpus	3.23M (32.29%)	8.52M	17.34M	34.97M (34.97%)	7.9M
EuroParl	3.61M (36.13%)	9.53M	19.4M	39.14M (39.24%)	8.85M
Total Words	10M	25M	50M	100M	22.4M
Approximate Age	1.4 yrs	3.5 yrs	7.1 yrs	14.3 yrs	

### ***Italian Evaluation Stimuli Examples***

(9) *3rd person singular and plural*

- a. Anna ha finito il libro. Lei/∅ pensa che il finale sia perfetto.

Anna has finished the book 3.sg.pl/3.sg thinks.3sg that the ending is perfect.

'Anna has finished the book. She thinks that the ending is perfect'

- b. I clienti hanno visto la proposta. Loro/∅ pensano che il budget sia accettabile.

The clients have seen the proposal. 3.pl think.3pl that the budget is acceptable.

'The clients have seen the proposal. They think that the budget is acceptable'

*(10) 2nd person singular and plural*

a. Marco, hai letto l'email. Tu/∅ pensi che abbiamo bisogno di più tempo.

Marco, had.2 read the.email. You think.2sg that have.1pl need of more time.

'Marco, you had read the email. 2.sg think that we need more time.'

b. Studenti, avete sentito la notizia. Voi/∅ pensate che la decisione sia giusta

Students, have.2pl heard the news. 2.pl think.2pl that the decision be.subj-3sg fair.

'Students, you heard the news. You think that the decision is fair.'

*(11) 1st person singular and plural*

a. Ho rivisto l'ordine del giorno. Lo/∅ penso che il programma sia troppo serrato

have-1sg reviewd the.order of.the day. 1.sg think.1sg that the schedule

be.subj-3sg too tight.

'I have reviewed the agenda. I think the schedule is too tight.'

b. Lo e il mio team abbiamo visto la demo. Noi/∅ pensiamo che il prodotto abbia potenziale.

Me and the my team have.1pl seen the demo. 1pl think.1pl that the product

have.subj-3sg potential

'Me and my team have seen the demo. We think that the product has potential.'

*(12) Subject and Object Control*

a. Maria ha convinto suo fratello ∅/\*lui a partire presto dalla festa.

Maria has convinced her brother 3.SG to leave early from.the party

b. Il regista ha promesso agli attori ∅/\*lui di rividere il copione.

The director has promised to.the actors 3.SG to revise the script

*(13) Expletive constructions*

∅/\*Ci sembra che gli studenti abbiano superato l'esame facilmente.

it seems that the students have.SUBJ passed the.exam easily

## (14) Distant antecedent in embedded finite clauses

Il cameriere ha detto che \*∅/lui aveva aspettato più di un'ora

The waiter has said that 3.SG had waited more of an hour

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

a. Giovanni si è svegliato tardi e ∅/lui ha perso il treno completamente.

Giovanni REFL is woken late and 3.SG has missed the train completely

b. Anna ha chiamato Marco e \*∅/lui ha rifiutato di rispondere alle sue domande.

Anna has called Marco and 3.SG has refused to answer to her questions

***English Evaluation Stimuli Examples***(16) *Third person singular and plural*

a. Anna finished the book. She/\*∅ thinks the ending is perfect.

b. The clients saw the proposal. They/\*∅ think the budget is acceptable

(17) *Second person singular and plural*

a. Marco, you read the email. You/\*∅ think we need more time.

b. Students, you heard the news. You all/\*∅ think the decision is fair.

(18) *First person singular and plural*

a. I reviewed the agenda. I/\*∅ think the schedule is too tight.

b. My team and I saw the demo. We/\*∅ think the product has potential

(19) *Subject and Object Control*

a. Maria convinced her brother ∅/\*him to leave the party early

b. The director promised the actors ∅/\*he to revise the script

(20) *Expletive constructions*

\*∅/It seems that the students passed the exam easily

(21) *Distant antecedent in embedded finite clauses*

The waiter mentioned that \*∅/he had waited over an hour

(22)Coordinate structures with and without topic shift

- a. Giovanni woke up late and  $\emptyset/\text{he}$  missed the train completely
- b. Anna called Mark and  $*\emptyset/\text{he}$  refused to answer her question