Grammar-Engineered Synthetic Languages for Studying Typological Biases in Language Models: A Case Study of Action Nominal Constructions

Iskar Deng

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2026

Committee:

Shane Steinert-Threlkeld

Gina-Anne Levow

Program Authorized to Offer Degree:
Department of Linguistics

University of Washington

Abstract

Grammar-Engineered Synthetic Languages for Studying Typological Biases in Language Models: A Case Study of Action Nominal Constructions

Iskar Deng

Chair of the Supervisory Committee:
Shane Steinert-Threlkeld
Department of Linguistics

This thesis proposes a grammar-engineered method for constructing synthetic corpora to study language models' learning preferences in acquiring action nominal constructions (ANCs) and related phenomena. The method uses the Grammar Matrix to construct Head-driven Phrase Structure Grammar (HPSG) grammars varying in typological parameters, with Minimal Recursion Semantics (MRS) as a shared semantic interface across grammars. This design makes it possible to control fine-grained grammatical parameters while preserving part of the semantic and lexical distribution of natural data. Using this method, I construct 96 parallel languages varying in clause word order, NP word order, alignment, complement system, and ANC strategy. For each language, I train GPT-2-small models from scratch and evaluate them using targeted minimal pairs covering 34 grammatical phenomena. The results show that models more easily learn systems that maintain consistent linear organization across related grammatical structures, which aligns with preferences in human language typology. At the same time, additional morphological cues can help models distinguish similar structures, but human languages do not always favor more such cues. As a result, some parameter combinations that are relatively easy for models to learn remain rare in human languages. Overall, these results highlight both the typological relevance of model learning preferences and the limits of learnability as an explanation for human language patterns.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Completing this thesis and my master's degree has been a tough but meaningful journey. I would like to thank the people who helped and supported me along the way.

First, I would like to thank my advisor, Prof. Shane Steinert-Threlkeld. Over the past two years, his guidance has helped me develop my research taste, and he has given me much support and encouragement throughout the development of this thesis.

Second, I would like to thank Prof. Gina-Anne Levow for serving as my second reader and for her careful reading and feedback on this thesis, and Prof. Emily M. Bender for her early suggestions on the direction of this thesis. I am also grateful to the other faculty members in UW Linguistics. During the program, the department has provided me with an academic environment in which I felt supported and at ease.

I would also like to thank my peers for our discussions, especially Nathalia Xu. Our collaboration on differential argument marking shaped the way I think about research and also inspired this thesis.

Finally, I would like to thank Seattle. Its natural scenery, city life, and the Mariners made my graduate school experience feel more grounded and warm, and gave me many memories that I will carry with me for a long time.

Chapter 1
INTRODUCTION

Recent language models (LMs) have shown substantial ability to acquire linguistic structure, including aspects of morphology, syntactic dependencies, and long-distance agreement, as demonstrated by targeted evaluations using grammatical minimal pairs [Linzen et al., 2016, Gulordava et al., 2018, Marvin and Linzen, 2018, Warstadt et al., 2020]. These findings suggest that LMs can acquire nontrivial structural generalizations from next-word prediction alone.

Therefore, a natural next question is not simply whether LMs can learn grammar, but which kinds of grammars are easier for them to learn. If several logically possible grammatical systems are all learnable, yet some are learned faster, yield lower perplexity, or support more robust generalization, then these differences can reveal the model's inductive biases [Ravfogel et al., 2019, White and Cotterell, 2021, Kallini et al., 2024, Xu et al., 2026].

This question matters because human languages occupy only a small and highly structured region of the space of possible grammars [Greenberg, 1963, Dryer and Haspelmath, 2013]. If model preferences align with human typology, this would suggest that some typological tendencies may not require strongly language-specific constraints, but can emerge partly from more general pressures such as prediction, memory, processing efficiency, or information locality [Hawkins, 2004, Kemp et al., 2018, Gibson et al., 2019, Hahn et al., 2020, Steinert-Threlkeld and Szymanik, 2020, Kuribayashi et al., 2024, Imel et al., 2026]. If they diverge from human typology, such divergence can also help identify where human language may depend on pragmatic, communicative, discourse-based, or diachronic pressures not captured by the model [White and Cotterell, 2021, Futrell and Mahowald, 2026, Deng et al., 2026].

Synthetic language experiments make it possible to frame these typological patterns as learnability questions, asking whether typologically rare or unattested grammatical systems are harder for LMs to learn. Such comparisons require synthetic languages that vary in their grammatical systems while holding semantic content, training scale, lexical statistics, and evaluation conditions approximately constant [Ravfogel et al., 2019, White and Cotterell, 2021, Kallini et al., 2024, Kuribayashi et al., 2024, Xu et al., 2026, Deng et al., 2026].

Existing work using synthetic languages can be broadly divided into two approaches. One approach starts from natural-language corpora and constructs counterfactual languages by systematically perturbing grammatical properties such as word order, case marking, or agreement [Wang and Eisner, 2016, Ravfogel et al., 2019, Kallini et al., 2024, Xu et al., 2026, Deng et al., 2026]. One advantage of this approach is that it preserves lexical, semantic, and distributional properties of the source corpus, making the input conditions relatively naturalistic. However, it is typically limited to modifying structures already present in the source data, making it difficult to construct globally coherent and typologically controlled grammatical systems.

A second approach generates corpora directly from artificial grammars, which makes it possible to create clean minimal contrasts and explore rare or unattested grammatical systems [White and Cotterell, 2021, Kuribayashi et al., 2024]. However, fully artificial corpora often lack the rich lexical, semantic, and structural distributions found in natural language. These distributions are crucial for LMs to learn grammatical phenomena such as argument structure.

In this thesis, I propose an approach based on grammar engineering that combines the strengths of these two approaches. Using the Grammar Matrix, a grammar engineering framework based on Head-Driven Phrase Structure Grammar (HPSG), I construct a set of parallel, typologically controlled grammars [Pollard and Sag, 1994, Bender et al., 2002, 2010], with Minimal Recursion Semantics (MRS) serving as a shared semantic interface across them [Copestake et al., 2005]. This makes it possible to derive predicate-argument structures from natural English data and realize the same underlying meanings through different artificial

grammars. The resulting corpora are therefore parallel in meaning but differ in grammatical form, retaining semantic and distributional properties derived from the source corpus while providing grammar-level experimental control.

I apply this method to action nominal constructions (ANCs), a class of nominalizations that denote events or processes and retain aspects of the corresponding predicate's argument structure [Koptjevskaja-Tamm, 1993]. For example, the event expressed by the finite clause *John destroyed the city* can also be expressed by nominalized forms such as *the destruction of the city* or *John's destroying the city*. ANCs are particularly useful for controlled typological comparison because they provide alternative ways of encoding predicate–argument relations between clause-like and noun-phrase-like structures, and these encoding patterns are closely connected to other grammatical dimensions such as word order, alignment, and complement structure. This makes ANCs a compact but multidimensional domain for testing how alternative grammatical systems organize these shared and distinguishing properties, and whether LMs show systematic learning preferences among them.

To implement different ANC systems in a controlled experimental setting, I build on the Grammar Matrix analysis of ANCs developed by Ruditsky [2024]. This work implements the ANC typology of Koptjevskaja-Tamm [1993] as a set of HPSG/Grammar Matrix strategies for generating nominalized clauses. I combine four ANC strategies with four additional typological dimensions: three clause word orders, two noun-phrase word orders, two alignment systems, and two complement-system types. Crossing these dimensions yields 96 target languages. The semantic input comes from the BabyLM 2026 Strict corpus [Choshen et al., 2026] and is converted into MRS representations. Each target grammar then realizes the same MRS bank according to its own typological parameters, producing synthetic corpora that are parallel in meaning but differ in morphosyntactic realization.

For each of the 96 target languages, I train three independently initialized GPT-2-small models [Radford et al., 2019] on the corresponding corpus. I then evaluate the models using targeted minimal pairs to determine whether they have learned the grammatical phenomena encoded in the language-specific grammar. The results show that the models can learn

many grammar-specific regularities, but still struggle with zero case marking and ANC-related phenomena. The target-language models also outperform the English baseline on the valency-related phenomena, suggesting that the corpora generated by this method do preserve part of the semantic distribution associated with predicates and arguments.

Overall, the models' learning behavior reveals two main patterns. First, the models generally learn related grammatical structures more easily when they maintain stable linear organization, and this preference is relatively consistent with several patterns in human language typology. Second, when existing surface patterns are insufficient to distinguish related structures, additional overt morphological cues generally improve model learnability. However, human typological distributions do not simply maximize such cues, and some typologically rare parameter combinations can therefore still be relatively easy for the models to learn. This combination of alignment and divergence between model preferences and human typology suggests that learnability may be one factor shaping typological distributions, while human languages are also influenced by other functional and communicative pressures.

The rest of the thesis is organized as follows. Chapter 2 reviews the background on LM typological biases, synthetic-language experiments, HPSG/MRS grammar engineering, and ANCs. Chapter 3 defines the experimental parameter space and presents predictions based on human language typology and prior LM experiments. Chapter 4 describes the pipeline for extracting shared MRS representations from English source data and generating the 96 parallel synthetic corpora. Chapter 5 presents the model training setup for the target-language GPT-2-small models. Chapter 6 reports the targeted minimal-pair evaluation and the main learnability patterns across the 96 grammars. Chapter 7 discusses how the model results align with and diverge from human typological tendencies. Chapter 8 concludes the thesis and discusses the broader implications of the findings. The code and materials used in this thesis are available at `https://github.com/Iskar-Deng/anc-typology-bias`.

Chapter 2

BACKGROUND

2.1 Typological Biases in Language Models

Early linguistic work on LMs focused on a basic question: whether models can acquire grammatical structure without explicit syntactic supervision. Researchers used targeted evaluations to test model behavior on phenomena such as subject–verb agreement, long-distance agreement, and filler-gap dependencies [Linzen et al., 2016, Gulordava et al., 2018, Wilcox et al., 2018, Goldberg, 2019], and used grammatical minimal pairs to test whether models assign higher probability to grammatical sentences than to minimally different ungrammatical ones [Marvin and Linzen, 2018, Warstadt et al., 2020, Hu et al., 2020]. This line of work shows that LMs can go beyond shallow statistical cues and acquire certain nontrivial structural generalizations from language-modeling objectives.

However, this success raises a new question: if LMs can succeed on syntactic evaluations, what kinds of generalizations are they relying on? Models may acquire hierarchical generalizations similar to those used in human grammar, but they may also achieve similar performance by relying on local cues, linear heuristics, or task-specific statistical regularities [McCoy et al., 2019, Linzen, 2020, Newman et al., 2021]. Prior work has shown that when the training data are compatible with both linear and hierarchical generalizations, neural models' choices are influenced by architecture and training conditions, and do not necessarily reflect human-like structural preferences [McCoy et al., 2020, Warstadt and Bowman, 2020]. This motivates a shift from asking whether models can learn a particular structure to asking which generalizations they rely on when several analyses are compatible with the data.

In learning theory, finite input typically does not uniquely determine the target generalization. Multiple hypotheses may be compatible with the observed data. A learner selects

some of these hypotheses because it has certain assumptions, constraints, or preferences, and these tendencies constitute its inductive biases [Mitchell, 1980, Linzen, 2020, McCoy et al., 2020, Warstadt and Bowman, 2020]. For LMs, studying inductive bias means asking which grammatical systems are easier or harder for models to learn, and whether these differences reflect stable learning biases [Ravfogel et al., 2019, White and Cotterell, 2021].

In human language learning, the source of inductive biases has been explained in different ways. The poverty-of-the-stimulus argument holds that the linguistic input available to children is insufficient to fully determine their eventual grammatical knowledge, and that language learning must therefore be constrained by some prior assumptions [Chomsky, 1980, 1986]. These constraints are often understood as language-specific biases that restrict the space of possible grammars available to human learners. Domain-general accounts, by contrast, emphasize that many preferences in linguistic structure need not arise from specifically linguistic priors, but may instead reflect more general cognitive, processing, memory, predictive, or communicative pressures [Hawkins, 2004, Christiansen and Chater, 2008, Kemp et al., 2018, Gibson et al., 2019, Hahn et al., 2020, Imel et al., 2026].

The idea of impossible languages provides a concrete perspective on this issue. If the human language faculty is constrained in specific ways, then some formal systems should fall outside the space of humanly learnable languages. Such systems violate organizational principles normally associated with natural language, such as hierarchical structure and structure-dependent rules, and often rely instead on purely linear order, arbitrary positional counting, or other non-natural-language rules [Moro, 2016]. A central concern in critiques of LMs is therefore that, if LMs could learn possible and impossible languages with equal ease, they would look less like models of human linguistic induction and more like powerful sequence fitters that fail to capture key inductive biases in human language learning [Chomsky et al., 2023, Kallini et al., 2024].

Linguistic typology provides a natural entry point for this question. Typological research uses cross-linguistic comparison to describe and explain structural diversity, commonalities, and correlations among grammatical features in human languages [Comrie, 1989, Croft,

2002]. A basic observation is that human languages do not uniformly cover the space of logically possible grammars. Some structures and structural combinations recur across languages, some are relatively rare, some have not been observed, and others appear highly unnatural [Greenberg, 1963, Dryer and Haspelmath, 2013]. Typological distributions therefore cannot simply be reduced to a binary distinction between possible and impossible languages, but instead constitute a continuous space involving frequency, correlation, and plausibility [Dryer, 1998]. In this sense, typologically implausible languages are often built from components that are independently attested in human languages, but combine them in ways that violate common typological tendencies [Xu et al., 2026]. For example, SOV order and prepositions are both attested in human languages, but SOV languages typically favor postpositions. A system that combines SOV order with prepositions may therefore be viewed as typologically implausible [Greenberg, 1963, Dryer, 1992, Xu et al., 2026].

Typological distributions can reflect human learning preferences to some extent. Human languages are not static systems formed only once, but are repeatedly learned and reconstructed through intergenerational transmission. If certain structures are harder to learn, remember, or transmit stably, they may be less likely to persist in a speech community. Conversely, structures that better align with learners' preferences may be more likely to be acquired, regularized, and repeatedly observed across languages [Kirby et al., 2008, Christiansen and Chater, 2008, Smith and Wonnacott, 2010, Steinert-Threlkeld and Szymanik, 2020]. Artificial language learning studies likewise show that learners do not always passively reproduce their input, but can systematically reshape it into systems that are more regular, more predictable, or more communicatively efficient [Culbertson et al., 2012, Fedzechkina et al., 2012]. Thus, common or rare structural patterns in typology can provide indirect evidence about human learning preferences. Nevertheless, typological distributions cannot be reduced entirely to learning preferences, since diachronic change, language contact, genealogical inheritance, and social factors may also shape the distribution of structures across the world's languages [Dunn et al., 2011, Thomason and Kaufman, 1988, Trudgill, 2011].

Against this background, research on LM typological biases can be understood as ask-

ing whether models' learning preferences among multiple possible grammatical systems are systematically related to distributional patterns in human language typology. If model preferences align with human typology, this suggests that similar preferences can arise in systems centered on predictive learning. Existing work has observed such alignment at different levels. LMs show greater learning difficulty for certain impossible languages [Kallini et al., 2024]. Typologically implausible languages are often learned more slowly or require more training than plausible variants [Xu et al., 2026]. In the domain of word order, typologically common structural combinations can correspond to lower perplexity or greater predictability in cognitively motivated LMs [Kuribayashi et al., 2024]. These results support the possibility that some typological tendencies may not depend entirely on strong language-specific constraints, but may arise at least in part from more general pressures of learning, prediction, memory, or processing.

Conversely, divergence between model preferences and human typology is also theoretically informative. Prior work has found that model inductive biases do not always align with typological tendencies in attested natural languages [White and Cotterell, 2021]. Such divergence suggests that the predictive learning mechanisms captured by current LMs may be insufficient on their own to explain the relevant typological patterns. These patterns may also depend on semantic and pragmatic constraints, communicative goals, discourse structure, social interaction, or diachronic mechanisms that the model does not fully capture [Futrell and Mahowald, 2026]. For complex grammatical phenomena, this comparison is especially useful, because typological distributions often result from multiple interacting dimensions rather than a single mechanism. A model may capture general learning pressures that resemble human language along some dimensions, while lacking other factors involved in human language along others [Deng et al., 2026, Popadich and Steinert-Threlkeld, 2026]. Fine-grained patterns of alignment and divergence can therefore help identify which factors contribute to different typological outcomes.

Therefore, this thesis takes action nominal constructions (ANCs) as a case study. ANCs, together with main clauses and clausal complements, form a set of related grammatical

structures that share certain properties while differing in others. Languages can maintain consistency across these structures, or distinguish them through particular grammatical cues. These choices are in turn related to multiple grammatical dimensions, such as word order, alignment, and complement system. Different languages combine these parameters in different ways, giving rise to rich typological variation. ANCs therefore provide a compact but multidimensional grammatical domain for testing whether LMs show systematic learning preferences among different typological combinations. More detailed discussion of ANCs is provided in Section 2.4.

2.2 Synthetic-Language Experiments

Studying typological preferences in LMs requires comparing the relative learnability of multiple grammatical systems, but direct comparisons across natural languages rarely provide sufficiently controlled evidence. Target typological differences in natural languages are usually bundled with other factors, such as corpus source and genre, training data size, tokenization, and lexical statistics. Even when parallel or translated corpora are used to reduce differences in semantic content, model performance across languages may still be affected by corpus statistics and resource conditions rather than by a specific grammatical parameter [Cotterell et al., 2018, Mielke et al., 2019]. In addition, typological features in natural languages are not independent of one another. Word order, adposition order, genitive order, and relative clause order often show cross-linguistic correlations, while the number of attested natural languages is too small to fully control for all relevant dimensions [White and Cotterell, 2021].

Synthetic languages offer a solution to this problem. They allow different corpora to be made as parallel as possible in content, scale, lexical distribution, and evaluation conditions, while varying primarily in the target grammatical system. Existing work has pursued this goal through two main strategies: perturbing natural-language corpora and generating corpora from artificial grammars.

2.2.1 Corpus Perturbation

Corpus perturbation constructs counterfactual languages by modifying syntactic structures in natural corpora. A typical procedure first obtains parse trees, dependency structures, POS tags, or morphological annotations, and then systematically modifies selected grammatical properties according to predefined rules, such as inserting markers or changing word order. Because the input is derived from natural language, this approach can construct variants that remain close to the source corpus while differing along the target grammatical dimension.

Early work in this tradition started from treebanks. Wang and Eisner [2016] use Universal Dependencies treebanks to generate large numbers of synthetic treebanks with different word-order profiles by reordering the dependents of nouns and verbs. This work established an important methodological idea: synthetic languages that may not be attested in natural language can be constructed from real annotated corpora through systematic dependency reordering. Similarly, Ravfogel et al. [2019] construct synthetic variants of English from a parsed English corpus by manipulating word order, case marking, and agreement systems, and use these variants to study the inductive biases of RNNs in agreement prediction. This design avoids the confounds involved in directly comparing natural languages such as English and Japanese, making it easier to isolate the effects of different grammatical parameters on model learning.

Recent work has extended corpus perturbation from annotated treebanks to more general natural-language corpora. Kallini et al. [2024] start from the BabyLM English corpus and construct a set of impossible languages using perturbation functions such as shuffling, reversal, and marker hopping. They train GPT-2-small models from scratch and show that these models have higher perplexity, slower learning, or weaker targeted syntactic behavior on impossible languages, challenging the claim that LMs can learn possible and impossible languages with equal ease. Xu et al. [2026] extend this approach to typologically implausible languages. Starting from English and Japanese corpora, they use dependency parses to perform span-level swaps involving word-order correlations, constructing harmonic and

non-harmonic variants and comparing LMs' learning difficulty across these counterfactual languages. Deng et al. [2026] apply this approach to differential argument marking, comparing whether LMs show typologically aligned biases in a semantically licensed phenomenon. Related methods have also been used to study how different kinds of linguistic evidence support model learning and generalization [Patil et al., 2024, Misra and Mahowald, 2024, Yao et al., 2025].

The main advantage of corpus perturbation is that it preserves relatively natural input distributions. Because it starts from real corpora, the resulting synthetic corpora often retain lexical frequencies, syntactic complexity, semantic content, and usage distributions found in natural language. This is important for LMs, since the input to language learning consists not only of abstract grammatical rules, but also of the semantic distributions associated with different grammatical positions.

However, the grammatical control provided by this approach is limited. First, perturbation methods typically modify structures already present in the source corpus, making it difficult to introduce a genuinely new and globally coherent grammatical system. Second, these methods are well suited to locally modifiable phenomena such as word order, case marking, or agreement, but are less able to model systematic interactions across multiple grammatical modules. Finally, corpus perturbation often faces a tradeoff between scale and accuracy. High-quality treebanks contain fewer annotation errors but are limited in size and coverage, while larger natural-language corpora provide richer input distributions but rely more heavily on automatic parsing and preprocessing, making them sensitive to parser errors, annotation ambiguities, and preprocessing choices [Xu et al., 2026].

2.2.2 Artificial Grammar

Artificial-grammar methods generate training corpora bottom-up from explicitly defined grammars or formal languages. This approach has often been used in formal-language experiments, where researchers define controlled formal systems to test whether RNNs, LSTMs, and Transformers can learn structural capacities such as counting, nesting, bracket matching,

long-distance dependencies, and length generalization [Weiss et al., 2018, Hahn, 2020, Bhattamishra et al., 2020, Ebrahimi et al., 2020, Hewitt et al., 2020, Delétang et al., 2023, Strobl et al., 2024, Wang and Steinert-Threlkeld, 2023]. Dyck languages, for example, require brackets to be correctly matched and nested, and are therefore commonly used to test whether models can acquire stack-like memory or hierarchy-sensitive generalization [Ebrahimi et al., 2020, Hewitt et al., 2020].

Building on this tradition, another line of work uses artificial grammars that more closely resemble natural language to generate synthetic languages. These grammars typically include word classes, phrase structures, agreement, or word-order parameters, making them more suitable for studying typological inductive biases. White and Cotterell [2021] use a PCFG with multiple word-order switches to construct 64 artificial languages, keeping the corpora parallel in scale and generation conditions while varying only the combination of word-order parameters. They train LSTMs and Transformers to compare the architectures' learning preferences across these word-order systems. Guerin et al. [2024] use a controlled artificial-language setup, extended with semantic-field manipulations, to study how syntactic and semantic proximity affect unsupervised translation. Kuribayashi et al. [2024] use a similar artificial-language space to ask whether LMs' word-order preferences align with typological distributions in WALS [Dryer and Haspelmath, 2013]. They find that models with cognitively motivated constraints, such as syntactic bias, left-corner parsing strategies, or memory limitations, more readily produce preferences aligned with human typology. El-Naggar et al. [2025] further use Generalized Categorial Grammar (GCG) to generate artificial languages with unbounded dependencies and mildly context-sensitive structures, and find that typologically plausible word orders better support length generalization.

Artificial grammar methods offer strong experimental control. Because the corpora are generated from explicitly specified grammars, the grammar parameter space can be defined directly, and different synthetic languages can be made highly parallel while varying only in the target parameter or parameter combination. This makes the approach well suited for minimal contrasts, causal comparisons, and the exploration of rare or unattested grammatical

systems that are difficult to obtain through corpus perturbation.

The cost of this control is reduced naturalism. In formal-language and artificial-grammar corpora, lexical distributions, generation probabilities, and construction frequencies are usually specified by the researcher, making it difficult to approximate the predicate-argument distributions, selectional restrictions, and syntax-semantics interactions found in natural language. Thus, artificial grammars are useful for testing clean causal hypotheses, but less suitable on their own for studying complex constructions whose learnability may depend on naturalistic semantic and distributional input.

2.3 HPSG, MRS, and the Grammar Matrix

This thesis proposes a grammar-engineered method for synthetic corpus construction, designed to combine the strengths of the two approaches discussed in Section 2.2. Specifically, I use the Grammar Matrix to construct typologically controlled HPSG grammars [Pollard and Sag, 1994, Bender et al., 2002, 2010], and use MRS as a shared semantic interface across grammars [Copestake et al., 2005]. In the implementation, I use PyDelphin to call ACE, so that the same set of MRS inputs can be realized by different target grammars as different morphosyntactic forms [Goodman, 2019, Packard, 2018]. With this design, the predicate-argument structures and semantic distributions of the synthetic languages can be derived from real corpora, while grammatical variation is controlled at the level of the target HPSG grammars.

2.3.1 Head-Driven Phrase Structure Grammar

Head-Driven Phrase Structure Grammar (HPSG) is a lexicalist, constraint-based grammatical framework [Pollard and Sag, 1994, Sag et al., 2003]. In HPSG, words and phrases are represented as typed feature structures. These structures encode not only syntactic category, but also multiple layers of linguistic information such as valence requirements, morphosyntactic features, and semantic contribution. HPSG therefore provides a formal framework for representing lexical information, phrase structure, argument structure, morphosyntactic

features, and semantics in a unified system. Compared with many more abstract formal grammars, HPSG representations include syntactic constraints, lexical constraints, and a semantic interface, making the framework well suited for describing interactions among multiple grammatical modules in natural language.

2.3.2 Minimal Recursion Semantics

Minimal Recursion Semantics (MRS) is a semantic representation commonly used in HPSG-based grammar engineering [Copestake et al., 2005]. MRS represents sentence meaning using elementary predications, argument relations, and scope constraints, thereby encoding predicate-argument structure while allowing certain scope relations to remain underspecified. In this work, MRS serves as a shared semantic interface across grammars. The same MRS can represent the same event structure, participant relations, and basic semantic content, while different target grammars can realize this semantic input in different surface forms. Figure 2.1 provides an example MRS for a sentence containing an ANC.

2.3.3 The Grammar Matrix

The Grammar Matrix is an HPSG-based grammar customization system designed to support the rapid development of typologically diverse computational grammars [Bender et al., 2002, 2010]. Researchers specify typological properties of the target language in a choices file, such as basic word order, case marking, tense/aspect/mood, and inflectional morphology. The system then generates a set of HPSG grammar source files, which can subsequently be compiled into a grammar for parsing and generation.

2.3.4 DELPH-IN Tools

ACE and PyDelphin are two DELPH-IN tools used in this thesis. ACE is a parsing and generation engine for compiled HPSG grammars: given a compiled grammar, ACE can parse surface strings into MRS representations and generate surface strings from MRS inputs

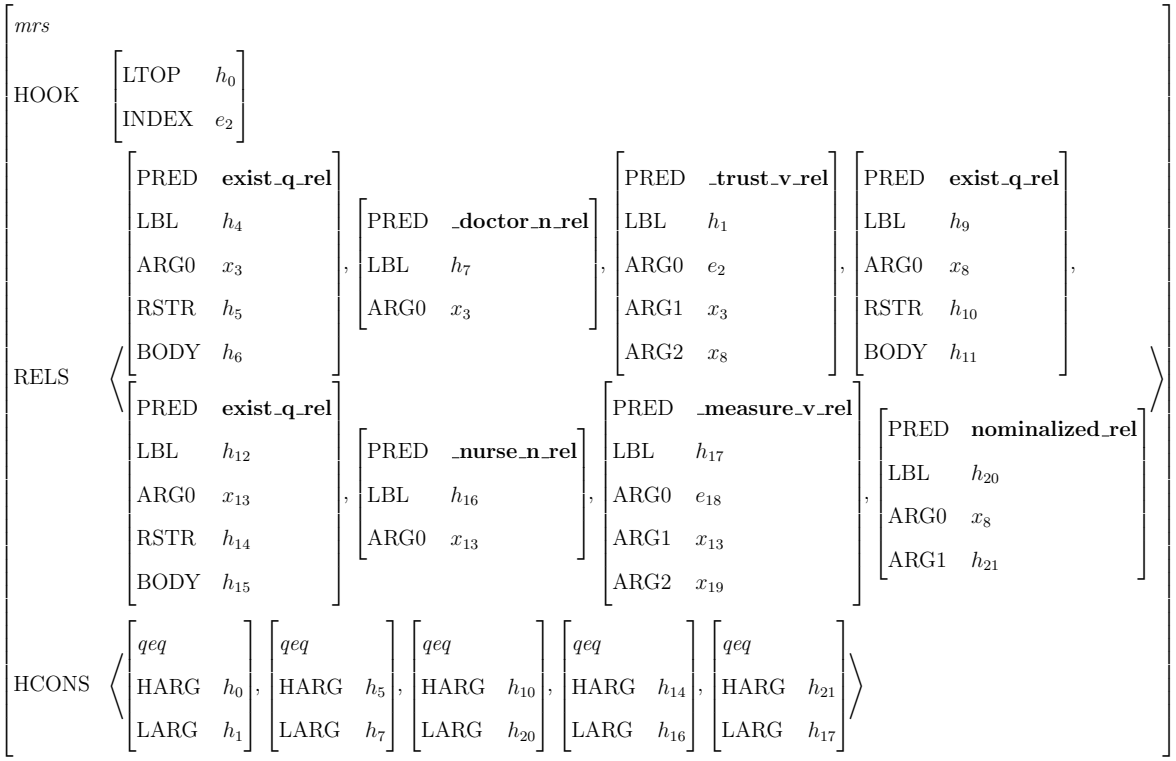

Figure 2.1: MRS representation of *The doctor trusted the nurse's measurement.*

[Packard, 2018]. PyDelphin provides an interface for working with DELPH-IN representations and for calling ACE within a Python experimental pipeline [Goodman, 2019]. The same set of MRS inputs is passed to different target grammars, and ACE generates the corresponding morphosyntactic realizations.

2.4 Action Nominal Constructions

Action nominal constructions (ANCs) are nominalizations that denote events or processes and retain aspects of the corresponding predicate's argument structure [Koptjevskaja-Tamm, 1993]. Unlike many ordinary entity-denoting nouns, ANCs remain associated with an event and its participants. Their head is typically derived from a verbal predicate, and the nominalized structure may still express participants licensed by that predicate [Grimshaw, 1990, Alexiadou, 2001]. ANCs therefore involve not only nominal derivation, but also the expres-

sion of event semantics and argument structure.

Syntactically, ANCs have a mixed categorial status. Externally, they pattern like noun phrases and can occur in nominal positions such as subjects, objects, or prepositional complements. Internally, they may preserve clausal or verbal properties, including argument selection, event interpretation, case marking, nonfinite morphology, or other verbal functional structure [Grimshaw, 1990, Koptjevskaja-Tamm, 1993, Alexiadou, 2001]. ANCs therefore differ both from ordinary noun phrases and from full clauses, carrying predicative content inside a nominal structure and linking nominal syntax with verbal argument structure.

From a broader grammatical perspective, ANCs are related to main clauses and clausal complements, while differing from them in grammatical status. All three express predicate–argument relations and may involve the same event and its participants, but they differ in their syntactic distribution and semantic functions. Main clauses typically constitute independent assertions, directly presenting an event or state. Clausal complements, by contrast, embed predicative content within a matrix clause as an argument of the matrix predicate; depending on the matrix predicate, this content may express a proposition, fact, event, or process. ANCs can likewise occur in argument positions, but they typically nominalize an event or process, allowing it to be referred to and further discussed rather than directly asserting its occurrence [Koptjevskaja-Tamm, 1993]. Thus, clausal complements and ANCs may overlap in their external syntactic positions, but their semantic functions differ. Languages also commonly distinguish between these related constructions through grammatical cues. Section 2.4.1 illustrates this partial overlap and distinction with sentential ANCs in Korean.

The main differences among ANC strategies lie in how the arguments of the nominalized predicate are encoded. The relevant arguments include the agent-like argument of a transitive predicate (A), the sole argument of an intransitive predicate (S), and the patient-like argument of a transitive predicate (P). These arguments may retain verbal marking similar to that found in finite clauses, or they may be reorganized as genitive, oblique, or adpositional expressions inside the noun phrase. These differences can be viewed as a continuum

Type	Transitive ANC	Intransitive ANC	Example language
SENT	John-NOM destruction city-ACC	John-NOM running	Korean
POSS-ACC	John-GEN destruction city-ACC	John-GEN running	Selkup
ERG-POSS	John-INSTR destruction city-GEN	John-GEN running	Russian
NOMN	John-GEN destruction city-OBL	John-GEN running	Icelandic

Table 2.1: Four major nominalization strategies illustrated with pseudo-examples. Case markers are simplified for expository purposes.

from more clause-like to more noun-phrase-like structures.

A further property of ANCs is that their arguments need not all be overtly expressed. ANCs often present events as known, identifiable, or backgrounded objects, while the sentence may focus on the evaluation, result, temporal relation, or broader discourse context of the event. As a result, when an argument is recoverable from context or irrelevant to the communicative goal of the sentence, it may remain unexpressed [Koptjevskaja-Tamm, 1993]. The ANC types discussed below should therefore be understood as patterns of argument encoding when arguments are overtly realized, rather than as requiring every ANC to express all of its possible arguments.

2.4.1 Major Types of ANC Strategies

Koptjevskaja-Tamm [1993] classifies action nominal constructions according to how the A, S, and P arguments of the nominalized predicate are encoded. These strategies form a continuum from more clause-like to more noun-phrase-like structures. Sentential ANCs are closest to clausal structure, nominal ANCs are closest to canonical noun-phrase structure, and possessive-accusative and ergative-possessive ANCs occupy intermediate positions, combining verbal and nominal resources. It is important to note that a single language may have more than one ANC strategy, depending on the nominalizer, lexical class, construction type,

or discourse context.

Here, argument encoding refers to how an ANC signals the relation between the action nominal and its participants. This relation can be expressed through dependent marking, head marking, or word order [Koptjevskaja-Tamm, 1993, Nichols, 1986]. This thesis focuses primarily on how the major ANC strategies differ in dependent marking, where role marking appears on the participant NP. Table 2.1 illustrates the four strategies with simplified dependent-marking patterns.

Sentential ANCs

Sentential (SENT) ANCs occupy the most clause-like end of the typology. Their defining property is that the arguments of the nominalized predicate retain the dependent-marking patterns found in finite clauses, rather than being recoded as genitive or possessive dependents inside an ordinary noun phrase [Koptjevskaja-Tamm, 1993].

Korean is an example of the SENT strategy. In ordinary finite clauses, Korean marks subjects with nominative case, direct objects with accusative case, and indirect objects with dative case. In corresponding nominalized constructions, these arguments retain the same case marking. Because of this property, the boundary between ANCs and clausal complements can be relatively subtle in SENT languages. The examples below show that a Korean ANC can occur externally as a nominal constituent and bear accusative case as the object of the matrix predicate, while internally preserving clause-like dependent marking.

(1) Korean finite clause [Koptjevskaja-Tamm, 1993, ex. 5.1]

John i Mary eke c^{h}æk ŭl cu-əs'-ta
John NOM Mary DAT book ACC give-PAST-FIN

'John gave a book to Mary.'

(2) Korean SENT ANC [Koptjevskaja-Tamm, 1993, ex. 5.6]

[Ku ton ŭl Pak sənsæŋ eke cu-kı] lŭl palæ-yo
that money ACC Pak mister DAT give-AN ACC hope-FIN

'I hope you gave that money to Mr Pak.'

Korean ANCs may also retain verbal properties such as tense, aspect, and adverbial modification. However, they do not preserve all properties of independent main clauses. For example, topic particles are more restricted in nominalized or subordinate contexts than in independent main clauses. Thus, although SENT ANCs are strongly clause-like in their internal argument encoding, they remain distinct from finite clauses in their external distribution and in some syntactic properties.

Possessive-Accusative ANCs

Possessive-Accusative (POSS-ACC) ANCs stand between the sentential and nominal types. In this type, the S and A arguments of the nominalized predicate are recoded as genitive-like dependents, while the P argument retains the object marking found in finite clauses [Koptjevskaja-Tamm, 1993]. This type therefore combines nominal and clausal resources, with A/S behaving like possessors inside the ANC and P continuing to behave like a direct object.

Selkup is an example of the POSS-ACC strategy. In Selkup, objects in ordinary finite clauses may bear accusative marking. In corresponding ANCs, A or S takes genitive-like marking, while P retains accusative marking. In the example below, 'my brother' is the A argument of the nominalized predicate and bears genitive marking, while 'wife' is the P argument and retains accusative marking.

(3) Selkup POSS-ACC ANC [Koptjevskaja-Tamm, 1993, ex. 6.10a]

Mat ašša tɛnymy-s-ak [timńa-n-y ima-p
I:NOM NEG know-PAST-1SG brother-GEN-1SG.POSS wife-ACC
qo-ptä-∅-ty]
find-AN-NOM-3SG.POSS

'I did not know about my brother's finding a wife.'

POSS-ACC ANCs illustrate the mixed categorial status of ANCs, since different arguments within a single nominalized predicate can be mapped onto nominal and clausal marking systems.

Ergative-Possessive ANCs

Ergative-Possessive (ERG-POSS) ANCs also stand between the sentential and nominal types. Unlike POSS-ACC ANCs, ERG-POSS groups S and P together by assigning them genitive-like marking, while A is marked separately, often by an oblique or instrumental expression [Koptjevskaja-Tamm, 1993]. ERG-POSS therefore creates an ergative-like alignment inside the ANC, contrasting with the accusative-like alignment in POSS-ACC ANCs.

Russian is an example of the ERG-POSS strategy. In Russian finite clauses, A is typically marked with nominative case and P with accusative case. In corresponding ANCs, however, P can be encoded as genitive, while A is expressed with instrumental case. The examples below show this contrast: in the finite clause, *Aleksandr* is the nominative A and *Egipet* is the accusative P; in the corresponding ANC, *Egipt-a* 'Egypt' is the genitive P, while *Aleksandr-om* 'Alexander' is the instrumental A.

(4) Russian finite clause [Koptjevskaja-Tamm, 1993, ex. 1.8]

Aleksandr zavoeva-l Egipet
Alexander:NOM conquer-PAST.PERF Egypt:ACC

'Alexander conquered Egypt.'

(5) Russian ERG-POSS ANC [Koptjevskaja-Tamm, 1993, ex. 1.9a]

zavoeva-nij-a Egipt-a Aleksandr-om
conquer-AN-GEN Egypt-GEN Alexander-INSTR

'the conquest of Egypt by Alexander'

Nominal ANCs

Nominal (NOMN) ANCs are the type most closely aligned with ordinary nominal structure. In this type, S and A are encoded as genitive-like dependents, while P is assimilated to a dependent type available in ordinary noun phrases [Koptjevskaja-Tamm, 1993].

NOMN can be divided into two subtypes: Double-Possessive and Possessive-Adnominal. In the Double-Possessive subtype, S, A, and P all receive genitive-like marking, and this pattern is especially natural in languages with multiple genitive positions. For example, English has two genitive positions, the prenominal *'s*-genitive and the postnominal *of*-genitive. When both A and P are overtly expressed, they can occupy these two positions, as in *Peter's shooting of the prisoners*. In the Possessive-Adnominal subtype, S/A receive genitive-like marking, while P is expressed by an oblique or adpositional dependent available in ordinary noun phrases.

Since ANCs in actual use often omit arguments that are recoverable from context or irrelevant to the communicative goal, a single genitive slot is often sufficient to express the most relevant event participant. Thus, when only one argument is overtly expressed, the surface difference between Double-Possessive and Possessive-Adnominal ANCs is limited.

Finnish provides an example of the Double-Possessive subtype. Finnish can accommodate two formally identical but functionally distinct genitive dependents within the same noun phrase. When two genitive dependents occur together, the genitive farther from the head typically has a determiner-like function, specifying the reference of the head, while the genitive closer to the head has a descriptive or characterizing function. In ANCs, the former typically corresponds to A and the latter to P. In the example below, *vanhempien* 'parents' is the genitive dependent farther from the action nominal *antaminen* 'giving' and functions as A, while *taloudellisen tuen* 'economic support' is adjacent to the head and functions as P.

(6) Finnish NOMN ANC [Koptjevskaja-Tamm, 1993, ex. 8.3]

[Vanhempien	taloudellisen	tuen	antaminen]	on	riippuvaista	tuloista.	
parents:GEN	economic	support:GEN	give:AN	is	dependent	incomes:PART	

'Parents' giving of economic support is dependent on their incomes.'

2.4.2 The ANC Library in the Grammar Matrix

Ruditsky [2024] builds on the nominalized-clause library in the Grammar Matrix [Howell et al., 2018] and extends it with an implementation specifically designed for ANCs. Users can define one or more nominalization strategies for a language, and for each strategy, specify the argument-marking pattern, internal word order, determiner requirements, and permitted types of modification. Each strategy is then associated with the relevant nominalizing morphology, which derives an action nominal from a verbal lexical entry. Figure 2.2 shows the questionnaire interface for one such nominalization strategy.

During this derivation, the syntactic category of the lexical item changes from verb to noun, while the event semantics and argument relations of the source verb are preserved. The system can therefore derive corresponding action nominal forms from verbal lexical entries through nominalization rules, rather than requiring a separate lexical entry for every nominalized form. The nominalization rules also adjust the valence of the derived item so that its participants can be realized according to the requirements of the selected ANC strategy.

The library allows action nominals to exhibit both nominal and verbal properties. A nominalized form may take nominal morphology, including case, number, and possessive marking, while retaining verbal morphology where the language permits it. The internal word order of an ANC can be specified independently of the basic word order of finite clauses, and its modification pattern can be configured to allow adjectival modification, adverbial modification, or both.

With respect to argument realization, the core arguments of an ANC are optional by default. If a language requires a particular argument to be overtly expressed, that argument can instead be specified as obligatory. The system also treats transitive and intransitive nominalizations separately, allowing them either to share a single strategy or to be subject to different constraints.

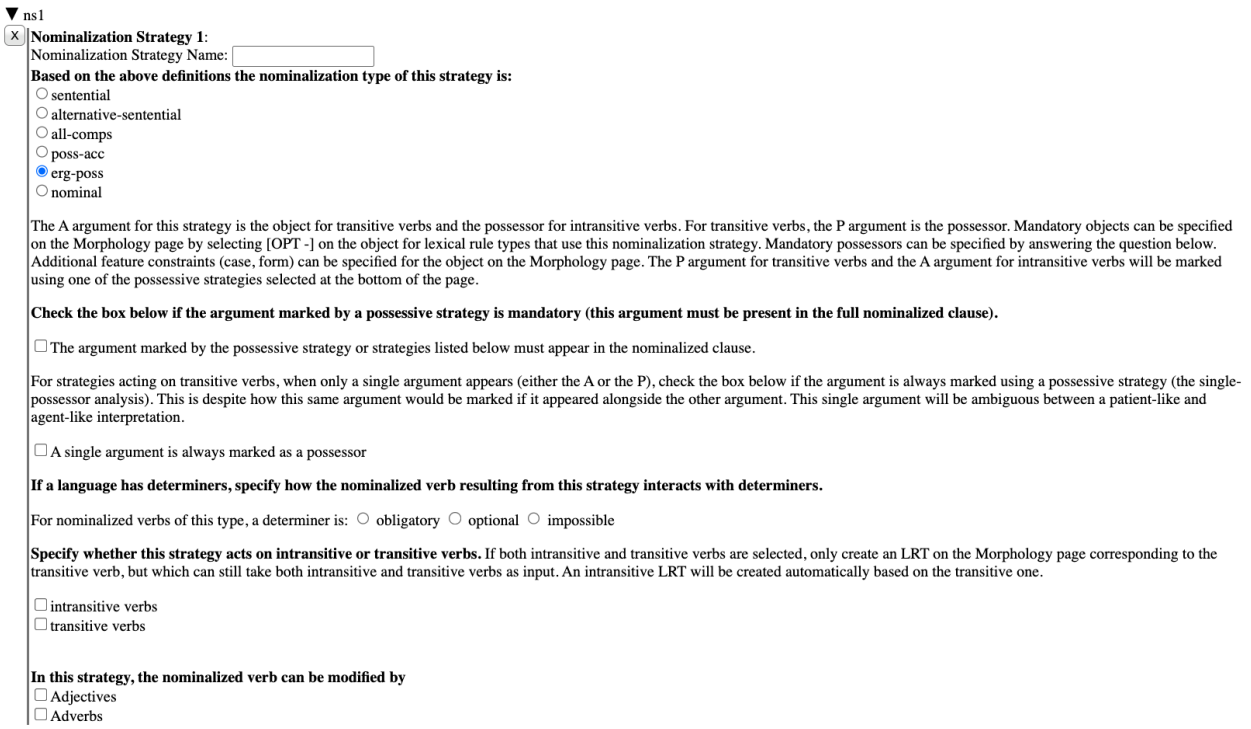

Figure 2.2: Grammar Matrix questionnaire interface for defining an ANC nominalization strategy.

Taken together, the work reviewed in this chapter motivates the experimental design of this thesis. Synthetic-language experiments offer a controlled way to study typological learning preferences, and the Grammar Matrix provides a systematic framework for building different grammatical systems while retaining shared semantic content. ANCs serve as a useful test domain because they connect different linguistic structures and allow us to examine interactions among several typological parameters. Building on this background, Chapter 3 defines the specific parameter space used in the experiment and summarizes the corresponding typological predictions.

Chapter 3

EXPERIMENTAL PARAMETERS AND TYPOLOGICAL PREDICTIONS

This thesis constructs synthetic languages that are parallel in semantic content and lexical basis, but differ in their grammatical systems, by controlling a set of typological parameters. This chapter introduces this experimental parameter space. It first defines the main parameters manipulated in the experiment, then discusses their distribution in human language typology and their possible interactions, and finally summarizes what prior work on LM typological biases suggests about their relative learnability.

3.1 Experimental Parameters

To construct synthetic languages with different ANC grammar systems, this thesis controls five main experimental parameters: clause word order, NP word order, alignment, complement-system type, and ANC strategy. Table 3.1 summarizes these parameters and their values. These parameters determine the core grammatical configuration of each target language, and their cross-product yields 96 target languages.

To describe these target languages more explicitly, I use a set of abstract profile variables to represent the core morphosyntactic properties of each language. Table 3.2 lists the profile variables tracked in this thesis. These profile variables are not all independent experimental parameters. Some correspond directly to the five crossed parameters, while others are derived from their combinations. A concrete example of this mapping is provided later in Table 3.7. The following sections describe these variables in more detail.

Dimension	Values	Size
Clause word order	SOV / SVO / VOS	3
NP word order	GN / NG	2
Alignment	NOM-ACC / ERG-ABS	2
Complement system	Balancing / Deranking	2
ANC strategy	SENT / POSS-ACC / ERG-POSS / NOMN	4

Table 3.1: Five main typological parameters defining the experimental grammar space.

3.1.1 Finite Clauses

Finite clauses are controlled by two main parameters: clause word order and alignment. Clause word order controls the relative order of the predicate and its core arguments in finite clauses. This thesis includes three values: SOV, SVO, and VOS. SOV and SVO are the two most frequent basic word orders cross-linguistically, while VOS serves as a representative verb-initial order [Dryer, 2013b].

Alignment determines the dependent-marking pattern of A, S, and P in finite clauses. Nominative–accusative alignment (NOM-ACC) groups S and A together, while P is marked as the distinct argument. Ergative–absolutive alignment (ERG-ABS) groups S and P together, while A is marked as the distinct argument.

This thesis adopts a simplified marked-argument design: the two arguments that pattern together receive zero marking, while the distinct argument receives overt marking. This design reflects a common cross-linguistic case-marking pattern, in which the argument role distinguished from the other two core arguments is often the one that receives overt case marking [Comrie, 2013]. For comparability, the same overt marker *-ca* is used in both alignment systems. Table 3.3 shows how alignment determines finite clause argument marking.

Profile variable	Meaning	Values
CLAUSE_WO	finite clause order of A/S, P, and V	SOV, SVO, VOS
FIN_S_MARK	dependent marking of S in finite intransitive clauses	0
FIN_A_MARK	dependent marking of A in finite transitive clauses	0, -ca
FIN_P_MARK	dependent marking of P in finite transitive clauses	0, -ca
FIN_V_MARK	verbal morphology on finite verbs	-s
NP_WO	order of genitive dependent and nominal head	GN, NG
GEN_MARK	marking of genitive dependents	-ge
COMP_V_MARK	verbal morphology in clausal complements	-s, -ing
ANC_IV_ORDER	internal order of S and the head in intransitive ANCs	SV, VS
ANC_TV_ORDER	internal order of A, P, and the head in transitive ANCs	APV, AVP, PVA, VPA
ANC_S_MARK	dependent marking of S in ANCs	0, -ge
ANC_A_MARK	dependent marking of A in ANCs	0, -ca, -ge, -ob
ANC_P_MARK	dependent marking of P in ANCs	0, -ca, -ge, -ob
ANC_V_MARK	morphology on nominalized predicates	-ing

Table 3.2: Abstract profile variables for the morphosyntactic configuration of each target language.

3.1.2 Noun Phrases

Within noun phrases, this thesis models only one type of genitive dependent relation. The only parameter manipulated here is NP word order, which determines the relative order of the possessor and the possessum. This parameter has two values: GN, in which the genitive dependent precedes the nominal head, and NG, in which the nominal head precedes the genitive dependent.

This parameter is needed because several ANC strategies recode some arguments of the nominalized predicate as genitive-like dependents. For simplicity, each nominal head is allowed only one direct genitive dependent. This means that two genitives cannot directly modify the same head, but a genitive dependent may itself contain another genitive relation.

Alignment	FIN_S_MARK	FIN_A_MARK	FIN_P_MARK
NOM-ACC	0	0	*-ca*
ERG-ABS	0	*-ca*	0

Table 3.3: Finite-clause argument marking under the two alignment systems.

Complement system	FIN_V_MARK	ANC_V_MARK	COMP_V_MARK
Balancing	*-s*	*-ing*	*-s*
Deranking	*-s*	*-ing*	*-ing*

Table 3.4: Verb-form profiles under the two complement-system settings.

3.1.3 Complement System

The target grammars in this thesis contain three relevant verbal environments: main-clause verbs, verbs in clausal complements, and action nominal predicates. The form of these verbs is determined by the complement-system parameter. Typological studies of subordination often distinguish between balancing and deranking constructions. In balancing constructions, the subordinate predicate has a form similar to one that can occur in an independent declarative clause. In deranking constructions, the subordinate predicate appears in a reduced form that cannot normally occur as the predicate of an independent main clause [Stassen, 1985, Cristofaro, 2003].

Therefore, in this experiment, main-clause verbs are assigned the finite suffix *-s*, while action nominal predicates are assigned the nonfinite nominalizing suffix *-ing*. The complement-system parameter determines the form of verbs in clausal complements. In the balancing condition, complement verbs take the finite suffix *-s*; in the deranking condition, complement verbs take the nonfinite suffix *-ing*. Table 3.4 summarizes this mapping.

3.1.4 Clausal Complements

Apart from verb form, this thesis does not introduce additional distinctions for clausal complements. Argument marking in clausal complements follows the same alignment system as in main clauses, so A, S, and P in embedded clauses are realized according to the finite-clause marking pattern of the target language.

For simplicity, the target grammars do not include overt complementizers by default. Overt complementizers can mark a clause as the complement of a matrix predicate and provide an explicit cue to the boundary between the embedded clause and the main clause [Noonan, 2007]. Since embedded clauses may occur without overt complementizers, this dimension is not manipulated in this thesis.

3.1.5 ANC Case Marking

The case marking of an ANC is controlled by the ANC strategy parameter. Natural languages may employ more than one ANC strategy, but each target language in this thesis is assigned exactly one ANC strategy for experimental control. The present experiment includes four ANC strategies: SENT, POSS-ACC, ERG-POSS, and NOMN.

Each strategy determines how the S, A, and P arguments of the nominalized predicate are marked. SENT preserves the argument-marking pattern of the corresponding finite clause. POSS-ACC realizes S and A as genitive-like dependents, while P retains the object marking of the finite clause. ERG-POSS realizes S and P as genitive-like dependents, whereas A receives oblique-like marking. NOMN realizes S and A as genitive-like dependents and P as an oblique-like dependent. Table 3.5 summarizes these marking profiles.

3.1.6 ANC Word Order

In this experiment, ANC-internal word order is derived from the interaction of clause word order, NP word order, and ANC strategy. This design is motivated by the observation that ANC-internal ordering can generally be related to ordering patterns already found in finite

ANC strategy	ANC_S_MARK	ANC_A_MARK	ANC_P_MARK
SENT	FIN_S_MARK	FIN_A_MARK	FIN_P_MARK
POSS-ACC	*-ge*	*-ge*	FIN_P_MARK
ERG-POSS	*-ge*	*-ob*	*-ge*
NOMN	*-ge*	*-ge*	*-ob*

Table 3.5: Dependent-marking profiles of the four ANC strategies.

clauses and ordinary noun phrases [Koptjevskaja-Tamm, 1993].

Concretely, SENT ANCs preserve the internal organization of the finite clause, so they directly retain finite-clause word order. The other three strategies combine structural properties of clauses and noun phrases, so their internal word order is determined according to the following three principles:

1. The position of the possessive-like argument is determined by NP word order. Under GN order, this argument precedes the nominalized predicate; under NG order, it follows the nominalized predicate.

2. The position of the non-possessive argument relative to the nominalized predicate follows the head–complement direction of the finite clause. In an OV language, this argument precedes the predicate; in a VO language, it follows the predicate.

3. When the preceding two principles place A and P on the same side of the nominalized predicate, P is placed closer to the predicate than A. This ordering reflects the structural asymmetry between the two arguments, with P forming a constituent with the predicate before A is introduced outside that constituent.

The ANC strategy determines which arguments occupy the possessive-like and non-possessive positions. In POSS-ACC and NOMN ANCs, S and A occupy the possessive-like

Word Order	SENT	POSS-ACC	ERG-POSS	NOMN
SOV + GN	SV/APV	SV/APV	SV/APV	SV/APV
SVO + GN	SV/AVP	SV/AVP	SV/PVA	SV/AVP
VOS + GN	VS/VPA	SV/AVP	SV/PVA	SV/AVP
SOV + NG	SV/APV	VS/PVA	VS/AVP	VS/PVA
SVO + NG	SV/AVP	VS/VPA	VS/VPA	VS/VPA
VOS + NG	VS/VPA	VS/VPA	VS/VPA	VS/VPA

Table 3.6: ANC-internal word orders determined by the interaction of clause word order, NP word order, and ANC strategy.

position, and P occupies the non-possessive position. In ERG-POSS ANCs, S and P occupy the possessive-like position, while A occupies the non-possessive position.

Based on these principles, I obtain the ANC word-order patterns shown in Table 3.6.

3.1.7 An Example Grammar Profile

Taken together, these five parameters define a space of 96 target languages ($3 \times 2 \times 2 \times 2 \times 4$). Each language is represented by a distinct combination of parameter values and corresponds to a unique morphosyntactic profile. Table 3.7 presents one complete example.

3.2 Typological Predictions from Human Languages

This section discusses typological tendencies associated with the five experimental parameters, the profile variables derived from them, and their combinations. It is important to note that the five parameters used in this thesis, as well as all of their values, are independently attested in human languages. I therefore do not treat any single parameter value as typologically unnatural. Although some values are less frequent than others, they are nevertheless attested types of human language. For example, VOS is rarer than SOV and SVO, and ERG-ABS alignment is less frequent than NOM-ACC alignment [Dryer, 2013b,

Parameter settings		Derived morphosyntactic profile	
Parameter	Value	Profile variable	Value
Clause word order	SVO	CLAUSE_WO	SVO
NP word order	GN	FIN_S_MARK	0
Alignment	NOM-ACC	FIN_A_MARK	0
Complement system	Deranking	FIN_P_MARK	*-ca*
ANC strategy	POSS-ACC	FIN_V_MARK	*-s*
		NP_WO	GN
		GEN_MARK	*-ge*
		COMP_V_MARK	*-ing*
		ANC_IV_ORDER	SV
		ANC_TV_ORDER	AVP
		ANC_S_MARK	*-ge*
		ANC_A_MARK	*-ge*
		ANC_P_MARK	*-ca*
		ANC_V_MARK	*-ing*

Table 3.7: An example of an experimental parameter combination and its derived morphosyntactic profile.

Comrie, 2013].

Accordingly, this section focuses on combinations of parameters and asks whether certain combinations are rare in human languages, violate common typological correlations, or give rise to more mixed or less coherent morphosyntactic profiles. These predictions should be understood as gradient typological tendencies rather than absolute grammatical constraints. If a combination is rare or unattested in existing samples, this does not mean it is an impossible language. Table 3.8 summarizes the main typological tendencies identified in this section.

Dimension	Typologically common	Typologically rare
Clause order × genitive order	SOV+GN VOS+NG	SOV+NG VOS+GN
Clause order × alignment	SVO+NOM-ACC	SVO+ERG-ABS
Word order × complement system	SOV/GN+Deranking	–
Complement system × ANC strategy	Deranking+SENT Balancing+ERG-POSS Balancing+NOMN	Deranking+NOMN
Alignment × ANC strategy	–	ERG-ABS+POSS-ACC ERG-ABS+NOMN
Word order × ANC strategy	SOV/GN+SENT SOV/GN+POSS-ACC SVO/NG+ERG-POSS SVO/GN+NOMN	V-initial+NOMN

Table 3.8: Summary of typological tendencies in the experimental parameter space.

3.2.1 Clause Word Order and Genitive Order

Clause word order and genitive order within the noun phrase are not distributed independently. For the three clause word orders included in this study, the distribution of GN and NG order in WALS is shown in Table 3.9 [Dryer, 2013b,a].

In SOV languages, GN is substantially more frequent than NG, whereas in VOS languages, NG is substantially more frequent than GN. These tendencies are consistent with head-order harmony between clause structure and noun phrase structure: both SOV and

Clause WO	GN	NG
SOV	398	26
SVO	106	249
VOS	2	21

Table 3.9: Distribution of genitive order by clause word order in WALS.

Clause WO	Neutral	NOM-ACC	ERG-ABS
SOV	29	29	16
SVO	39	9	1
VOS	5	1	2

Table 3.10: Distribution of alignment type by clause word order in WALS.

GN exhibit dependent–head order, while both VOS and NG exhibit head–dependent order. Accordingly, SOV+GN and VOS+NG are defined as harmonic profiles, whereas SOV+NG and VOS+GN are defined as non-harmonic profiles [Greenberg, 1963, Dryer, 1992].

SVO languages are generally classified as VO languages on the basis of the verb–object relation, and SVO+NG is therefore more frequent than SVO+GN. Nevertheless, SVO+GN is also common in human languages.

3.2.2 Clause Word Order and Alignment

Clause word order may also be related to alignment type. Table 3.10 shows the distribution of the three clause word orders and alignment types in WALS [Dryer, 2013b, Comrie, 2013]. Here, neutral refers to systems without an overt case-marking distinction.

As shown in the table, SOV languages with overt alignment marking are more often NOM-ACC than ERG-ABS, although ERG-ABS systems are also well represented. The NOM-ACC preference is stronger in SVO languages, where only one language in the sample

Word-order profile	Deranking	Balancing
SOV/GN	24	7
SOV/NG	0	4
V-initial/NG	2	9
SVO/GN	1	5
SVO/NG	1	13

Table 3.11: Distribution of complement systems by word-order profile in the sample of Koptjevskaja-Tamm [1993].

is ERG-ABS. VOS shows the opposite direction, with more ERG-ABS than NOM-ACC languages, but the number of VOS languages is too small for this to be treated as a major trend.

3.2.3 Word Order and Complement System

Clause and NP word order also show an association with complement-system type. In the sample of Koptjevskaja-Tamm [1993], 24 of the 28 deranking languages have SOV/GN order, whereas balancing languages are distributed across a wider range of word-order profiles, as shown in Table 3.11. The combination of SOV/GN order and deranking can therefore be regarded as typologically well supported. However, Koptjevskaja-Tamm [1993] notes that the source of this correlation remains insufficiently understood.

3.2.4 Complement System and ANC Strategy

Complement system and ANC strategy show a clear typological correlation. Their distribution in the sample of Koptjevskaja-Tamm [1993] is shown in Table 3.12.

As shown in the table, SENT occurs predominantly in deranking languages, whereas ERG-POSS and NOMN occur predominantly in balancing languages. POSS-ACC is dis-

ANC strategy	Deranking	Balancing
SENT	14	3
POSS-ACC	12	13
ERG-POSS	3	22
NOMN	0	14

Table 3.12: Distribution of complement-system type by ANC strategy in the sample of Koptjevskaja-Tamm [1993].

tributed almost evenly across the two complement-system types. The distribution therefore forms a gradient, in which the proportion of balancing languages increases from SENT through POSS-ACC and ERG-POSS to NOMN.

This distribution can be understood in terms of the different functions that ANCs serve in the two complement systems. In deranking languages, ANCs often serve a broad range of complement functions and therefore tend to retain sentential dependent marking. SENT represents the clearest case of this pattern. In balancing languages, by contrast, finite clauses are independently available for complementation, so ANCs can assimilate more closely to ordinary noun phrases, as reflected in ERG-POSS and NOMN.

3.2.5 Alignment and ANC Strategy

Alignment and ANC strategy show a clear but asymmetric typological relationship. In the sample of Koptjevskaja-Tamm [1993], NOM-ACC languages are found with all four ANC strategies. By contrast, ERG-ABS languages are concentrated mainly in SENT and ERG-POSS. POSS-ACC is represented only by Agul, where it occurs as an alternative to the more frequent SENT pattern, while NOMN is unattested. NOM-ACC languages with ERG-POSS ANCs, however, are not uncommon. The restriction is therefore not a simple effect of alignment mismatch, but shows a clear directional asymmetry.

ANC strategy	SOV/GN	SOV/NG	V-initial/NG	SVO/GN	SVO/NG
SENT	14	0	1	0	0
POSS-ACC	12	0	6	0	5
ERG-POSS	3	4	8	1	7
NOMN	2	0	1	4	1

Table 3.13: ANC strategy by word-order profile in the sample of Koptjevskaja-Tamm [1993].

Koptjevskaja-Tamm [1993] suggests that this asymmetry may reflect different motivations for the two ANC patterns. In ERG-ABS languages, the special marking of A serves not only to encode syntactic relations but also to distinguish semantic roles, and may therefore persist under nominalization. POSS-ACC and NOMN require A to take possessive-like marking, thereby eliminating the original ergative marking, and are consequently disfavored. By contrast, ERG-POSS in NOM-ACC languages need not be inherited from finite-clause alignment. Transitive ANCs often occur with P as the only overt argument. Such constructions can be assimilated to ordinary possessive NPs, with P encoded as a possessive-like dependent. When A is also expressed, it is then introduced separately with oblique agent marking. NOM-ACC+ERG-POSS is therefore a natural and well-attested combination, whereas ERG-ABS+POSS-ACC is substantially less frequent.

3.2.6 Word Order and ANC Strategy

ANC strategies are not evenly distributed across different clause and NP word-order profiles. The distribution in the sample of Koptjevskaja-Tamm [1993] is shown in Table 3.13.

The association between SENT and SOV/GN is the clearest: almost all SENT languages have SOV/GN order. This distribution also overlaps with the previously discussed correlation between SOV/GN order and complement deranking, so SOV/GN+Deranking+SENT is supported by several converging typological tendencies. POSS-ACC is found mainly with

SOV/GN, V-initial/NG, and SVO/NG, with no clear instances in SOV/NG or SVO/GN [Koptjevskaja-Tamm, 1993].

ERG-POSS has the broadest word-order distribution and is attested in all five profiles shown in the table, with particularly high frequencies in V-initial/NG and SVO/NG. NOMN is found mainly in SVO/GN, with fewer cases in SOV/GN, V-initial/NG, and SVO/NG [Koptjevskaja-Tamm, 1993].

Overall, SOV/GN is strongly associated with SENT and POSS-ACC, SVO/NG with POSS-ACC and ERG-POSS, and SVO/GN most clearly with NOMN. ERG-POSS is compatible with the widest range of word-order profiles. However, the reasons behind these distributions remain insufficiently understood.

3.2.7 Derived ANC Word Order

ANC word order is not sampled as an independent parameter in this study, but is derived from clause word order, NP word order, and ANC strategy. This design reflects the typological observation that ANC word order is influenced by both clause structure and noun phrase structure. In particular, when finite clauses and ordinary noun phrases differ in head–dependent order, ANCs often follow the noun phrase pattern. Hixkaryana provides one such example: S follows the predicate in finite clauses, while the possessor precedes the noun in ordinary possessive NPs. Accordingly, an S encoded as a possessive-like argument in an ANC follows NP order rather than clause order and precedes the nominalized predicate [Koptjevskaja-Tamm, 1993, Ruditsky, 2024].

Under some parameter combinations, these derivation rules produce different relative orders for S and A across intransitive and transitive ANCs. For example, SVO/GN + ERG-POSS yields SV/PVA, while SOV/NG + ERG-POSS yields VS/AVP; in both cases, S and A occur on opposite sides of the nominalized predicate. Comparable splits between transitive and intransitive word order are nevertheless attested in human languages. In Muna, for example, transitive clauses have SVO order, whereas intransitive clauses generally have VS order, likewise placing A before the predicate and S after it [van den Berg, 1989, Dryer,

2013c]. I therefore treat these ANC profiles as derived patterns for which direct typological evidence is limited, rather than as grounds for a negative typological prediction.

3.3 Typological Predictions from Prior LM Experiments

This section summarizes prior findings on typological biases in LMs. Existing research has focused mainly on word-order parameters. Ravfogel et al. [2019] found that RNNs performed better on subject–verb agreement prediction under SVO order than under SOV order, attributing this difference to a recency advantage from the shorter subject–verb distance. White and Cotterell [2021], by contrast, found that LSTMs showed little consistent preference across word orders, while Transformers exhibited clearer preferences that did not align with the typological distribution of human languages.

Evidence for typological harmony among word-order parameters is more consistent. Kuribayashi et al. [2024] found that models with syntactic biases, particular parsing strategies, or memory limitations generally assigned lower perplexity to typologically common word-order combinations. Similarly, Xu et al. [2026] found that models learned languages that departed from common word-order correlations more slowly, although some differences diminished later in training.

Beyond word order, prior work has also examined case marking, but mainly as an overt cue to argument relations rather than as a comparison between alignment systems. Ravfogel et al. [2019] found that overt case marking significantly improved agreement prediction under both SVO and SOV order, suggesting that overt cues to argument roles can reduce the difficulty of learning syntactic dependencies. Deng et al. [2026] studied model preferences in differential argument marking and found that models do not reproduce the strong typological preference for systems in which overt marking more often targets objects than subjects.

Existing LM typological-bias experiments have not directly examined clausal complements and ANCs. This thesis therefore tests whether the word-order harmony effects previously reported can be reproduced, while extending the study of LM typological biases to more complex morphosyntactic combinations involving complement structure and ANCs.

Chapter 4
CORPUS GENERATION

This chapter describes how the abstract grammar space defined in Chapter 3 was converted into parallel synthetic corpora. The pipeline takes natural English as a source of predicate-argument content, maps it into a shared MRS bank, and then realizes the same meanings through 96 typologically distinct HPSG grammars.

4.1 Pipeline Overview

Figure 4.1 shows the corpus generation pipeline used in this thesis. The pipeline begins with raw English, which is automatically parsed using spaCy [Honnibal et al., 2020]. Based on these parsed records, the experiment derives two resources: pseudo-English and a shared lexicon.

Pseudo-English is a controlled input representation used for later semantic parsing. It keeps the core predicate-argument relations from the original sentences, while removing additional surface phenomena such as adjectival modification and auxiliary structure, making it easier for the pseudo-English grammar to parse.

The shared lexicon contains the lexical items needed by the later grammars, including nouns, different classes of verbs, and action-nominal candidates. For action nominal constructions, nominal suffixes and stem similarity between nouns and verbs are used to identify possible nominalization sources, so that these lexical items can be treated as nominalizable predicates in the grammars.

On the grammar side, the experiment starts from the typological parameters defined in Chapter 3. These parameters are encoded as Grammar Matrix choices files. For the 96 target languages, each file corresponds to one parameter combination and specifies the full

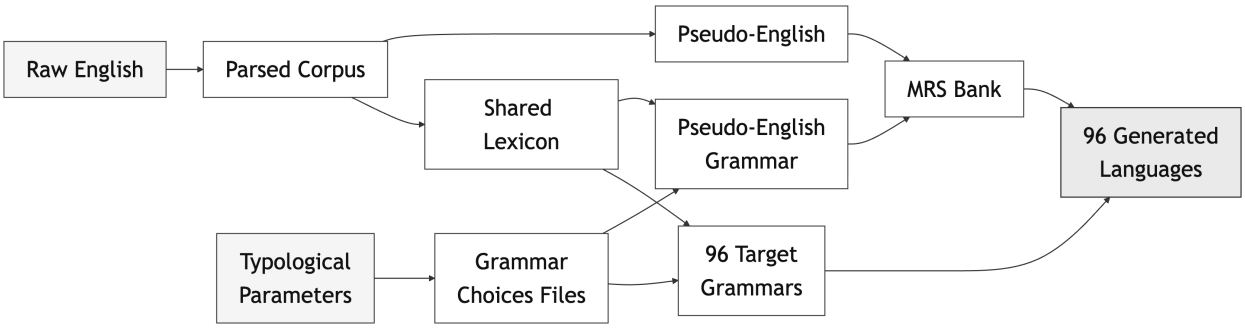

Figure 4.1: Overview of the corpus-generation pipeline.

grammar settings for that language. The pipeline also includes a separate pseudo-English choices file, which is used to build the pseudo-English grammar.

The shared lexicon is then inserted into both the pseudo-English grammar and the 96 target grammars, ensuring that they use the same lexical inventory. The pseudo-English corpus is parsed with the updated pseudo-English grammar to produce the MRS bank. MRS serves as the shared semantic interface across grammars.

Finally, the same set of MRS inputs is passed to the 96 target grammars. Each grammar realizes these inputs according to its own typological settings, producing one generated corpus. The final output is therefore a set of 96 synthetic corpora that share semantic content and lexical source material, but differ systematically in their surface morphosyntactic form. The following sections describe each part of the pipeline in more detail.

4.2 English Parsing and Pseudo-English

In the English parsing stage, the pipeline uses spaCy dependency parses [Honnibal et al., 2020] to extract the predicate–argument structures that can be expressed by the current grammar fragment. I do not directly parse the raw English corpus with an ACE-based HPSG grammar such as the English Resource Grammar [Flickinger et al., 2014], because such grammars provide rich and fine-grained semantic analyses of English, which would increase the complexity of aligning the resulting MRS representations with the target grammars used

in this experiment. Using dependency parses at this stage also allows the pipeline to filter out structures that are not modeled, including passives, adjectival predication, and clauses with multiple object candidates.

The experiment retains four constructions: intransitive clauses, transitive clauses, clausal-complement constructions, and nominal-predicate copular clauses. For each retained sentence, the system records the predicate and its core arguments. For clausal complements, the system records the matrix predicate and recursively analyzes the embedded clause. Both *xcomp* and *ccomp* dependencies are included. In the current pipeline, an *xcomp* is treated as a subject-control structure, with the matrix subject supplying the subject of the embedded predicate.

The system also records three types of nominal modifiers for all nouns: possessives, *of*-phrases, and *by*-phrases. These records are used later in the treatment of action nominal constructions. They provide cues to the possible argument structure of an action nominal: possessives and *by*-phrases can correspond to agent-like arguments, while *of*-phrases can correspond to patient-like arguments.

Parsed English is then linearized as pseudo-English. Pseudo-English uses a fixed, English-like surface format to express the extracted predicate–argument structures. Intransitive clauses are realized as S–V, transitive clauses as A–V–P, and clausal-complement constructions are realized with an overt complementizer. PP objects and particles are incorporated into the predicate form. For example, *look for* is represented as *lookfor*. Further details on the pseudo-English grammar are given in Section 4.3.

At the same time, the system constructs the shared lexicon. The lexicon contains nouns, intransitive verbs (*iv*), transitive verbs (*tv*), clausal-complement verbs (*cv*), and action-nominal candidates. Predicates extracted from nominal-predicate copular clauses (*cop_n*) are incorporated into the transitive-verb class in the lexicon, since the Grammar Matrix configuration does not directly support nominal complements in copular constructions. Lexical items preserve forms extracted from the English corpus, retaining some of the lexical distribution and word-form regularities of the source corpus.

Action-nominal candidates are identified heuristically. The system first searches for nouns with common nominalizing suffixes, including *-ation*, *-tion*, *-sion*, *-ment*, *-ance*, *-ence*, *-ure*, *-al*, and *-ing*. The suffix is then removed, and the remaining stem is compared with verbs in the verb lexicon. A noun is treated as a possible action nominal only when the noun stem and a verb share a sufficiently similar prefix, with the threshold set to 0.70. Nouns that are not identified as action nominals can still enter the lexicon as ordinary nouns, but their *of*- and *by*-modifiers are not interpreted as ANC arguments.

The current system uses only *iv* and *tv* verbs as nominalization sources. This is because *cv* verbs must retain their clausal complement after nominalization. In the POSS-ACC strategy, this complement occupies an object-like position and requires the appropriate dependent marking. The current Grammar Matrix setup does not reliably assign the required case marking to such nominalized clausal complements. To preserve parallelism across the generated corpora, *cv* verbs are therefore excluded from the action-nominal source set.

Table 4.1 summarizes the main outputs of English parsing and lexicon construction. Table 4.2 gives examples of the pseudo-English realization.

4.3 Choices Files and Grammars

On the grammar-construction side, the pipeline converts the typological parameters defined in Chapter 3 into HPSG grammars that can be used for parsing and generation with ACE [Packard, 2018]. Each grammar is first specified by a Grammar Matrix choices file. A choices file is the input format used by the Grammar Matrix to record the main grammatical settings of a language, such as basic word order, case system, genitive order, clausal-complement system, nominalization strategy, and the relevant morphological rules.

For the 96 target grammars, the choices files are generated automatically from the parameter grid. Each choices file corresponds to one parameter combination, and its language ID records the parameter values for that combination. In addition to these explicit parameters, the script derives more specific grammar settings according to the rules defined in Chapter 3, including ANC-internal word order, verb form, and case marking. Grammatical dimensions

English Parsing		Shared Lexicon	
Category	Count	Category	Count
Parsed sentences	12,441,836	Nouns	225,939
Retained sentences	4,095,236	Intransitive verbs	12,252
iv constructions	883,158	Transitive verbs	35,312
tv constructions	1,812,451	Clausal-complement verbs	5,220
cv constructions	674,525	Action-nominal candidates	6,298
cop_n constructions	725,102		
ANC External Position		**ANC Internal Arguments**	
Category	Count	Category	Count
S	16,084	No argument	220,501
A	43,155	One S/A-like argument	24,712
P	156,671	One P-like argument	18,264
Copular complement	45,675	Two arguments	719
Genitive possessor	643		

Table 4.1: Summary of English parsing, shared-lexicon construction, and action-nominal distribution for the training split.

that are not part of the experiment are kept inactive.

The pipeline then injects the shared lexicon extracted from the source corpus in Section 4.2 into the grammars. After the Grammar Matrix generates an initial Type Description Language (TDL) grammar from a choices file, the pipeline replaces the lexical entries with the shared lexicon. Nouns are written as common-noun lexical entries, while intransitive verbs, transitive verbs, and clausal-complement verbs are written into the corresponding verbal lexical types. Predicates extracted from *cop_n* constructions are added to the transitive-verb class.

In addition to the 96 target grammars, the pipeline builds a pseudo-English grammar for parsing the pseudo-English corpus from the previous section into MRS. This grammar uses

Original sentence	Pseudo-English	Generated form
Mary is sleeping.	*Mary sleeps*	*Mary sleeps*
I eat an apple.	*I eats apple*	*I eats appleca*
He wants to eat an apple.	*He wants that he eats apple*	*He wants he eating appleca*
I think Mary sleeps.	*I thinks that mary sleeps*	*I thinks mary sleeping*
He is looking for a book.	*He lookfors book*	*He lookfors bookca*
Why don't you get up a collection?	*You getups collectnmz*	*You getups collectingca*
John's destruction of the city shocked everyone.	*Johnge destroynmz cityob shocks everyone*	*Johnge destroying cityca shocks everyoneca*

Table 4.2: Examples of Pseudo-English and target-language realization under the profile in Table 3.7.

a separate pseudo-English choices file, which is different from the target-language choices files generated from the typological parameter grid.

The pseudo-English grammar differs from the target grammars in several implementation details. It uses an overt complementizer *that*, allowing clausal-complement structures to be parsed more reliably. It also uses the dedicated nominalization marker *-nmz*. Since *-nmz* does not occur in ordinary English word forms, it prevents nominalized predicates in pseudo-English from being confused with common nouns. The target grammars instead use *-ing* as the nominalization suffix, so that nominalized forms in the generated corpora retain word-form cues similar to English nominalization morphology.

Finally, the TDL grammars generated by the Grammar Matrix are compiled into ACE grammar images. The compiled pseudo-English grammar is used to parse pseudo-English into MRS, while the target grammars are used to generate the 96 target-language corpora from MRS. The full set of choices files is included in the project repository.

4.4 Grammar Patching

The ANC library in the Grammar Matrix has a limitation in word-order control. In this system, ANC-internal word order must be selected as a single word-order profile, such as SOV or SVO [Ruditsky, 2024]. However, as discussed in Section 3.2.7, under some parameter combinations, the S and A arguments within ANCs may appear on opposite sides of the nominalized predicate. For example, SVO/GN + ERG-POSS yields SV/PVA, and SOV/NG + ERG-POSS yields VS/AVP. These split patterns cannot be naturally represented by a single Grammar Matrix ANC-WO setting.

To address this problem, the pipeline first sets a single ANC word-order value in the choices file and uses it to generate the initial TDL grammar. In split patterns, this setting is used only to determine the position of the possessive-like argument relative to the nominalized predicate. For example, in split configurations with GN order, it is set to SOV.

The pipeline then changes the `HEAD.ANC-WO` value of the three non-sentential transitive ANC lexical rules from `+` to `-`, allowing the non-possessive argument to use the ordinary complement-head path. As a result, the possessive-like argument is still ordered through the specifier path according to NP order, while the non-possessive argument is ordered through the ordinary complement-head path according to the finite-clause direction. This makes it possible for the generated grammars to realize the split orders defined in the experiment.

However, this treatment introduces overgeneration. Since the non-possessive argument of a transitive ANC is allowed to use the ordinary complement-head path, both A-first and P-first derivations may be available when both ANC-internal arguments occur on the same side of the nominalized predicate. As a result, a single MRS can generate multiple surface candidates that differ only in the relative order of A and P. For example, under an SOV, GN, NOM-ACC, Balancing, ERG-POSS configuration, one MRS generates the following two outputs:

1. `I herob sermonge tellingca hears`
2. `I sermonge herob tellingca hears`

For this reason, the pipeline applies post-generation selection. If the candidates differ only in the relative order of the A and P arguments, it selects the candidate that matches the language's effective ANC transitive order. In the example above, the parameter configuration gives the target ANC order APV. The markers identify *herob* as A and *sermonge* as P, so the system keeps the first output.

4.5 MRS Bank and Surface Selection

After pseudo-English has been generated, it is parsed with the pseudo-English grammar to produce the MRS bank. The MRS bank then serves as the shared semantic input to the 96 target grammars. Each target grammar realizes the same MRSs according to its own morphosyntactic settings, producing one target-language corpus. Table 4.2 shows the realizations produced by the example grammar profile in Table 3.7.

During MRS construction, the pipeline applies a small amount of MRS normalization. For example, the Grammar Matrix uses `COG-ST` to represent the discourse status of nominal referents: `uniq-id` means that the referent is uniquely identifiable, while `in-foc` means that the referent is in focus. The pipeline replaces these more specific subtypes with the more general value `cog-st` to keep the MRS compatible with the different ANC strategies during generation.

After generation, the pipeline gathers all outputs for the same input item and selects one final sentence. Overgeneration can arise at two stages in the pipeline: pseudo-English parsing can produce multiple MRSs for the same pseudo-English item, and target grammar generation can produce multiple surface candidates from the same MRS.

The first main source is ANC-internal order overgeneration, discussed in the previous section. The second comes from pseudo-English parsing. The pseudo-English grammar does not always produce a unique MRS for each input item. A common case is that the same stem occurs in both the *iv* and *tv* lexicons, so an action nominal with only one overt argument can receive two analyses. The argument may be analyzed as the S of an intransitive nominalization, or as the remaining A or P of a transitive nominalization, with the other

argument omitted. When these two MRSs are generated by a target grammar, they may surface with different argument markers. For example, under the ERG-POSS strategy:

```
Theirob suffering seems
Theirge suffering seems
```

Here, *theirob* corresponds to the A-like analysis, while *theirge* corresponds to the S-like analysis. In such cases, the pipeline keeps the S-like candidate by default, that is, the output with `ANC_S_MARK`. This choice follows the typological tendencies discussed in Chapter 3, where single-argument action nominals tend to use S-like marking.

In addition, a small number of overgeneration cases cannot be resolved by the rules above. These usually come from accidental attachment ambiguity. For example, the generator may produce both *the mention seems that the mention of the rejection sinks into the mind* and *the mention of the rejection seems that the mention sinks into the mind*, which differ in whether *of the rejection* attaches to the matrix nominal or the embedded nominal. For these cases, the pipeline chooses one output with a fixed random seed, so that the result remains reproducible.

Chapter 5
MODEL TRAINING

This chapter describes the language-model training setup. I train three independently initialized GPT-2-small models for each target language generated in Chapter 4, using different random seeds.

5.1 Training Corpus

Model training uses the 100M-word strict-track corpus from BabyLM 2026 [Choshen et al., 2026]. The original corpus is split into train, dev, and test sets with a 90/5/5 ratio. The train split is used to build the shared lexicon and is then passed through the corpus-generation pipeline described in Chapter 4, producing 96 target-language training corpora of approximately 4.1 million sentences each, with sentence counts differing by fewer than 40 across languages. The dev and test splits are processed using the lexicon and grammars built from the train split, and do not add new lexical items to the shared lexicon. Each model is trained only on the corpus for its corresponding target language. The dev set is used for evaluation during training. The test set is held out and is not used in the current experiment.

5.2 Tokenization

This experiment uses the GPT-2 byte-level BPE tokenizer [Radford et al., 2019]. In the target grammars, morphological markers are designed to attach directly to lexical stems. Because these stems are themselves derived from English, this design keeps the input closer to the surface conditions under which models learn raw English. The fixed tokenizer represents these novel surface forms as sequences of subword units. Inspection of the tokenizer output shows that the experimental morphological markers remain separable subword cues in forms

such as stem+-*ca*, stem+-*ge*, and stem+-*ing*, indicating that the attached markers remain available to the model as subword-level cues.

After tokenization, the 96 training corpora contain approximately 24.2–27.6 million tokens each, with a mean of approximately 25.4 million. Some of this variation appears to reflect how the GPT-2 tokenizer segments different surface forms. In ordinary English clauses, verbs do not typically occur at the beginning of the sentence, so verb forms more commonly occur with preceding whitespace in the distribution from which the GPT-2 BPE merges were learned. In the VOS languages in this experiment, however, verbs frequently appear at the beginning. The absence of preceding whitespace can therefore make some of these forms less familiar to the tokenizer and lead to finer subword segmentation, which may partly account for the higher token counts in the VOS corpora.

5.3 Training Configuration

The current experiment uses the GPT-2-small configuration [Radford et al., 2019]. Model parameters are randomly initialized before training, so all models are trained from scratch. For each target language, I train three models with different random seeds for multi-seed validation. The three runs use the same target-language corpus, tokenizer, architecture, and training hyperparameters, differing only in random initialization and data shuffling. Because the tokenized corpus sizes vary slightly across languages, the fixed training schedule corresponds to approximately 2.59–2.96 corpus-equivalent epochs. The training configuration is shown in Table 5.1.

Parameter	Value	Parameter	Value
Model configuration	GPT-2-small	Tokenizer	GPT-2 tokenizer
Initialization	Random	Seeds	42, 43, 44
Block size	32	Training steps	70,000
Train batch size	16	Eval batch size	16
Gradient accumulation	2	Learning rate	5×10^{-4}
Warmup steps	590	Weight decay	0.01
Logging steps	100	Evaluation steps	5,000
Save steps	5,000	Save total limit	1
Saved model	Final-step model		

Table 5.1: Training configuration for each target-language model.

Chapter 6

EVALUATION

I use targeted minimal-pair evaluations to test whether the models learned the grammatical phenomena encoded in their corresponding grammars. Based on the grammatical parameters defined in Chapter 3, I designed 34 evaluation phenomena in six groups, covering basic finite clauses, noun phrases, clausal complements, and action nominal constructions. For each phenomenon, I generate source sentences from large language model prompts and then use source templates and perturbation rules to construct minimal pairs. I use minimal-pair accuracy to measure how well each model learned the phenomenon.

6.1 Minimal-Pair Construction

For each phenomenon to be tested, I define two components: source templates and a perturbation rule. The source templates specify the structural environment in which the phenomenon must occur. For example, when testing the finite form of an intransitive verb, the source template is an intransitive clause containing only S and V. The perturbation rule specifies how a GOOD sentence should be modified to construct a corresponding BAD sentence for the target grammatical phenomenon. When a phenomenon uses multiple source templates or multiple perturbation variants, these variants are balanced whenever possible.

After defining the source templates, I use GPT-4o with a controlled prompt to generate natural English source sentences [OpenAI, 2024]. The prompt template is shown in Appendix A. These source sentences are then passed through the same semantic generation pipeline described in Chapter 4. They are first simplified into pseudo-English, then parsed into MRS, and finally generated by each of the 96 target grammars. Items are filtered out if they cannot be extracted, parsed, or generated, if they contain words outside the shared

lexicon, or if the generated output cannot be reliably matched to the target template.

BAD sentences are constructed from the GOOD sentences using the perturbation rule. Different types of phenomena use different perturbation strategies. For word-order phenomena, the perturbation swaps the linear order of the relevant constituents. For morphological phenomena, the perturbation replaces the suffix with one that is ungrammatical in that position. Special phenomena such as valency and ANC omission use dedicated perturbation rules.

Finally, for each phenomenon and each target language, I randomly select 100 GOOD–BAD minimal pairs from the filtered set as the evaluation set. Each pair is scored with the model trained on the corresponding grammar. If the model assigns a higher length-normalized log probability to the GOOD sentence, the pair is counted as correct. The minimal-pair accuracy is the proportion of pairs in which the model selects the GOOD sentence.

The following example shows a minimal pair constructed under the parameter configuration in Table 3.7. This configuration uses the POSS-ACC ANC strategy, so the P argument inside a transitive ANC should receive the accusative marker *-ca*. The BAD sentence changes the P marker to *-ge*. Additional examples for each phenomenon are provided in Appendix A. The full implementation of the templates and perturbation rules is available in the project repository.

(7) a. GOOD: *johnge destroying cityca shocks maryca*

b. BAD: *johnge destroying cityge shocks maryca*

6.2 Evaluation Phenomena

Table 6.1 lists the 34 targeted phenomena and their perturbation rules. Further details about the source templates, perturbation rules, and examples for each phenomenon are provided in Appendix A.

The phenomena fall into six sets. The first three sets test the models' basic syntactic

ID	Phenomenon	Perturbation
1.1	Intransitive finite verb form	replace finite V with nonfinite V
1.2	Intransitive S marker	add an overt marker to S
1.3	Intransitive S–V order	swap S and V
1.4	Transitive finite verb form	replace finite V with nonfinite V
1.5	Transitive A marker	replace A marking with P marking or genitive marking
1.6	Transitive P marker	replace P marking with A marking or genitive marking
1.7	Transitive A–V order	swap A and the VP chunk
1.8	Transitive V–P order	swap V and P
1.9	Intransitive V valency	replace V with a transitive verb
1.10	Transitive V valency	replace V with an intransitive verb
2.1	Intransitive genitive marker	replace or remove genitive marking
2.2	Transitive genitive marker	replace or remove genitive marking
2.3	Genitive possessor–head order	swap possessor and head noun
3.1	Clausal complement verb form	replace embedded V with the opposite complement form
3.2	Clausal complement S marker	add genitive–like marking to embedded S
3.3	Clausal complement A marker	replace embedded A marking with ANC A marking*
3.4	Clausal complement P marker	replace embedded P marking with ANC P marking*
3.5	Clausal complement–V order	swap matrix V and clausal complement
3.6	Clausal complement CV valency	replace matrix CV with a transitive verb
4.1	ANC verb form	replace ANC V with finite V
4.2	ANC S marker	switch ANC S marking between SENT and non-SENT marking
4.3	ANC A marker	replace ANC A marking with finite A marking*
4.4	ANC P marker	replace ANC P marking with finite P marking*
4.5	ANC S–V order	swap ANC-internal S and V
4.6	ANC A–V order	swap ANC-internal A and the VP chunk
4.7	ANC V–P order	swap ANC-internal V and P
5.1	ANC external S marker	add an overt marker to the external ANC head
5.2	ANC external A marker	replace external A marking with P or genitive marking
5.3	ANC external P marker	replace external P marking with A or genitive marking
5.4	ANC external genitive marker	replace or remove external genitive marking
6.1	ANC omitted S	replace bare IV ANC head with a finite IV predicate
6.2	ANC omitted A	replace A-omitting TV ANC with a finite TV predicate frame
6.3	ANC omitted P	replace P-omitting TV ANC with a finite TV predicate frame
6.4	ANC omitted A+P	replace bare TV ANC head with a finite TV predicate

Table 6.1: Phenomena tested in the targeted minimal-pair evaluation.

abilities, while the last three sets focus on ANC-related phenomena:

1. **Basic finite clauses**: Test whether the models learned the basic structure of finite clauses, including finite verb form, S/A/P marking, core word order, and verbal valency.

2. **Noun phrases**: Test genitive marking and possessor-head order.

3. **Clausal complements**: Test verb form in complement clauses, embedded argument marking, the relative order of the clausal complement and the matrix verb, and the valency of clausal-complement verbs.

4. **ANC internal**: Test case marking, word order, and verb form inside ANCs.

5. **ANC external**: Test whether an ANC receives the correct case or genitive marking when it functions as a matrix argument or genitive possessor.

6. **ANC omission**: Test whether argument omission can occur inside ANCs but not in main or subordinate clauses.

Most phenomena follow the minimal-pair construction pipeline described in Section 6.1. Beyond this, three additional points should be noted.

6.2.1 Valency Phenomena

For the valency phenomena, the perturbation replaces the target verb with another verb whose valency does not match the environment. I first construct verb pairs for these phenomena. For 1.9 and 1.10, I extract intransitive/transitive verb pairs from the `intransitive.jsonl` and `transitive.jsonl` paradigms in BLiMP [Warstadt et al., 2020]. For 3.6, I use the MegaAcceptability dataset [White and Rawlins, 2020] to construct pairs of clausal-complement verbs and ordinary transitive verbs. This dataset provides acceptability ratings for English verbs in clausal complement frames and transitive frames.

Next, I use prompts to generate new source sentences for the target verbs, and then use source templates to filter out items that do not match the target structure or contain words outside the shared lexicon. I then replace the target verb in the GOOD sentence with the mismatched verb to form the BAD sentence.

Since many English verbs can occur in multiple valency frames, a grammar may still parse the BAD sentence if the relevant verb–valency combination is attested in the training set and therefore recorded in the lexicon. I therefore use an English baseline model as a reference point to estimate how much valency information can be learned directly from the original English data. The baseline uses the same GPT-2-small architecture and training settings as the target-language models in Section 5.3, but is trained directly on the raw English training split. Because the raw English corpus is larger, the English baseline is trained for 450,000 steps, corresponding to approximately 3 epochs and thus a comparable amount of training to the target-language models. It is evaluated on the English versions of the same valency minimal pairs.

6.2.2 Case Marking between Clausal Complements and ANCs

In the clausal-complement and ANC-internal case marking tests, the perturbation is mainly constructed by replacing markers across constructions. The embedded argument marker in a clausal complement is replaced with the corresponding ANC marker, while the ANC-internal argument marker is replaced with the corresponding marker in a non-ANC clause. When the two markers are identical on the surface in a particular configuration, I instead use another marker that is ungrammatical in that position. These cases are marked with an asterisk in Table 6.1.

It is important to note that identical surface marking does not mean that the two constructions are identical. Under the Deranking + SENT strategy, the verb form and argument marking inside an ANC are the same as those in a clausal complement. However, the ANC as a whole can occur in argument position and can receive external case or genitive marking, whereas a clausal complement can only occur as the complement of a clausal-complement

verb.

6.2.3 BAD Sentence Parsability

In minimal-pair construction, the BAD sentence is created by an external perturbation rule and therefore should not be accepted by the grammar under the target structure. However, some BAD sentences can still receive an alternative analysis in the grammar. I record such cases as BAD parses. A BAD parse means that the perturbed surface string has some available analysis under the current grammar and lexicon. The BAD parse rate is the proportion of BAD sentences that can be parsed by the grammar for a given phenomenon and grammar.

For example, under SOV/GN + ERG-ABS, the GOOD sentence *Nurseca bandage touchs* is perturbed into *Nursege bandage touchs*. The BAD sentence can be reanalyzed as something like "the nurse's bandage touches": the original A is analyzed as a genitive possessor, while the original P becomes the intransitive S. Such reanalysis is sometimes semantically natural, but in other cases it is only syntactically available and semantically unnatural.

In this study, BAD parse rate is used as a diagnostic measure alongside minimal-pair accuracy. It identifies phenomena in which alternative grammatical analyses are available and provides additional context for interpreting the evaluation results. Appendix A discusses BAD parse patterns for the relevant phenomena and reports BAD parse rate alongside the accuracy heatmaps.

6.3 Evaluation Results

This section reports the results of the targeted minimal-pair evaluation across the 34 phenomena. For each phenomenon, Table 6.2 summarizes the three-seed mean accuracy, grammar-based BAD parse rate, and main weak configurations across the 96 grammars. Figures 6.1 and 6.2 show the accuracy patterns for all phenomena and for the ANC-related phenomena only. The full evaluation details and per-phenomenon heatmaps are provided in Appendix A. The multi-seed validation is provided in Appendix B.

ID	Phenomenon	Acc. (%)	BAD (%)	Weak configurations
1.1	Intransitive finite verb form	99.4	0.0	–
1.2	Intransitive S marker	87.7	3.3	SOV/NG + ERG-ABS
1.3	Intransitive S–V order	100.0	0.0	–
1.4	Transitive finite verb form	99.7	0.0	–
1.5	Transitive A marker	96.6	12.7	SOV/GN + NOM-ACC
1.6	Transitive P marker	97.9	11.2	SOV/NG + ERG-ABS
1.7	Transitive A–V order	100.0	0.0	–
1.8	Transitive V–P order	100.0	0.1	–
1.9	Intransitive V valency	70.4	50.0	Non-SVO
1.10	Transitive V valency	76.2	93.0	Non-SVO
2.1	Intransitive genitive marker	89.3	16.0	SOV/NG + NOM-ACC; SOV/GN + ERG-ABS
2.2	Transitive genitive marker	91.6	0.1	SOV/NG + ERG-ABS
2.3	Genitive possessor–head order	94.9	23.7	VOS/NG + NOM-ACC
3.1	Clausal complement verb form	98.6	8.5	–
3.2	Clausal complement S marker	75.2	37.3	SOV/NG + NOM-ACC
3.3	Clausal complement A marker	83.9	30.8	SOV/GN + NOM-ACC
3.4	Clausal complement P marker	97.7	15.6	SOV/NG + ERG-ABS
3.5	Clausal complement–V order	100.0	0.4	–
3.6	Clausal complement CV valency	74.9	76.4	All grammars
4.1	ANC verb form	91.9	21.3	Balancing + SENT
4.2	ANC S marker	82.1	25.9	SENT
4.3	ANC A marker	68.7	20.7	SENT + NOM-ACC; ERG-POSS + ERG-ABS
4.4	ANC P marker	90.1	31.8	POSS-ACC + ERG-ABS
4.5	ANC S–V order	90.3	28.9	Non-harmonic[1] + non-SENT; SVO + SENT + ERG-ABS
4.6	ANC A–V order	93.1	16.7	Non-harmonic + POSS-ACC/NOMN; SVO/GN + ERG-POSS
4.7	ANC V–P order	92.9	13.9	SVO + Deranking + ERG-ABS + non-SENT
5.1	ANC external S marker	62.9	7.1	All except SVO + NOM-ACC + non-SENT
5.2	ANC external A marker	82.7	11.6	Non-SVO + NOM-ACC
5.3	ANC external P marker	91.4	15.8	SOV/NG + ERG-ABS
5.4	ANC external genitive marker	75.9	4.2	SOV + ERG-ABS
6.1	ANC omitted S	98.5	11.0	–
6.2	ANC omitted A	75.8	16.1	Balancing
6.3	ANC omitted P	83.9	16.7	Balancing + ERG-ABS
6.4	ANC omitted A+P	90.5	25.8	Balancing + ERG-ABS

Table 6.2: Summary of targeted minimal-pair evaluation results across the 96 grammars. Accuracy and BAD parse rate are mean values across grammars.

Figure 6.1: Overall accuracy heatmap across all targeted minimal-pair phenomena.

Overall, the models learn most of the grammatical phenomena in the synthetic languages, with a mean accuracy of 88.4% across the 34 phenomena and 96 grammars. However, there are clear differences across phenomena. As Table 6.2 shows, verb-form and basic word-order phenomena are relatively easy, while argument-marking phenomena are harder overall. Comparing Figures 6.1 and 6.2, the ANC-related phenomena show broadly similar patterns across grammar configurations, but configurations that are already weak in the full

[1]Non-harmonic profiles are SOV/NG and VOS/GN, as defined in Section 3.2.1, where clause-level word order and NP-internal order have different head directions.

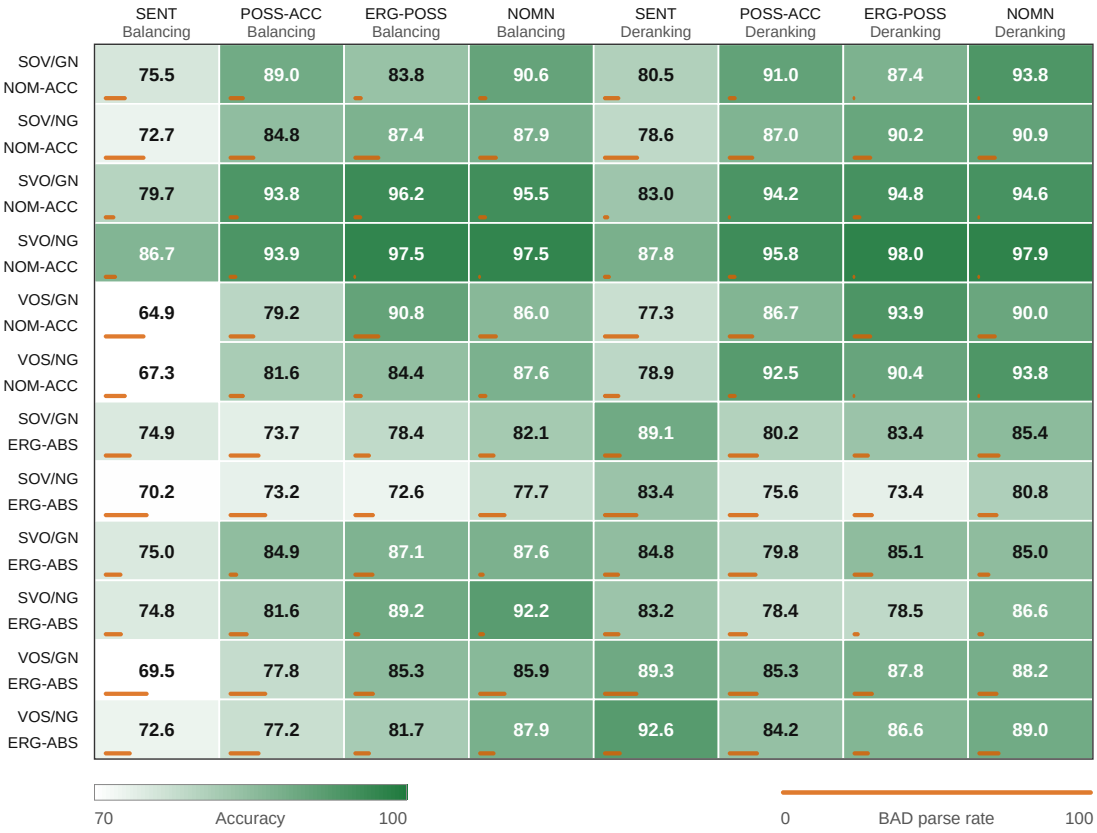

Figure 6.2: Accuracy heatmap for ANC-related phenomena only.

evaluation often become weaker in the ANC-related phenomena. Lexical valency is another difficult group, but the models outperform the corresponding English baselines on all three valency tests (IV: 70.4% vs. 56.0%; TV: 76.2% vs. 66.0%; CV: 74.9% vs. 68.0%), suggesting that the grammar-generated training data provide a clearer valency signal than the original English corpus.

Across grammar configurations, Figures 6.1 and 6.2 reveal systematic differences among parameter combinations. SVO configurations perform best overall and maintain relatively high accuracy across ANC strategies and complement systems. For alignment, NOM-ACC

performs better than ERG-ABS overall, but under SENT + Deranking, ERG-ABS instead has an advantage. Across ANC strategies, SENT performs worst overall, especially under Balancing, while the other three non-SENT strategies generally perform better.

Specifically, two recurring patterns stand out. First, the models generally perform better when related grammatical structures exhibit stable linear patterns. Second, overt morphological marking provides additional learning cues and helps the models distinguish grammatical structures with similar surface forms. The following two subsections analyze these patterns in detail. Finally, I examine the relationship between BAD parses and evaluation accuracy.

6.3.1 Linear Pattern Consistency

In the experiment, the models are highly sensitive to surface structure. They prefer stable linear cues, and more easily learn linear patterns that recur across construction types with stable argument–predicate relations. By contrast, learning becomes harder when the same surface position corresponds to different argument roles across structures.

The clearest pattern is the interaction between clause word order and alignment. Under SVO, NOM-ACC is much better than ERG-ABS, with mean accuracies of 94.1% and 89.3%, respectively. This is because, in SVO + NOM-ACC, finite intransitive and finite transitive clauses both begin with an unmarked NP followed by a finite verb. By contrast, in SVO + ERG-ABS, intransitives have the form *NP V*, while transitives have the form *NP-ca V NP*. The model must use the following verb and argument pattern to decide whether the initial NP is S or A, and therefore whether it should take *-ca*. Similarly, SOV also favors NOM-ACC, with mean accuracies of 87.0% for NOM-ACC and 84.6% for ERG-ABS. VOS shows the reverse pattern, with 89.0% for ERG-ABS and 86.2% for NOM-ACC.

The same pattern also appears in ANC-internal order. Its learnability is directly related to whether it preserves the corresponding order in finite clauses. In combinations of non-SENT strategies and non-harmonic profiles, ANCs are affected by both clause word order and NP word order. When the two point in different directions, ANC-internal order can diverge from the familiar argument–predicate pattern in finite clauses. These combinations

are especially weak in ANC S–V Order and ANC A–V Order, with mean accuracies of 73.5% and 81.4%, well below the overall means for these two phenomena, 90.3% and 93.1%.

Another related effect is that SVO separates the two core arguments in transitive clauses. Since the verb separates A and P, marker perturbations are less likely to create a local genitive-like sequence or a different argument span. This is visible in Clausal Complement A Marker in Section A.3.3, where SVO has a mean accuracy of 96.8%, compared with 71.9% for SOV and 83.0% for VOS. Similar local reanalysis effects also appear in Intransitive S Marker in Section A.1.2, Transitive A and P Marker in Sections A.1.5 and A.1.6, and Clausal Complement P Marker in Section A.3.4.

Finally, in lexical valency phenomena, SVO also performs better than the two non-SVO orders. SVO has a mean accuracy of 77.3%, higher than non-SVO at 72.1%. This advantage is clearest in Intransitive V Valency, where SVO reaches 76.6%, compared with 67.3% for non-SVO. This suggests that, in SVO, the verb occurs between A and P, which helps the models identify argument roles and verb frames.

6.3.2 Morphological Cues

Morphological cues are another major source of cross-phenomenon differences. The models show a clear asymmetry between overt and zero argument marking, with overtly marked roles usually easier to learn and zero marking less stable. This pattern appears across several phenomena. In Transitive A Marker in Section A.1.5, the overt A marker *-ca* under ERG-ABS reaches 100.0%, while zero-marked A under NOM-ACC has a mean accuracy of 93.2%. Transitive P Marker in Section A.1.6 shows the mirror pattern: the overt P marker *-ca* under NOM-ACC reaches 100.0%, while zero-marked P under ERG-ABS has a mean accuracy of 95.8%. The same asymmetry is stronger in clausal complement and ANC external marking. In Clausal Complement A Marker in Section A.3.3, ERG-ABS reaches 99.7%, while NOM-ACC reaches only 68.0%; in ANC External A Marker in Section A.5.2, ERG-ABS reaches 99.0%, while NOM-ACC reaches only 66.4%. This suggests that the models learn where overt markers appear more reliably than where markers must be absent.

Morphological cues also affect the strength of constructional boundaries. SENT reuses finite-clause marking to the greatest extent, so the boundary between ANCs and finite clauses or clausal complements is the weakest. This corresponds to its lowest overall ANC-related accuracy, at 78.9%. POSS-ACC partially reuses finite P marking and performs slightly better, with a mean accuracy of 84.2%. By contrast, ERG-POSS and NOMN provide more distinctive ANC-internal argument marking, and therefore perform somewhat better overall, with mean accuracies of 86.8% and 88.9%. However, ERG-POSS and NOMN do not show different alignment patterns. Both perform better under NOM-ACC than under ERG-ABS: ERG-POSS has 91.2% versus 82.4%, and NOMN has 92.2% versus 85.7%.

The complement system shows a similar mechanism. Deranking uses nonfinite *V-ing* to distinguish ANC and complement predicates from finite predicates, and is therefore especially helpful in ANC omission. In ANC omission, Deranking has a mean accuracy of 94.6%, much higher than Balancing at 79.7%.

6.3.3 BAD Parses and Accuracy

In the experiment, BAD parse rate does not directly determine accuracy. Some phenomena with high BAD parse rates still have high accuracy, while some low-accuracy cases have few or no BAD parses. The following examples illustrate this mismatch.

First, a high BAD parse rate does not necessarily lead to low accuracy. For example, in Clausal Complement S Marker in Section A.3.2, under SOV/GN + ERG-ABS + Deranking + POSS-ACC, the BAD parse rate reaches 100.0%, but the accuracy is still 82.3%. In this configuration, the BAD sentence can be reanalyzed by treating the original clausal complement as an ANC in matrix P position, roughly like "He thinks [the guest's leaving]." This alternative parse is available partly because of CV/TV overlap in the shared lexicon, since 4,140 of the 5,220 CV verbs also appear in the TV class. Nevertheless, the model still usually prefers the GOOD sentence, showing that it can use distributional differences between verbs and complement types.

Second, when high BAD parse rate and low accuracy occur together, the alternative

analysis usually competes strongly with the target structure. This should not be treated simply as a failure to learn the grammar. For example, in ANC P Marker in Section A.4.4, POSS-ACC + ERG-ABS has a mean accuracy of only 64.3%, but the BAD parse rate reaches 83.5%. In these configurations, incorrectly adding *-ge* can cause ANC-internal arguments to be reanalyzed as a genitive-like nominal structure, for example from "the reporter notes [the scientist's measuring of the sample]" to "the reporter notes [the measuring of the scientist's sample]." The BAD sentence is therefore competitive both grammatically and on the surface. This shows that, in some configurations, the boundary inside ANCs is weak enough that a local perturbation can create a plausible alternative analysis.

Finally, low BAD parse rate can also occur with low accuracy. This shows that even when the BAD sentence cannot be parsed by the grammar, the model may still prefer it. For example, in Clausal Complement A Marker in Section A.3.3, SOV/GN + NOM-ACC + Balancing + POSS-ACC has an accuracy of only 13.3%, but a BAD parse rate of 0.0%. In this configuration, the GOOD sentence *Lawyer witness attackca sees claims* roughly corresponds to "the lawyer claims that the witness sees the attack." The perturbed BAD sentence is *Lawyer witnessge attackca sees claims*. This BAD sentence cannot receive a valid parse, but *witnessge attackca* is locally close to a marked P with a possessor. As a result, the beginning of the BAD sentence is similar to a common A + P sequence in this language. The model may assign a high score to the BAD sentence because of this local surface pattern, even though the whole sentence is not parsable by the grammar.

Taken together, BAD parses help explain some difficult conditions, but they do not account for the overall learnability patterns. Rather, the contrast between parsability and model preference further shows that the models are sensitive both to grammatical distributions in the training data and to local surface cues.

Chapter 7
DISCUSSION

In Section 6.3, I reported the overall experimental results. The models perform better when linear structure is stable, argument roles have overt cues, and ANCs are morphologically easier to distinguish from finite clauses. They perform worse when the target contrast depends on zero marking, weak constructional boundaries, or local surface sequences that support alternative analyses. They also show some advantage for SVO over SOV. This is consistent with the general findings of previous work on LM learning, which shows that model learning difficulty is systematically affected by surface structure, word-order correlations, and overt cues to argument roles [Ravfogel et al., 2019, White and Cotterell, 2021, Kuribayashi et al., 2024, Xu et al., 2026].

Beyond these overall patterns, comparison with the human typological predictions in Chapter 3 highlights three parameter interactions that require further discussion: clause word order and alignment, word order and complement system, and complement system and ANC strategy. Across these interactions, the model results align closely with human typology in patterns involving stable linear organization, but diverge in the preference for additional morphological cues.

7.1 Clause Word Order and Alignment

The experiment shows that the model's preference for alignment depends on clause word order. The strongest contrast appears under SVO. In SVO grammars, NOM-ACC is substantially easier than ERG-ABS, with mean accuracies of 94.1% and 89.3%, respectively. The other two clause word orders show the same type of interaction, but the differences are smaller. SOV favors NOM-ACC, with 87.0% for NOM-ACC and 84.6% for ERG-ABS,

whereas VOS favors ERG-ABS, with 89.0% for ERG-ABS and 86.2% for NOM-ACC.

The model's preference for NOM-ACC under SVO is consistent with human language typology [Trask, 1979, Siewierska, 1996]. As shown in Table 3.10, among SVO languages with overt alignment marking, NOM-ACC systems are much more frequent than ERG-ABS systems. O'Grady [2022] uses Sidedness Harmony to explain this pattern. In accusative systems, S and A tend to occur on the same side of the verb. In ergative systems, S and P tend to occur on the same side of the verb. SVO is therefore harmonic for NOM-ACC, because both S and A precede the verb. By contrast, SVO is not harmonic for ERG-ABS, because S precedes the verb while P follows it.

In this thesis, I propose a more general surface-overlap account to explain why the model prefers different alignments under different clause word orders. For each word-order/alignment combination, I define overlap as the number of morphologically identical positions shared by the finite intransitive frame and the finite transitive frame, counted from the beginning of the clause. Two positions count as matching only when they have the same morphological form, including NP case marking and verb form. Table 7.1 shows the IV/TV frames, overlap scores, model accuracies, and human language proportions for the six combinations.

The overlap difference between the two alignments is largest under SVO, with a difference of 2. Correspondingly, the model learns NOM-ACC much better than ERG-ABS under SVO. Under SOV and VOS, the overlap difference between the two alignments is only 1, and the model's alignment preference is weaker. This suggests that surface overlap may be associated with model learnability. When finite intransitive and transitive clauses share stable surface cues earlier in the clause, the model learns the corresponding word-order/alignment combination more easily.

Compared with previous accounts, this surface-overlap account provides a more general perspective. It captures not only why SVO disfavors ERG-ABS, but also the alignment preferences under SOV and VOS, as well as the differences in preference strength. The model results support this account. They suggest that stable initial surface cues may be one

Clause WO	Alignment	IV frame	TV frame	Overlap	Model acc.	Human proportion
SVO	NOM-ACC	*NP V-s*	*NP V-s NP-ca*	2	94.1	90.0%
SVO	ERG-ABS	*NP V-s*	*NP-ca V-s NP*	0	89.3	10.0%
SOV	NOM-ACC	*NP V-s*	*NP NP-ca V-s*	1	87.0	64.4%
SOV	ERG-ABS	*NP V-s*	*NP-ca NP V-s*	0	84.6	35.6%
VOS	NOM-ACC	*V-s NP*	*V-s NP-ca NP*	1	86.2	33.3%
VOS	ERG-ABS	*V-s NP*	*V-s NP NP-ca*	2	89.0	66.7%

Table 7.1: Surface overlap between finite intransitive and transitive frames under different clause word order and alignment combinations. The human language proportions exclude neutral systems.

factor shaping language learnability for both models and humans.

7.2 Word Order and Complement System

The experimental results show that the learnability of complement systems depends on word-order profile. Table 7.2 shows the mean accuracy on ANC-related phenomena for the two SVO profiles and the two harmonic non-SVO profiles under the two complement systems. In the SVO profiles, the two complement systems are close. By contrast, in the harmonic non-SVO profiles, Deranking is clearly easier than Balancing.

This model pattern is only partly consistent with human typology. In the sample of Koptjevskaja-Tamm [1993] in Table 3.11, Deranking strongly clusters with SOV/GN order, with 24 of the 28 deranking languages having SOV/GN order. Balancing languages, by contrast, are distributed across a wider range of word-order profiles. The model also finds SOV/GN + Deranking easy to learn, which is consistent with the human distribution. However, in the model results, Deranking is not limited to SOV/GN, and SVO + Deranking is not especially difficult either. Conversely, the model treats non-SVO + Balancing as relatively harder, although this combination is not rare in human languages.

Complement system	SVO/GN	SVO/NG	SOV/GN	VOS/NG
Balancing	87.5	89.2	81.0	80.0
Deranking	87.7	88.3	86.4	88.5

Table 7.2: Mean ANC-related accuracy by word-order profile and complement system.

This discrepancy has two aspects. First, SOV + Balancing, which is well attested in human languages, performs relatively weakly in the model. This is possibly because the experimental grammars in this thesis do not include complementizers. One core function of complementizers is to mark a clause as the complement of a matrix predicate [Noonan, 2007]. This cue may be especially important for Balancing systems, where the embedded predicate has the same verb form as predicates in independent clauses [Cristofaro, 2003]. Without an additional complement-clause marker, non-SVO Balancing configurations must rely mainly on word order and case marking to distinguish embedded clauses, matrix arguments, and ANCs. This may be more difficult in non-SVO orders such as SOV and VOS, because the arguments of the embedded clause and the matrix clause are more likely to be linearly adjacent, creating sequences of adjacent NPs that are easier to confuse. Complementizers or other clause-boundary markers in real languages may help reduce this ambiguity.

Second, and more importantly, the model learns a combination that is rare in human languages: SVO + Deranking. As shown in Section 6.3.2, Deranking provides an explicit cue to embedded predicates. It marks the embedded predicate with nonfinite *V-ing*, making complement clauses and ANCs easier to distinguish from finite matrix clauses. For the model, this additional cue is usually helpful.

However, typological distributions in human languages do not depend only on whether a structure is learnable. They also depend on whether the structure provides enough additional functional benefit. Previous work suggests that different cues in a language system can compensate for one another, and that when one cue is sufficiently reliable, additional overt

coding may provide less communicative benefit [Fedzechkina and Jaeger, 2020, Haspelmath, 2021, Levshina, 2021]. In SVO order, the linear relations among the matrix verb, embedded predicate, and core arguments are already relatively stable, so Balancing may already be sufficient to distinguish matrix clauses from embedded clauses. As a result, SVO languages may not face the same pressure to use Deranking. Therefore, although SVO + Deranking is learnable for the model, it is not common in human languages.

7.3 Complement System and ANC Strategy

The interaction between complement system and ANC strategy is very clear in human languages. As shown in the sample of Koptjevskaja-Tamm [1993] in Table 3.12, Deranking languages mainly use more clause-like ANC strategies, especially SENT and POSS-ACC, whereas Balancing languages mainly use more nominal-like ANC strategies, including POSS-ACC, ERG-POSS, and NOMN. Conversely, Deranking + NOMN and Balancing + SENT are rare.

The model results reproduce this pattern only partially. Table 7.3 shows that Balancing + SENT is indeed the weakest combination, with a mean accuracy of only 73.7%. Deranking substantially improves the learnability of SENT, raising its accuracy to 84.0%. This is consistent with the fact that SENT more often appears in Deranking systems in human languages. However, the model does not treat Deranking and NOMN as incompatible. Instead, NOMN performs well under both complement systems, reaching 88.2% under Balancing and 89.7% under Deranking. This shows that, for the model, Deranking + NOMN is not a difficult combination.

The reason is similar to the explanation in Section 7.2. Balancing and SENT are closer to the original argument–verb frame, whereas Deranking and NOMN both provide additional explicit cues. Deranking distinguishes embedded predicates from finite predicates, while NOMN distinguishes ANC arguments from finite-clause arguments. For the model, combining Deranking and NOMN therefore does not create learning difficulty; instead, it provides more surface cues.

Complement system	SENT	POSS-ACC	ERG-POSS	NOMN
Balancing	73.7	82.6	86.2	88.2
Deranking	84.0	85.9	87.5	89.7

Table 7.3: Mean ANC-related accuracy by complement system and ANC strategy.

For human languages, however, if Deranking plus a SENT-like ANC is already sufficient to distinguish ANCs from finite clauses, there may be little motivation to further turn the ANC into a more NOMN-like structure. By contrast, in Balancing systems, main clauses and embedded clauses are already similar, so there is stronger pressure to mark ANCs in a more clearly distinct way. In this sense, Deranking + NOMN is learnable for the model, but may lack an independent functional motivation in human languages.

7.4 Summary

Overall, these three parameter interactions reveal two different patterns in learnability. First, the models generally learn related grammatical structures more easily when they maintain stable linear organization, and this preference is relatively consistent with several patterns in human language typology. Second, when existing surface cues are insufficient to distinguish related structures, additional overt morphological marking can provide more reliable distinguishing cues and improve model learnability.

However, human languages do not optimize learnability alone. Previous work has treated linguistic systems as shaped by multiple competing pressures rather than by a single objective. Many typological distributions have been analyzed as trade-offs between simplicity and informativeness, in which linguistic systems must remain sufficiently simple while also preserving useful distinctions [Kemp et al., 2018, Steinert-Threlkeld and Szymanik, 2020, Steinert-Threlkeld, 2021, Imel et al., 2026]. The learning process can create pressure toward simpler systems, while communicative interaction creates pressure to preserve informative

distinctions [Carr et al., 2020].

In grammatical coding, additional cues are usually helpful for model learning. Deranking, NOMN, and other explicit distinguishing devices can make the target structure easier to identify, and stacking multiple cues does not necessarily create learning difficulty. For human languages, however, if existing word order, case marking, or constructional patterns are already sufficient for the relevant function, a new marking strategy may lack independent functional benefit and therefore have weaker pressure to spread [Haspelmath, 2021, Fedzechkina and Jaeger, 2020, Levshina, 2021]. As a result, some configurations can be learnable, or even easy to learn, for the model while still being rare in human languages.

Chapter 8

CONCLUSION

This thesis proposes a grammar-engineered method for constructing synthetic corpora and uses it to study language models' learning preferences in acquiring action nominal constructions (ANCs) and related phenomena. I first use the Grammar Matrix to define a set of choices files and generate target grammars that vary systematically along five dimensions: basic clause word order, noun phrase-internal word order, alignment system, complement system, and argument-coding strategy in ANCs. I then extract shared MRS representations from an English source corpus and realize them through different target grammars to produce training corpora for different synthetic languages. This method combines the advantages of two previous approaches to synthetic-language experiments. On the one hand, unlike simple perturbations of natural corpora, it uses executable grammars to control deeper typological parameters rather than only changing surface order. On the other hand, it preserves a relatively natural semantic and lexical distribution by relying on shared semantic input.

I then train three independently initialized GPT-2-small models for each target grammar and evaluate different grammatical phenomena using minimal pairs. The evaluation covers basic word order, noun-phrase structure, argument marking, lexical valency, clausal complements, and ANC-related phenomena. The results show that the models perform robustly on many of these phenomena, but they still show greater difficulty with zero-marking phenomena and non-harmonic word-order profiles, and they are easily influenced by local surface heuristics. These results are consistent with previous research showing that language models prefer stable surface-linear regularities and local structural patterns, suggesting that this method can produce effective corpora for synthetic-corpus experiments on language-model learning. In addition, in the valency-related phenomena, the target-language models out-

perform the English baseline, suggesting that the generated corpora do preserve part of the semantic distribution associated with predicates, arguments, and event structure.

At the same time, this experiment reveals two more general patterns in model learnability. First, the models generally learn related grammatical structures more easily when they maintain stable linear organization. This preference is relatively consistent with several patterns in human language typology. For example, in the interaction between clause word order and alignment, both the model results and human language distributions show a relative preference for SVO with nominative–accusative alignment. These results suggest that stable and recurring linear patterns across related structures may be one source of the correspondence between model learnability and some human typological tendencies.

Second, when existing surface patterns are insufficient to distinguish related structures, additional overt morphological cues generally improve model learnability. Deranking, non-SENT ANC strategies, and other distinguishing devices can make the target structure easier for the models to identify, while stacking multiple reliable cues does not necessarily create additional learning difficulty. However, human typological distributions do not simply maximize such cues. Human language systems are also shaped by multiple competing pressures, such as the trade-off between coding economy and the functional benefit provided by additional marking. If a structure can already be sufficiently distinguished by existing word order, case marking, or constructional patterns, additional overt marking may not provide enough functional benefit, and therefore may not spread in human languages. For the model, by contrast, such additional cues usually reduce learning difficulty. As a result, some configurations can be learnable for the model while still being rare in human languages.

In conclusion, this thesis proposes and demonstrates a grammar-engineered synthetic-language method for studying the inductive biases of language models. By comparing the learnability of different grammatical systems in a controlled grammatical space, this method not only helps identify which structures are easier or harder for models to learn, but also makes it possible to compare where these learning preferences align with or diverge from human language typology. The results suggest that model learnability can reflect some of

the pressures underlying human typological distributions, such as a preference for stable linear organization. At the same time, divergence between model preferences and human typology shows that learnability alone is not sufficient to explain the distribution of grammatical systems in human languages, which also reflects other functional and communicative pressures. This comparison therefore provides information about both models and language. It reveals the surface regularities that current language models rely on during learning. It also helps identify which human typological patterns may arise partly from learnability and which require additional factors to explain.

BIBLIOGRAPHY

Artemis Alexiadou. *Functional Structure in Nominals: Nominalization and Ergativity*. John Benjamins Publishing Company, 2001. ISBN 9789027298256. doi: 10.1075/la.42. URL https://doi.org/10.1075/la.42.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *COLING-02: Grammar Engineering and Evaluation*, 2002. URL https://aclanthology.org/W02-1502/.

Emily M. Bender, Scott Drellishak, Antske Fokkens, Michael Wayne Goodman, Daniel P. Mills, Laurie Poulson, and Safiyyah Saleem. Grammar prototyping and testing with the LinGO Grammar Matrix customization system. In Sandra Kübler, editor, *Proceedings of the ACL 2010 System Demonstrations*, pages 1–6, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-4001/.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7096–7116, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.576. URL https://aclanthology.org/2020.emnlp-main.576/.

Jon W. Carr, Kenny Smith, Jennifer Culbertson, and Simon Kirby. Simplicity and informativeness in semantic category systems. *Cognition*, 202:104289, 2020. ISSN 0010-0277. doi: 10.1016/j.cognition.2020.104289. URL https://www.sciencedirect.com/science/article/pii/S0010027720301086.

Noam Chomsky. *Rules and Representations*. Woodbridge Lectures Delivered at Columbia University. Columbia University Press, 1980. ISBN 9780231048279.

Noam Chomsky. *Knowledge of Language: Its Nature, Origin, and Use*. Convergence. Praeger, 1986. ISBN 9780030055539.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. The false promise of ChatGPT. *The New York Times*, March 2023. URL `https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html`.

Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Jaap Jumelet, Tal Linzen, Aaron Mueller, Suchir Salhan, Raj Sanjay Shah, Alex Warstadt, and Ethan Gotlieb Wilcox. BabyLM turns 4 and goes multilingual: Call for papers for the 2026 BabyLM workshop, 2026. URL `https://arxiv.org/abs/2602.20092`.

Morten H. Christiansen and Nick Chater. Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489–509, 2008. doi: 10.1017/S0140525X08004998.

Bernard Comrie. *Language Universals and Linguistic Typology: Syntax and Morphology*. University of Chicago Press, 1989. ISBN 9780226114330.

Bernard Comrie. Alignment of case marking of full noun phrases. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/chapter/98`.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3:281–332, 2005. doi: 10.1007/s11168-006-6327-9.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. Are all languages equally hard to language-model? In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceed-*

ings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 536–541, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2085. URL https://aclanthology.org/N18-2085/.

Sonia Cristofaro. *Subordination*. Oxford Studies in Typology and Linguistic Theory. Oxford University Press, Oxford, 2003.

William Croft. *Typology and Universals*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, 2nd edition, 2002.

Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. Learning biases predict a word order universal. *Cognition*, 122(3):306–329, 2012. doi: 10.1016/j.cognition.2011.10.017.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural networks and the Chomsky Hierarchy. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WbxHAzkeQcn.

Iskar Deng, Nathalia Xu, and Shane Steinert-Threlkeld. Differences in typological alignment in language models' treatment of differential argument marking. In *Proceedings of the 30th Conference on Computational Natural Language Learning*, pages 268–283, July 2026. doi: 10.18653/v1/2026.conll-main.16. URL https://aclanthology.org/2026.conll-main.16/.

Matthew S. Dryer. The Greenbergian word order correlations. *Language*, 68:81–138, 1992. doi: 10.2307/416370.

Matthew S. Dryer. Why statistical universals are better than absolute universals. In *Papers from the 33rd Regional Meeting of the Chicago Linguistic Society: The Panels*, pages 123–145. Chicago Linguistic Society, 1998.

Matthew S. Dryer. Order of genitive and noun. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013a. URL https://wals.info/chapter/86.

Matthew S. Dryer. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013b. URL https://wals.info/chapter/81.

Matthew S. Dryer. Order of subject and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013c. URL https://wals.info/chapter/82.

Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online (v2020.4)*. Zenodo, 2013. doi: 10.5281/zenodo.13950591. URL https://doi.org/10.5281/zenodo.13950591.

Michael Dunn, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345): 79–82, 2011. doi: 10.1038/nature09923. URL https://doi.org/10.1038/nature09923.

Javid Ebrahimi, Dhruv Gelda, and Wei Zhang. How can self-attention networks recognize Dyck-n languages? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4301–4306, November 2020. doi: 10.18653/v1/2020.findings-emnlp.384. URL https://aclanthology.org/2020.findings-emnlp.384/.

Nadine El-Naggar, Tatsuki Kuribayashi, and Ted Briscoe. Which word orders facilitate length generalization in LMs? An investigation with GCG-based artificial languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35599–35613, November 2025. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1803. URL https://aclanthology.org/2025.emnlp-main.1803/.

Maryia Fedzechkina, T. Florian Jaeger, and Elissa L. Newport. Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences of the United States of America*, 109(44):17897–17902, 2012. doi: 10.1073/pnas.1215776109.

Masha Fedzechkina and T. Florian Jaeger. Production efficiency can cause grammatical change: Learners deviate from the input to better balance efficiency against robust message transmission. *Cognition*, 196:104115, 2020. ISSN 0010-0277. doi: 10.1016/j.cognition.2019.104115. URL https://www.sciencedirect.com/science/article/pii/S0010027719302896.

Dan Flickinger, Emily M. Bender, and Stephan Oepen. ERG semantic documentation, 2014. URL https://delph-in.github.io/docs/erg/ErgSemantics.

Richard Futrell and Kyle Mahowald. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*, 49:e198, 2026. doi: 10.1017/S0140525X2510112X.

Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger P. Levy. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407, 2019. doi: 10.1016/j.tics.2019.02.003. URL https://doi.org/10.1016/j.tics.2019.02.003.

Yoav Goldberg. Assessing BERT's syntactic abilities, 2019. URL https://arxiv.org/abs/1901.05287.

Michael Wayne Goodman. A Python library for deep linguistic resources. In *2019 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*, pages 1–7, 2019. URL https://ieeexplore.ieee.org/document/8939628.

Joseph H. Greenberg. Some universals of grammar with particular reference to the order of

meaningful elements. In *Universals of Language*, pages 73–113. MIT Press, Cambridge, MA, 1963.

Jane Grimshaw. *Argument Structure*. MIT Press, Cambridge, Mass., 1990. ISBN 9780262071253.

Nicolas Guerin, Emmanuel Chemla, and Shane Steinert-Threlkeld. The impact of syntactic and semantic proximity on machine translation with back-translation. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=6DflIABPQP.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 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)*, pages 1195–1205, June 2018. doi: 10.18653/v1/N18-1108. URL https://aclanthology.org/N18-1108/.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. doi: 10.1162/tacl_a_00306. URL https://aclanthology.org/2020.tacl-1.11/.

Michael Hahn, Dan Jurafsky, and Richard Futrell. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America*, 117(5):2347–2353, 2020. doi: 10.1073/pnas.1910923117. URL https://doi.org/10.1073/pnas.1910923117.

Martin Haspelmath. Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability. *Journal of Linguistics*, 57(3):605–633, 2021. doi: 10.1017/S0022226720000535.

John A. Hawkins. *Efficiency and Complexity in Grammars*. Oxford University Press, Novem-

ber 2004. ISBN 9780199252695. doi: 10.1093/acprof:oso/9780199252695.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199252695.001.0001.

John Hewitt, Michael Hahn, Surya Ganguli, Percy Liang, and Christopher D. Manning. RNNs can generate bounded hierarchical languages with optimal memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1978–2010, November 2020. doi: 10.18653/v1/2020.emnlp-main.156. URL https://aclanthology.org/2020.emnlp-main.156/.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python, 2020. URL https://doi.org/10.5281/zenodo.1212303.

Kristen Howell, Olga Zamaraeva, and Emily M. Bender. Nominalized clauses in the Grammar Matrix. In *Proceedings of the 25th International Conference on Head-Driven Phrase Structure Grammar*, pages 66–86, 2018. doi: 10.21248/hpsg.2018.5. URL https://proceedings.hpsg.xyz/article/view/857.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P. Levy. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, July 2020. doi: 10.18653/v1/2020.acl-main.158. URL https://aclanthology.org/2020.acl-main.158/.

Nathaniel Imel, Qingxia Guo, and Shane Steinert-Threlkeld. An efficient communication analysis of modal typology. *Open Mind*, 10:1–28, 2026. doi: 10.1162/OPMI.a.313. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC13053023/.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714,

August 2024. doi: 10.18653/v1/2024.acl-long.787. URL https://aclanthology.org/2024.acl-long.787/.

Charles Kemp, Yang Xu, and Terry Regier. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4:109–128, 2018. doi: 10.1146/annurev-linguistics-011817-045406. URL https://www.annualreviews.org/doi/10.1146/annurev-linguistics-011817-045406.

Simon Kirby, Hannah Cornish, and Kenny Smith. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686, 2008. doi: 10.1073/pnas.0707835105. URL https://www.pnas.org/doi/abs/10.1073/pnas.0707835105.

Maria Koptjevskaja-Tamm. *Nominalizations*. Theoretical Linguistics Series. Routledge, London and New York, 1993. ISBN 0415060206.

Tatsuki Kuribayashi, Ryo Ueda, Ryo Yoshida, Yohei Oseki, Ted Briscoe, and Timothy Baldwin. Emergent word order universals from cognitively-motivated language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14522–14543, August 2024. doi: 10.18653/v1/2024.acl-long.781. URL https://aclanthology.org/2024.acl-long.781/.

Natalia Levshina. Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations. *Frontiers in Psychology*, 12:648200, 2021. ISSN 1664-1078. doi: 10.3389/fpsyg.2021.648200. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.648200.

Tal Linzen. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 5210–5217, July 2020. doi: 10.18653/v1/2020.acl-main.465. URL https://aclanthology.org/2020.acl-main.465/.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016. doi: 10.1162/tacl_a_00115. URL https://aclanthology.org/Q16-1037/.

Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, October-November 2018. doi: 10.18653/v1/D18-1151. URL https://aclanthology.org/D18-1151/.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, July 2019. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

R. Thomas McCoy, Robert Frank, and Tal Linzen. Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140, 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00304. URL https://doi.org/10.1162/tacl_a_00304.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, July 2019. doi: 10.18653/v1/P19-1491. URL https://aclanthology.org/P19-1491/.

Kanishka Misra and Kyle Mahowald. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida,

USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.53. URL https://aclanthology.org/2024.emnlp-main.53/.

Tom M. Mitchell. The need for biases in learning generalizations. Technical Report CBM-TR-117, Rutgers University, New Brunswick, NJ, May 1980.

Andrea Moro. *Impossible Languages*. The MIT Press, September 2016. ISBN 9780262034890. doi: 10.7551/mitpress/9780262034890.001.0001. URL https://doi.org/10.7551/mitpress/9780262034890.001.0001.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, June 2021. doi: 10.18653/v1/2021.naacl-main.290. URL https://aclanthology.org/2021.naacl-main.290/.

Johanna Nichols. Head-marking and dependent-marking grammar. *Language*, 62(1):56–119, 1986. doi: 10.2307/415601. URL https://doi.org/10.2307/415601.

Michael Noonan. Complementation. In Timothy Shopen, editor, *Language Typology and Syntactic Description, Volume II: Complex Constructions*, pages 52–150. Cambridge University Press, Cambridge, 2nd edition, 2007. doi: 10.1017/CBO9780511619434.002.

William O'Grady. The syntax of alignment: An emergentist typology. In Thiago Costa Chacon, Nala H. Lee, and W. D. L. Silva, editors, *Language Change and Linguistic Diversity: Studies in Honour of Lyle Campbell*, pages 260–280. Edinburgh University Press, 2022. doi: 10.1515/9781474488143-017.

OpenAI. GPT-4o System Card, August 2024. URL https://openai.com/index/gpt-4o-system-card/.

Woodley Packard. ACE: The answer constraint engine, 2018. URL https://sweaglesw.org/linguistics/ace/.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615, 2024. doi: 10.1162/tacl_a_00720. URL https://aclanthology.org/2024.tacl-1.87/.

Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. Number 4 in Studies in Contemporary Linguistics. CSLI Publications and University of Chicago Press, Stanford and Chicago, 1994.

Amanda Popadich and Shane Steinert-Threlkeld. Language models generalize to human-like word order preferences, 2026. URL https://arxiv.org/abs/2608.05028.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. Studying the inductive biases of RNNs with synthetic variations of natural languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3532–3542, June 2019. doi: 10.18653/v1/N19-1356. URL https://aclanthology.org/N19-1356/.

Keren Ruditsky. Action nominals in the Grammar Matrix. Master's thesis, University of Washington, Seattle, WA, 2024. URL https://hdl.handle.net/1773/52815.

Ivan A. Sag, Thomas Wasow, and Emily M. Bender. *Syntactic Theory: A Formal Introduction*. Number 152 in CSLI Lecture Notes. CSLI Publications, Stanford, CA, 2nd edition, 2003. ISBN 9781575864006.

Anna Siewierska. Word order type and alignment type. *Sprachtypologie und Universalienforschung*, 49:149–176, 1996.

Kenny Smith and Elizabeth Wonnacott. Eliminating unpredictable variation through iterated learning. *Cognition*, 116(3):444–449, 2010. ISSN 0010-0277. doi: 10.1016/j.cognition.2010.06.004.

Leon Stassen. *Comparison and Universal Grammar*. Basil Blackwell, Oxford, 1985.

Shane Steinert-Threlkeld. Quantifiers in natural language: Efficient communication and degrees of semantic universals. *Entropy*, 23(10), 2021. ISSN 1099-4300. doi: 10.3390/e23101335. URL https://www.mdpi.com/1099-4300/23/10/1335.

Shane Steinert-Threlkeld and Jakub Szymanik. Ease of learning explains semantic universals. *Cognition*, 195:104076, 2020. ISSN 0010-0277. doi: 10.1016/j.cognition.2019.104076. URL https://www.sciencedirect.com/science/article/pii/S0010027719302495.

Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. What formal languages can transformers express? A survey. *Transactions of the Association for Computational Linguistics*, 12:543–561, 2024. doi: 10.1162/tacl_a_00663. URL https://aclanthology.org/2024.tacl-1.30/.

Sarah Grey Thomason and Terrence Kaufman. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley, 1988. ISBN 9780520057890.

R. L. Trask. On the origins of ergativity. In Frans Plank, editor, *Ergativity: Towards a Theory of Grammatical Relations*, pages 385–404. Academic Press, New York, 1979.

Peter Trudgill. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford Linguistics. Oxford University Press, Oxford, 2011. ISBN 9780199604340.

René van den Berg. *A Grammar of the Muna Language*, volume 139 of *Verhandelingen van*

het Koninklijk Instituut voor Taal-, Land- en Volkenkunde. Foris, Dordrecht, 1989. ISBN 9789067654548.

Dingquan Wang and Jason Eisner. The Galactic Dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505, 2016. doi: 10.1162/tacl_a_00113. URL https://aclanthology.org/Q16-1035/.

Shunjie Wang and Shane Steinert-Threlkeld. Evaluating Transformer's ability to learn mildly context-sensitive languages. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 271–283, December 2023. doi: 10.18653/v1/2023.blackboxnlp-1.21. URL https://aclanthology.org/2023.blackboxnlp-1.21/.

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*, pages 1737–1743, 2020. URL https://escholarship.org/uc/item/3rx3h34n.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL https://aclanthology.org/2020.tacl-1.25/.

Gail Weiss, Yoav Goldberg, and Eran Yahav. On the practical computational power of finite precision RNNs for language recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 740–745, July 2018. doi: 10.18653/v1/P18-2117. URL https://aclanthology.org/P18-2117/.

Aaron Steven White and Kyle Rawlins. Frequency, acceptability, and selection: A case study of clause-embedding. *Glossa: a journal of general linguistics*, 5(1):105, 2020. doi: 10.5334/gjgl.1001. URL https://doi.org/10.5334/gjgl.1001.

Jennifer C. White and Ryan Cotterell. Examining the inductive bias of neural language models with artificial languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 454–463, August 2021. doi: 10.18653/v1/2021.acl-long.38. URL https://aclanthology.org/2021.acl-long.38/.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 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*, pages 211–221, November 2018. doi: 10.18653/v1/W18-5423. URL https://aclanthology.org/W18-5423/.

Tianyang Xu, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, and Alex Warstadt. Can language models learn typologically implausible languages? *Transactions of the Association for Computational Linguistics*, 14:588–611, 2026. doi: 10.1162/tacl.a.640. URL https://aclanthology.org/2026.tacl-1.27/.

Qing Yao, Kanishka Misra, Leonie Weissweiler, and Kyle Mahowald. Both direct and indirect evidence contribute to dative alternation preferences in language models, 2025. URL https://openreview.net/forum?id=h5SRsDax8v.

Appendix A

EVALUATION DETAILS

This appendix provides construction details and result summaries for the targeted minimal-pair evaluation described in Chapter 6. For each phenomenon, I report the target phenomenon, source template, perturbation rule, representative minimal-pair example, results summary, and heatmaps. In all heatmaps, darker green indicates higher accuracy on a 50–100% scale, and orange bars indicate the BAD parse rate on a 0–100% scale.

Box 1 shows the prompt template for generating English source sentences. All phenomena use the same general instructions, with a specific construction block. The minimal-pair examples below use the parameter configuration closest to English, SVO/GN/NOM-ACC/Balancing/NOMN, detailed in Table A.1.

Property	Value	Property	Value
Clause word order	SVO	Finite S marker	0
NP word order	GN	Finite A marker	0
Alignment	NOM-ACC	Finite P marker	*-ca*
Complement system	Balancing	Finite verb form	*-s*
ANC strategy	NOMN	Genitive marker	*-ge*
Complement verb form	finite	ANC verb form	*-ing*
ANC intransitive order	SV	ANC transitive order	AVP
ANC S marker	*-ge*	ANC A marker	*-ge*
ANC P marker	*-ob*		

Table A.1: Parameter values for the example configuration used in Appendix A.

Box 1. Source sentence generation prompt

You are helping generate English source sentences for a controlled grammar evaluation. Generate natural English sentences that instantiate the target construction described below. The sentences should be grammatical, semantically plausible, and close to ordinary corpus English. Avoid producing a mechanical list with repeated lexical items or repeated sentence frames.

General requirements:

- Each sentence must clearly instantiate the target construction.
- Avoid structures that are irrelevant to the target construction and may interfere with extraction or evaluation.
- Use ordinary, high-frequency vocabulary whenever possible.
- Vary lexical items and noun phrase types where appropriate.
- Return one sentence per line, with no numbering, explanation, or section headings.

Target construction (example: 1.1 Intransitive finite verb form):

- Generate 160 ordinary finite intransitive clauses.
- Each sentence must contain one overt subject and one finite intransitive verb.
- The target verb should be a lexical intransitive verb; do not make an auxiliary, modal, or aspectual marker the target.
- Generate approximately half of the sentences with simple subjects and half with possessive/genitive subjects.
- Possessive/genitive subjects should contain one possessor and one head noun.
- Avoid transitive clauses, clausal complements, passives, coordination, relative clauses, nominalizations, idioms, and unstable argument structures.
- Use ordinary, high-frequency intransitive verbs.

Output format:

- Return one sentence per line.
- Do not include numbering, explanations, or section headings.

A.1 Basic Finite Clauses

A.1.1 Intransitive Finite Verb Form

This phenomenon tests whether the models learned the finite form of intransitive verbs in finite intransitive clauses. The source templates include two types of subjects: simple NPs and NPs with a genitive possessor.

In all target grammars, the finite verb form is realized with the suffix *-s*. The BAD sentence therefore replaces the target intransitive verb with its nonfinite form, changing *V-s* to *V-ing*. An example is shown below:

(8) a. GOOD: *Girlge teacher laughs*

b. BAD: *Girlge teacher laughing*

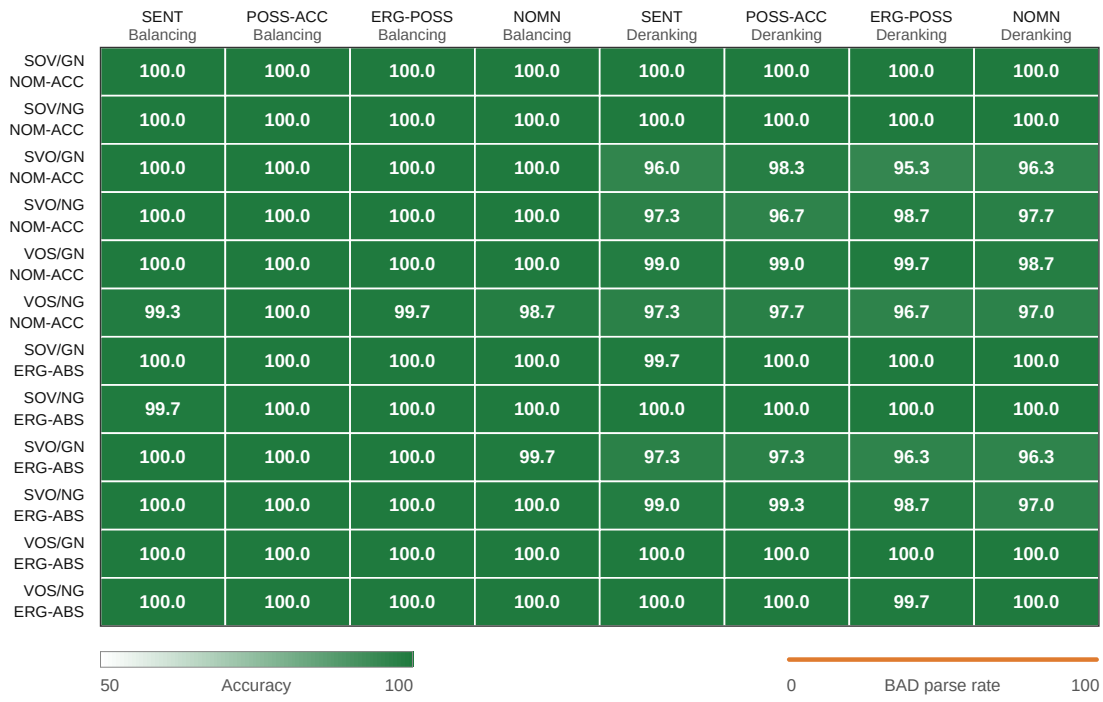

Figure A.1: Accuracy heatmap for 1.1 Intransitive finite verb form.

Figure A.1 shows that the models perform near ceiling on this phenomenon. The mean

accuracy across the 96 grammars is 99.4%, with a minimum of 95.3% and a maximum of 100%. The BAD parse rate is 0 for all grammars. Overall, this result shows that the models robustly learned that finite intransitive clauses require the finite verb form.

A.1.2 Intransitive S Marker

This phenomenon tests whether the models learned that the S argument in a finite intransitive clause should not receive an overt marker. The source templates include two types of subjects: simple NPs and NPs with a genitive possessor.

In all target grammars, the finite intransitive S marker is zero. The BAD sentence therefore adds an overt marker, either *-ca* or *-ge*, to the head of S. An example is shown below:

(9) a. GOOD: *Doctorge nurse waits*

b. BAD: *Doctorge nurseca waits*

Figure A.2 shows that models generally learned this phenomenon well. The mean accuracy across the 96 grammars is 87.7%, with a minimum of 51.7% and a maximum of 99.7%. The BAD parse rate is low, ranging from 3.0% to 4.0% across grammars.

The weakest configuration is SOV/NG + ERG-ABS, with a mean accuracy of 59.7%. Another weaker configuration is VOS/NG + NOM-ACC, with a mean accuracy of 74.5%. These errors mainly reflect an interaction between alignment and word order. In these configurations, adding a marker to S can make the BAD sentence resemble a transitive argument-marking pattern. For example, under SOV/NG + ERG-ABS, the GOOD sentence *Nurse doctorge waits* is perturbed into *Nurseca doctorge waits*. Since *-ca* is the legal A marker in this configuration, the model may treat *Nurseca* as a transitive A and *doctorge* as a transitive P. This suggests that the model does not robustly distinguish genitive possessor marking from transitive argument marking in this configuration.

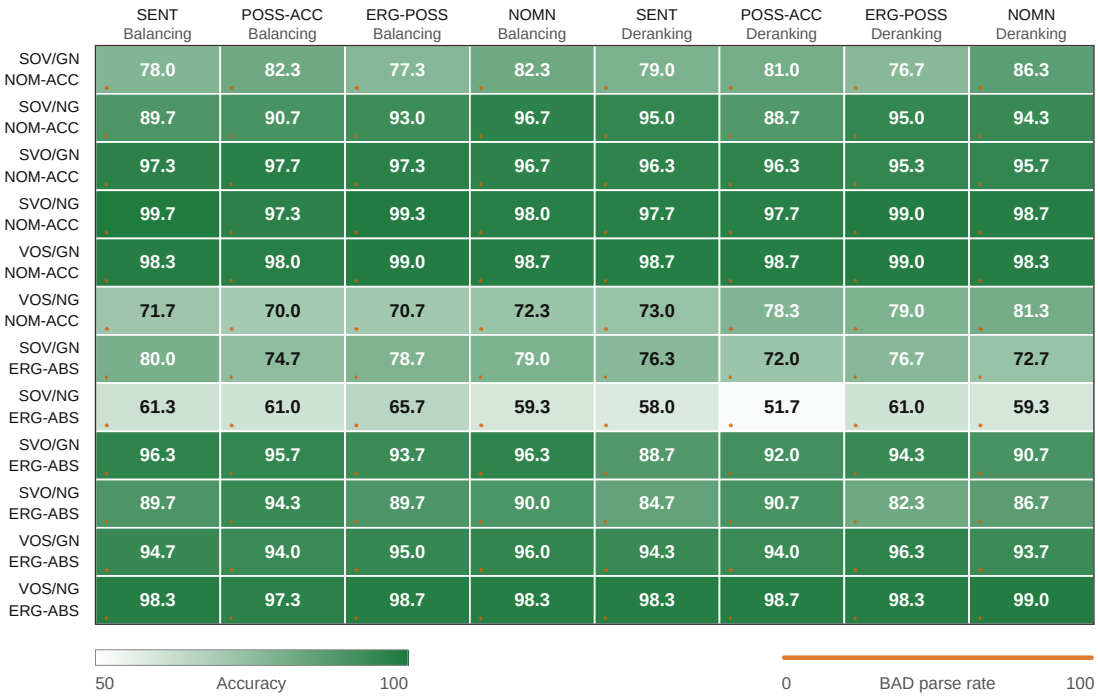

Figure A.2: Accuracy heatmap for 1.2 Intransitive S marker.

A.1.3 Intransitive S–V Order

This phenomenon tests whether the models learned the linear order between S and the finite verb in finite intransitive clauses. Intransitive clauses distinguish two surface orders: SOV and SVO yield S–V order, while VOS yields V–S order.

The BAD sentence is constructed by swapping the entire subject span and the finite verb. An example is shown below:

(10) a. GOOD: *Bakerge bread rises*

b. BAD: *Rises bakerge bread*

Figure A.3 shows that the models perform at ceiling on this phenomenon. The mean accuracy across the 96 grammars is 99.99%, with a minimum of 99.3% and a maximum of 100%. The BAD parse rate is 0 for all grammars. Overall, this result shows that the models robustly learned the basic S–V order of finite intransitive clauses.

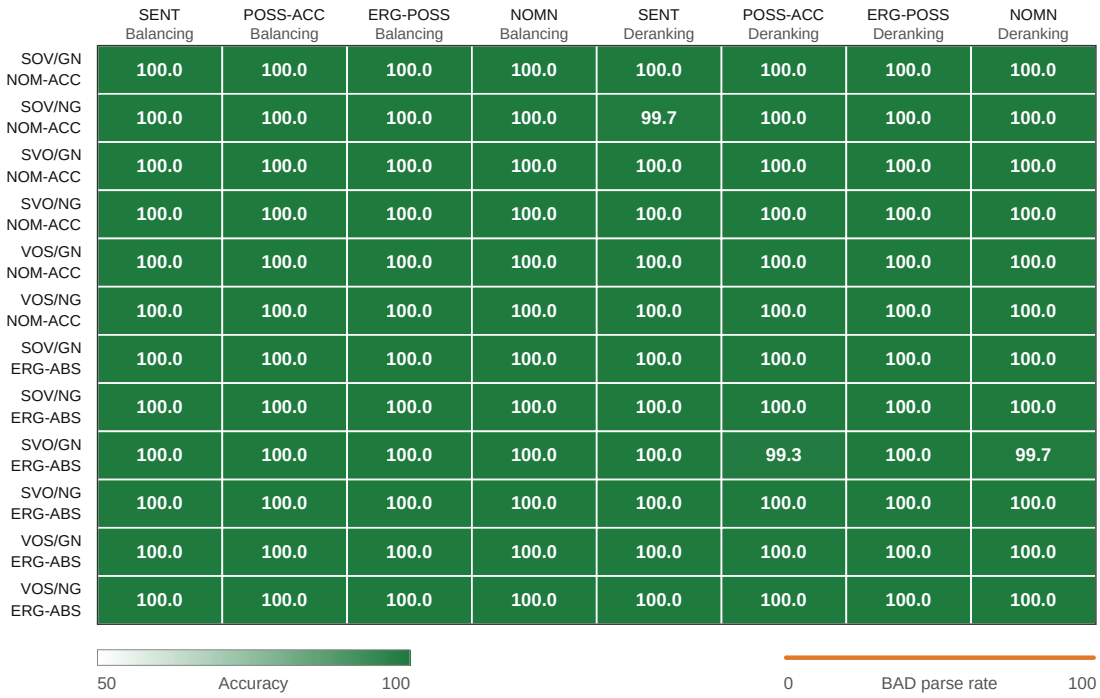

Figure A.3: Accuracy heatmap for 1.3 Intransitive S–V order.

A.1.4 Transitive Finite Verb Form

This phenomenon tests whether the models learned the finite form of transitive verbs in finite transitive clauses. The source templates include transitive clauses in which A and P may be simple NPs or NPs with a genitive possessor.

In all target grammars, the finite verb form is realized with the suffix *-s*. The BAD sentence therefore replaces the target transitive verb with its nonfinite form, changing *V-s* to *V-ing*. An example is shown below:

(11) a. GOOD: *Nurse holds cupca*

b. BAD: *Nurse holding cupca*

Figure A.4 shows that the models perform near ceiling on this phenomenon. The mean accuracy across the 96 grammars is 99.7%, with a minimum of 97.0% and a maximum of 100%. The BAD parse rate is 0 for all grammars. The few remaining errors are mostly in

SVO grammars with Deranking, especially SVO/NG + ERG-ABS + Deranking. This reflects the fact that *V-ing* is more common in deranking systems than in balancing systems.

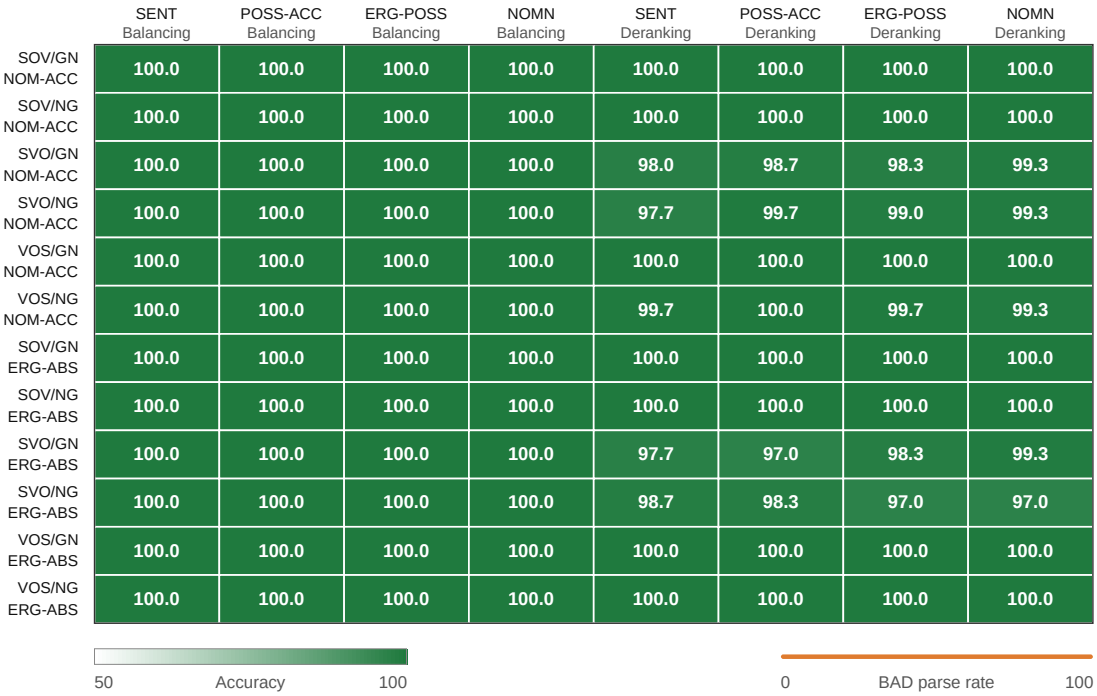

Figure A.4: Accuracy heatmap for 1.4 Transitive finite verb form.

A.1.5 Transitive A Marker

This phenomenon tests whether the models learned the marking of the A argument in basic finite transitive clauses. The source templates include four types of transitive clauses, where A and P can each be either a simple NP or an NP with a genitive possessor.

In the target grammars, A marking is determined by alignment. In NOM-ACC grammars, A has zero marking, while in ERG-ABS grammars, A has *-ca*. So in NOM-ACC grammars, the BAD sentence adds *-ca* or *-ge* to A. In ERG-ABS grammars, the BAD sentence either removes *-ca* from A or replaces it with *-ge*. An example is shown below:

(12) a. GOOD: *Doctorge assistant finds fileca*

b. BAD: *Doctorge assistantca finds fileca*

Figure A.5 shows that the models generally learned transitive A marking. The mean accuracy across the 96 grammars is 96.6%, with a minimum of 75.0% and a maximum of 100%. The BAD parse rate ranges from 0.0% to 53.0%. However, BAD parse rate and model accuracy show an asymmetry with respect to alignment: NOM-ACC grammars have a lower mean BAD parse rate of 7.3%, but a lower mean accuracy of 93.2%; ERG-ABS grammars have a higher mean BAD parse rate of 18.1%, but reach 100% accuracy.

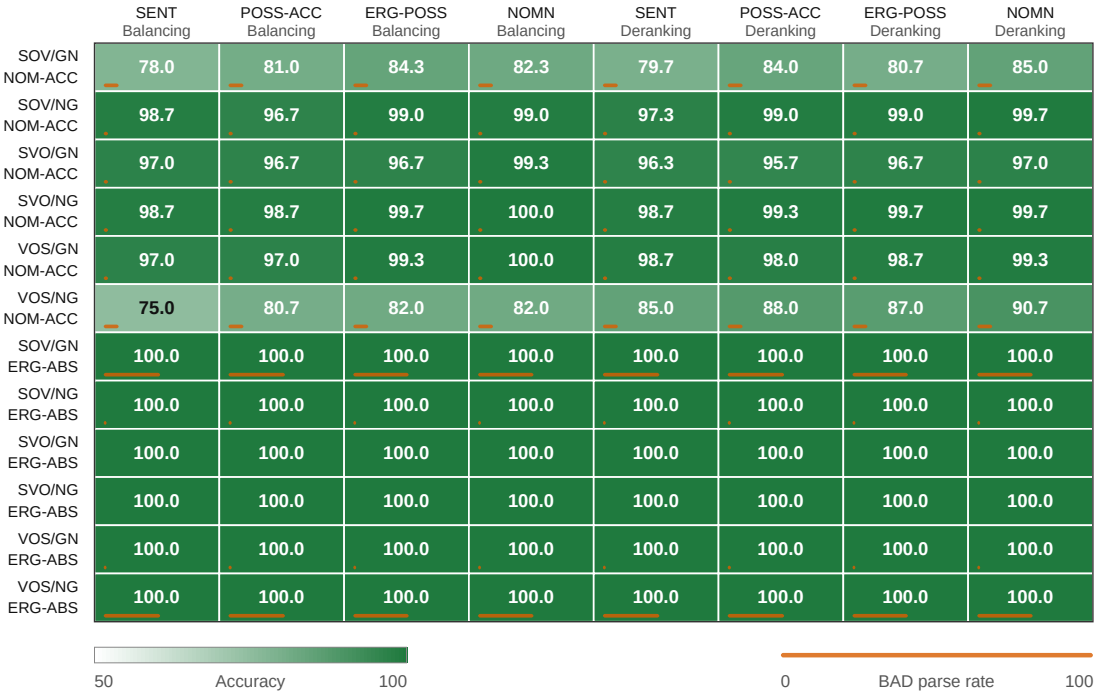

Figure A.5: Accuracy heatmap for 1.5 Transitive A marker.

Within NOM-ACC, the weakest configuration is SOV/GN + NOM-ACC, with a mean accuracy of 81.9% and a BAD parse rate of 14.0%. For example, *Nurse childge kneeca cleans* is perturbed into *Nursege childge kneeca cleans*. This BAD sentence cannot be parsed by the grammar, but the model still assigns it a higher score. A likely reason is that the model does not robustly identify *-ge* as an illegal marker in the A position.

In contrast, although ERG-ABS grammars have a higher BAD parse rate, these parses often require a larger semantic reanalysis. Under SOV/GN + ERG-ABS, the mean accuracy is 100% even though the BAD parse rate is 52.5%. For example, *Nurseca bandage touchs* is perturbed into *Nursege bandage touchs*. This BAD sentence can be parsed as "the nurse's bandage touches", where the original A is reanalyzed as a genitive possessor, while the original P becomes the intransitive S. This analysis is syntactically available, but semantically unnatural, since many inanimate NPs that occur as P are not suitable as intransitive S.

A.1.6 Transitive P Marker

This phenomenon tests whether the models learned the marking of the P argument in basic finite transitive clauses. The source templates include four types of transitive clauses, where A and P can each be either a simple NP or an NP with a genitive possessor.

In the target grammars, P marking is determined by alignment. In NOM-ACC grammars, P bears *-ca*, while in ERG-ABS grammars, P has zero marking. In NOM-ACC grammars, the BAD sentence either removes *-ca* from P or replaces it with *-ge*. In ERG-ABS grammars, the BAD sentence incorrectly adds *-ca* or *-ge* to P. The perturbation applies only to the P head. An example is shown below:

(13) a. GOOD: *Shop sells breadca*

b. BAD: *Shop sells bread*

Figure A.6 shows that the models generally learned transitive P marking. The mean accuracy across the 96 grammars is 97.9%, with a minimum of 81.7% and a maximum of 100%. The BAD parse rate ranges from 0.0% to 50.0%. Specifically, NOM-ACC grammars all reach 100% accuracy, but have a higher mean BAD parse rate of 16.7%; ERG-ABS grammars have a lower mean accuracy of 95.8%, but a lower mean BAD parse rate of 5.7%.

This phenomenon mirrors the pattern in Section A.1.5. For example, SOV/NG + NOM-ACC reaches 100% accuracy, but has a BAD parse rate of 50.0%. The GOOD sentence *Student answerca copys* is perturbed into *Student answerge copys*. This BAD sentence can

Figure A.6: Accuracy heatmap for 1.6 Transitive P marker.

be parsed as roughly "the answer's student copies", where the original P is reanalyzed as a genitive possessor. Since this parse also involves semantic reanalysis, the model still consistently prefers the GOOD sentence. In contrast, in ERG-ABS grammars, P has zero marking and therefore lacks an overt P cue, making it more vulnerable to interference from incorrect markers. The weakest configuration is SOV/NG + ERG-ABS, with a mean accuracy of 84.6% and a BAD parse rate of 11.0%.

A.1.7 Transitive A–V Order

This phenomenon tests whether the models learned the relative order between the A argument and the finite transitive verb in basic finite transitive clauses. The source templates include four types of transitive clauses, where A and P can each be either a simple NP or an NP with a genitive possessor.

The BAD sentence is constructed by swapping the A span and the V+P chunk. An

example is shown below:

(14) a. GOOD: *Bakerge daughter slices breadca*

b. BAD: *Slices breadca bakerge daughter*

Figure A.7 shows that the models reach ceiling on this phenomenon. All 96 grammars have 100% accuracy, and the BAD parse rate is 0 for all grammars. This shows that the models learned the order relation between A and V very robustly.

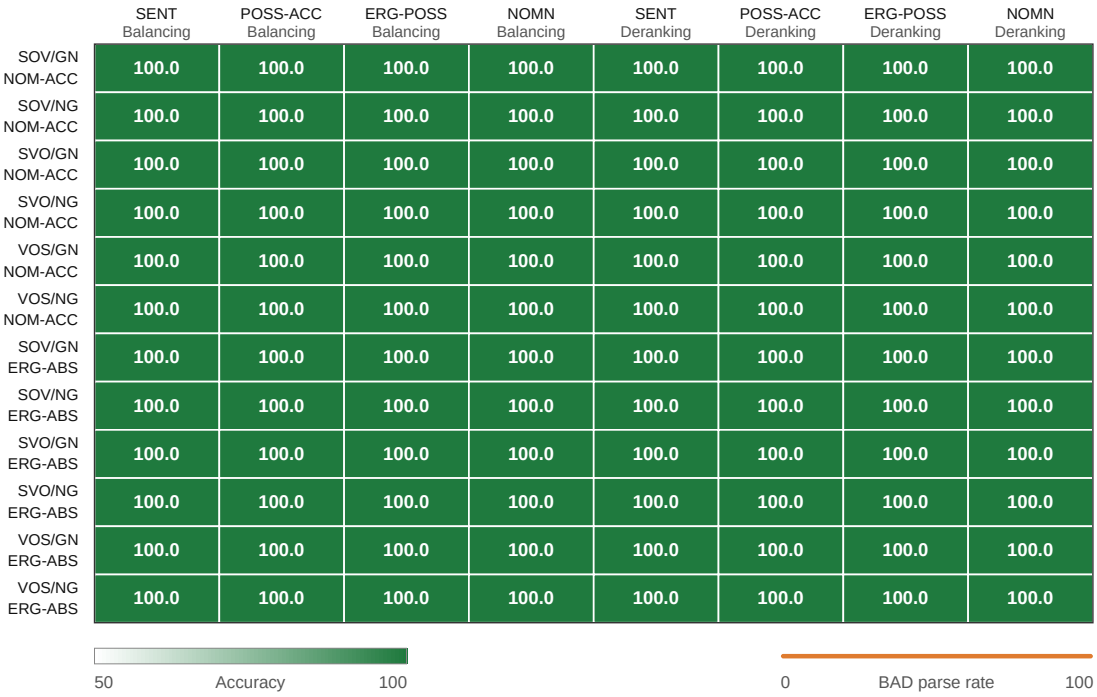

Figure A.7: Accuracy heatmap for 1.7 Transitive A–V order.

A.1.8 Transitive V–P Order

This phenomenon tests whether the models learned the relative order between the finite transitive verb and the P argument in basic finite transitive clauses. The source templates include four types of transitive clauses, where A and P can each be either a simple NP or an NP with a genitive possessor.

The BAD sentence is constructed by swapping V and the P span, while the A span remains in place. An example is shown below:

(15) a. GOOD: *Shopkeeper closes windowca*

b. BAD: *Shopkeeper windowca closes*

Figure A.8 shows that the models reach ceiling on this phenomenon. All 96 grammars have 100% accuracy. The BAD parse rate is also nearly zero, with a mean of 0.08%. This shows that the models learned the order relation between V and P very robustly.

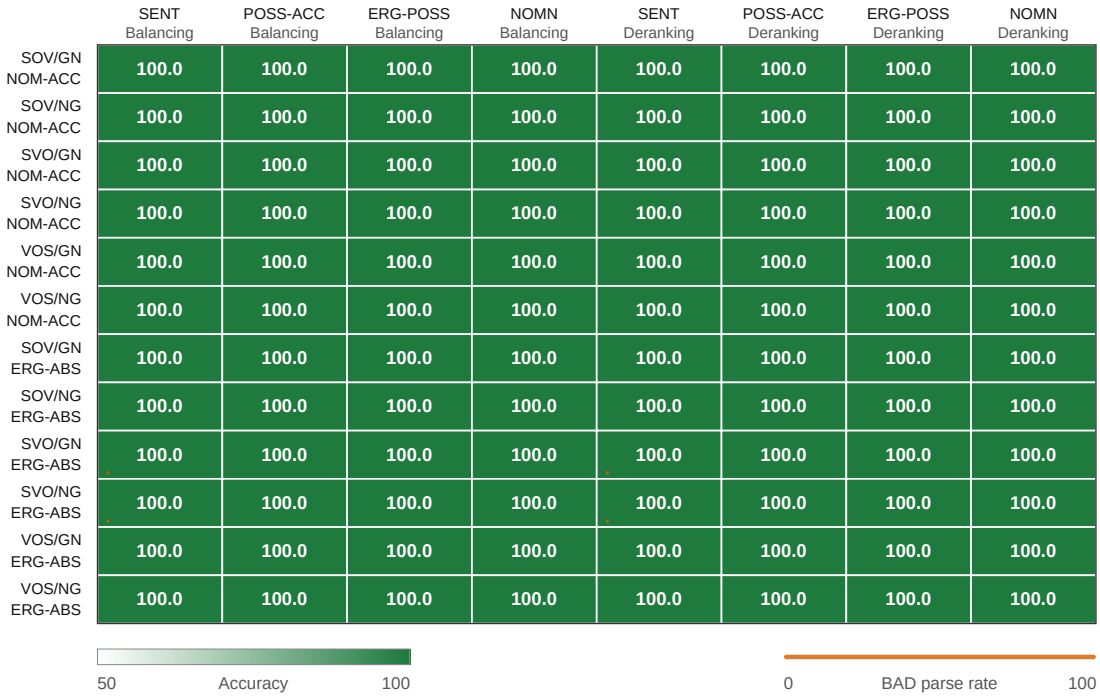

Figure A.8: Accuracy heatmap for 1.8 Transitive V–P order.

A.1.9 Intransitive V Valency

This phenomenon tests whether the models learned verb valency in basic finite intransitive clauses. The GOOD sentence is a finite clause containing S and an intransitive verb. The source templates include simple-NP subjects and subjects with a genitive possessor.

This phenomenon uses a valency pair table for perturbation. The verb pairs come from the BLiMP intransitive paradigm [Warstadt et al., 2020]. The GOOD verb is an intransitive verb, while the BAD verb is a transitive verb that does not match the environment. The source sentences are regenerated with prompts around the GOOD verbs. The BAD sentence is constructed by replacing only the target verb stem while keeping the finite form unchanged. An example is shown below:

(16) a. GOOD: *Clerk smiles*

b. BAD: *Clerk admires*

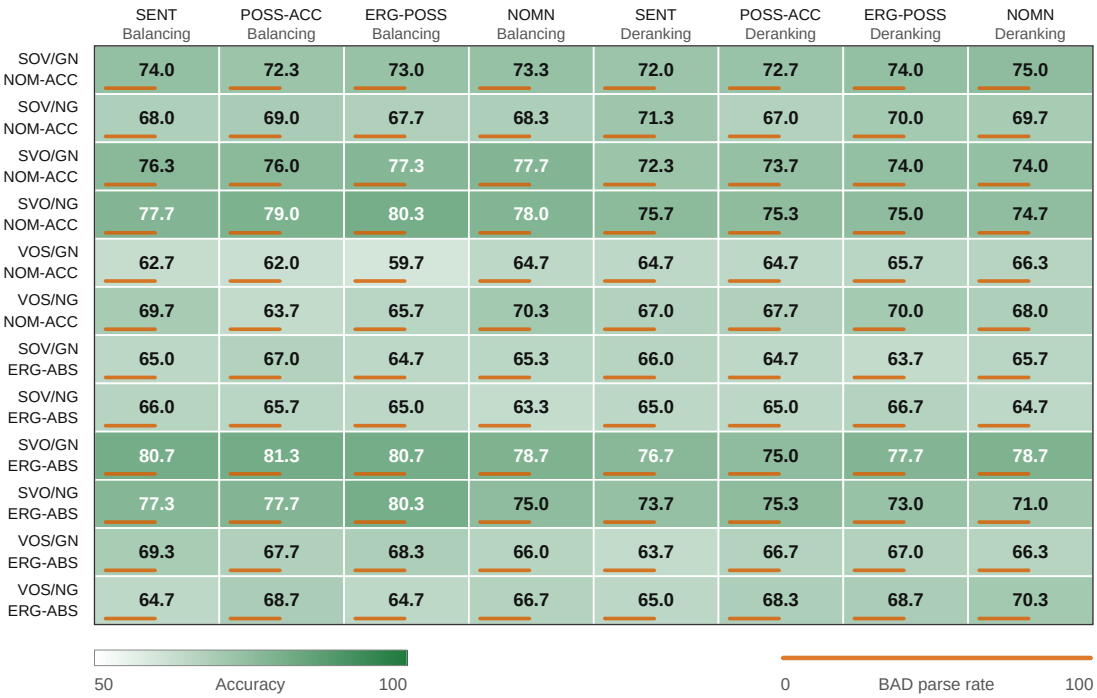

Figure A.9: Accuracy heatmap for 1.9 Intransitive V valency.

Figure A.9 shows that this phenomenon is much harder than the preceding form, marker, and word-order phenomena. The mean accuracy across the 96 grammars is 70.4%, with a minimum of 59.7% and a maximum of 81.3%. The BAD parse rate is 50.0% for all grammars,

showing that about half of the BAD verbs occurred without a complement in the original training corpus. The English baseline accuracy is 56.0%.

Across clause word orders, SVO grammars have the highest mean accuracy at 76.6%. SOV and VOS grammars are lower, at 68.1% and 66.4%, respectively. The other parameters have only small effects. This suggests that SVO word order helps the models distinguish the environments of intransitive and transitive verbs. At the same time, the target-language models perform better than the English baseline, indicating that semantic simplification and grammar generation provide a clearer valency signal. However, the overall accuracy is still not high, showing that lexical argument structure remains difficult for small models trained on small corpora.

A.1.10 Transitive V Valency

This phenomenon tests whether the models learned verb valency in basic finite transitive clauses. The GOOD sentence is a finite clause containing A, a transitive verb, and P. The source templates include four types of transitive clauses, where A and P can each be either a simple NP or an NP with a genitive possessor.

This phenomenon also uses a valency pair table for perturbation. The verb pairs come from the BLiMP transitive paradigm [Warstadt et al., 2020]. The GOOD verb is a transitive verb, while the BAD verb is an intransitive verb that does not match the environment. The source sentences are regenerated with prompts around the GOOD verbs. The BAD sentence is constructed by replacing only the target verb stem while keeping the finite form unchanged. An example is shown below:

(17) a. GOOD: *Guest insults hostca*

b. BAD: *Guest sighs hostca*

Figure A.10 shows that this phenomenon is also difficult, but the models perform slightly better here than on Intransitive V Valency. The mean accuracy across the 96 grammars is 76.2%, with a minimum of 67.7% and a maximum of 86.7%. The BAD parse rate is 93.0%

for all grammars, showing that most BAD verbs also occurred in transitive frames in the original training corpus. The English baseline accuracy is 66.0%.

Figure A.10: Accuracy heatmap for 1.10 Transitive V valency.

Across clause word orders, SVO grammars have the highest mean accuracy at 79.4%. SOV and VOS grammars are lower, at 74.4% and 74.8%, respectively. The other parameters have only small effects. This pattern is consistent with the intransitive valency result. The target-language models also perform better than the English baseline, suggesting that the synthetic training data provide a clearer signal for transitive valency as well.

A.2 Noun Phrases

A.2.1 Intransitive Genitive Marker

This phenomenon tests whether the models learned the genitive marker on the possessor inside the subject NP of an intransitive clause. The GOOD sentence is a finite intransitive

clause in which S is an NP with one genitive possessor.

In all target grammars, the genitive marker is *-ge*. The BAD sentence is constructed by modifying only the *-ge* on the possessor, either deleting it or replacing it with *-ca*. An example is shown below:

(18) a. GOOD: *Studentge father laughs*

b. BAD: *Student father laughs*

Figure A.11 shows that the models generally learned genitive marking inside intransitive subjects. The mean accuracy across the 96 grammars is 89.3%, with a minimum of 73.3% and a maximum of 100%. The BAD parse rate ranges from 0% to 48%, with a mean of 16.0%.

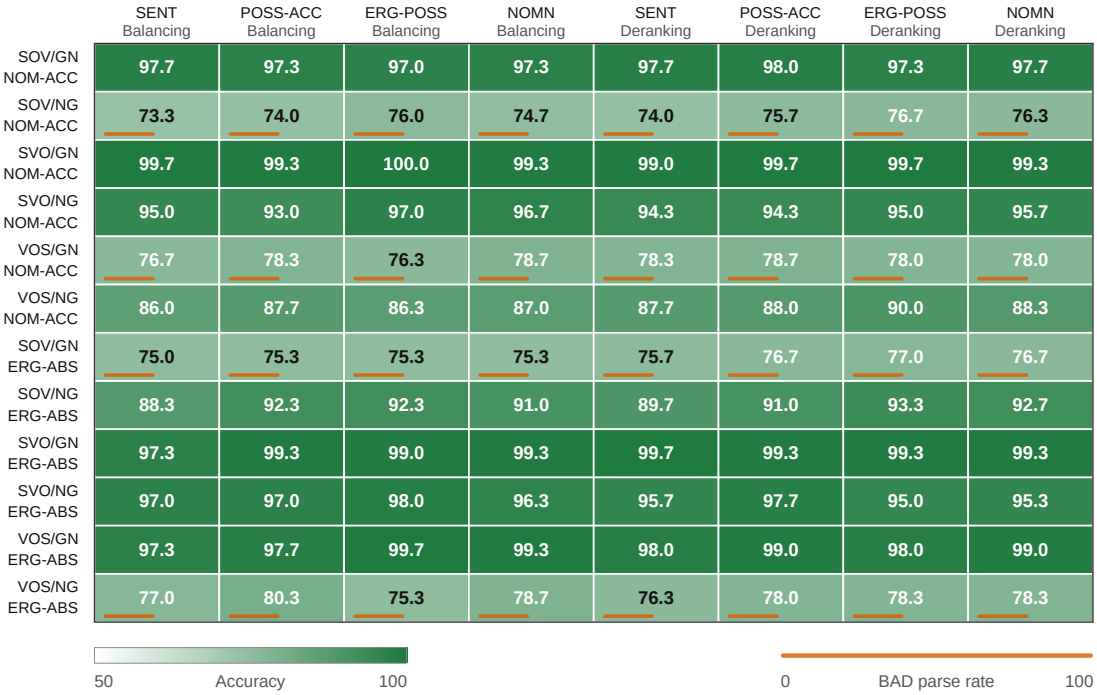

Figure A.11: Accuracy heatmap for 2.1 Intransitive genitive marker.

SVO grammars perform best, with a mean accuracy of 97.6%. The weaker configurations are mainly found under SOV and VOS grammars. The four weakest groups are SOV/NG +

NOM-ACC, VOS/GN + NOM-ACC, SOV/GN + ERG-ABS, and VOS/NG + ERG-ABS, with mean accuracies of 75.1%, 77.9%, 75.9%, and 77.8%, respectively.

The main errors occur when the model treats BAD sentences with *-ca* as legal transitive patterns. For example, under SOV/GN + ERG-ABS, the GOOD sentence *Studentge father laughs* can be perturbed into *Studentca father laughs*. Since *-ca* is the A marker in this configuration, the BAD sentence resembles an SOV transitive clause, roughly interpretable as "the student laughs the father".

A.2.2 Transitive Genitive Marker

This phenomenon tests whether the models learned the genitive marker on possessors inside A/P argument NPs in transitive clauses. The GOOD sentence is a finite transitive clause in which at least one argument NP has a genitive possessor.

The source templates and perturbation rules cover four cases: targeting the possessor inside A when A has a possessor, targeting the possessor inside P when P has a possessor, and targeting the possessor inside A or P separately when both A and P have possessors.

In all target grammars, the genitive marker is *-ge*. The BAD sentence is constructed by modifying only the *-ge* on the target possessor, either deleting it or replacing it with *-ca*. An example is shown below:

(19) a. GOOD: *Olivia calls studentge roommateca*

b. BAD: *Olivia calls studentca roommateca*

Figure A.12 shows that the models generally learned genitive marking inside transitive argument NPs. The mean accuracy across the 96 grammars is 91.6%, with a minimum of 79.3% and a maximum of 100%. The BAD parse rate is very low, with a mean of only 0.1%.

The result is broadly consistent with Section A.2.1. SVO grammars perform best, with a mean accuracy of 95.3%; SOV and VOS grammars are lower, at 87.9% and 91.5%, respectively. The weakest group is SOV/NG + ERG-ABS, with a mean accuracy of 81.5%. Overall accuracy is slightly higher than in Section A.2.1, possibly because after the marker

Figure A.12: Accuracy heatmap for 2.2 Transitive genitive marker.

perturbation, the model can more easily identify that the verb cannot accommodate three arguments.

A.2.3 Genitive Possessor–Head Order

This phenomenon tests whether the models learned the relative order of the possessor and the head noun inside genitive NPs. The source templates place the target genitive NP in one of three positions: intransitive S, transitive A, or transitive P.

In the target grammars, NP word order is either GN or NG. In GN grammars, the possessor precedes the head noun; in NG grammars, the head noun precedes the possessor. The BAD sentence is constructed by swapping only the possessor and the head noun inside the target genitive NP. An example is shown below:

(20) a. GOOD: *Reporter interviews clerkge bossca*

b. BAD: *Reporter interviews bossca clerkge*

Figure A.13 shows that the models generally learned genitive possessor–head order. The mean accuracy across the 96 grammars is 94.9%, with a minimum of 71.7% and a maximum of 100%. The BAD parse rate ranges from 0% to 36%, with a mean of 23.7%.

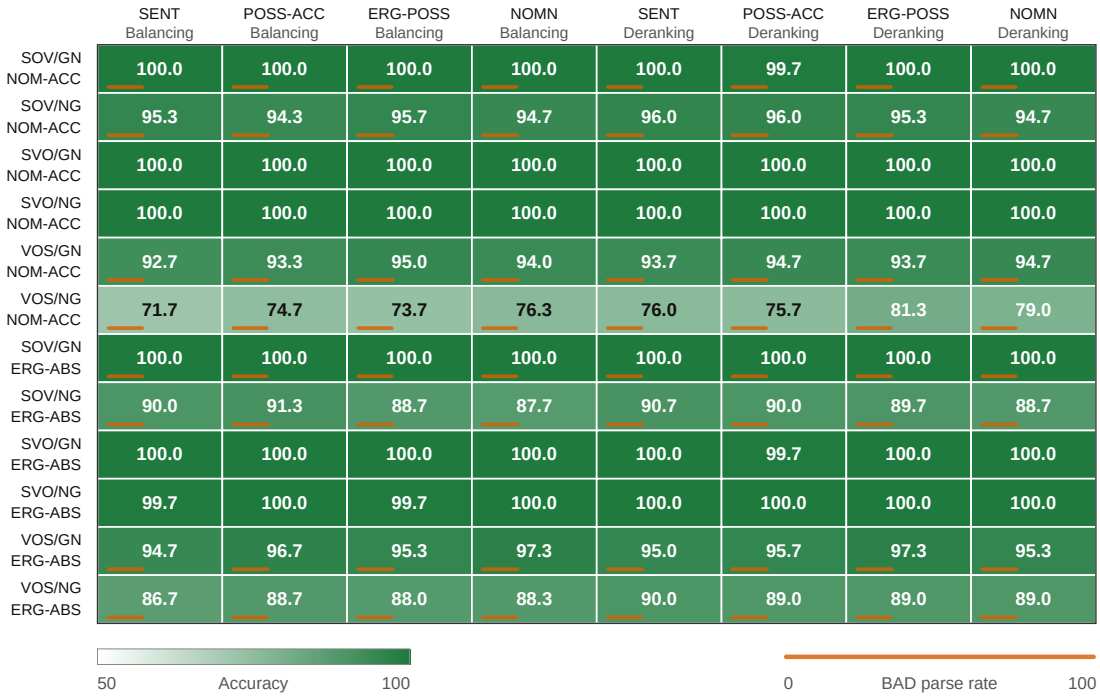

Figure A.13: Accuracy heatmap for 2.3 Genitive possessor–head order.

Across typological parameters, SVO grammars perform best, with a mean accuracy of 100%. SOV grammars are also high, with a mean accuracy of 96.2%, while VOS grammars are weaker, at 88.6%. NP order also has an effect. GN grammars have a mean accuracy of 98.3%, compared with 91.6% for NG grammars.

The weaker performance mainly comes from cases where the perturbed genitive sequence can be reanalyzed as part of another argument. For example, the VOS/NG + NOM-ACC configuration has a mean accuracy of 76.0%. Under this configuration, the GOOD sentence *Chases rabbitca dog noahge* is perturbed into *Chases rabbitca noahge dog*. The GOOD sen-

tence has the structure *Chases [rabbitca] [dog noahge]*, roughly corresponding to "Noah's dog chases the rabbit." However, the BAD sentence can be reanalyzed as *Chases [rabbitca noahge] [dog]*, roughly corresponding to "the dog chases Noah's rabbit." Both analyses are semantically reasonable, so this case should not be treated as a simple model mistake. Nevertheless, it suggests that genitive order is more prone to confusion in this configuration.

A.3 Clausal Complements

A.3.1 Clausal Complement Verb Form

This phenomenon tests whether the models learned the form of the embedded verb in clausal complements. The GOOD sentence is a finite matrix clause whose matrix verb is a clausal-complement verb, and whose complement is an embedded intransitive or transitive clause. The source templates cover both embedded IV and embedded TV clauses.

In the target grammars, embedded verb form is determined by the complement system. In Balancing grammars, embedded V uses the finite form *V-s*; in Deranking grammars, embedded V uses the nonfinite form *V-ing*. The BAD sentence changes the embedded verb form to the other form. An example is shown below:

(21) a. GOOD: *Captain notices boat drifts*

b. BAD: *Captain notices boat drifting*

Figure A.14 shows that the models generally learned clausal-complement verb form well. The mean accuracy across the 96 grammars is 98.6%, with a minimum of 92.7% and a maximum of 100%. The BAD parse rate ranges from 0% to 100%, with a mean of 8.5%. Balancing and Deranking differ only slightly, with mean accuracies of 98.7% and 98.4%, respectively.

Among these configurations, Balancing + ERG-ABS + SENT has a BAD parse rate of 100%, but its mean accuracy remains high at 96.7%. In this configuration, once the embedded verb is perturbed to *V-ing*, the whole embedded sequence can be reanalyzed by the grammar as an ANC functioning as matrix P. For example, the GOOD sentence *Farmerca cow calves*

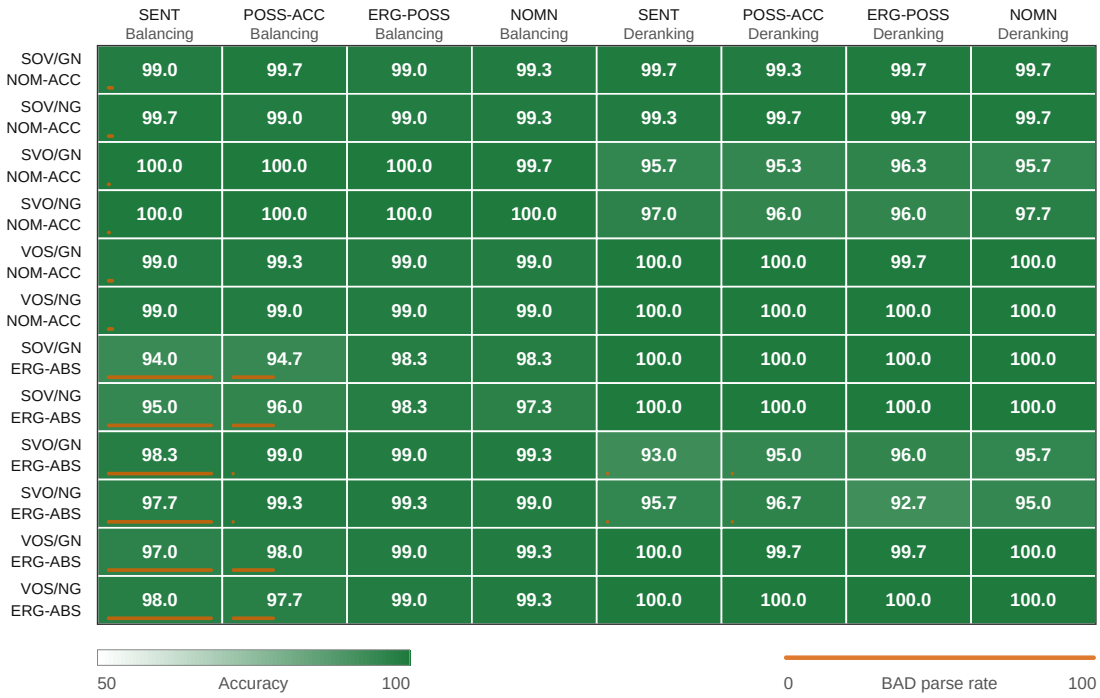

Figure A.14: Accuracy heatmap for 3.1 Clausal complement verb form.

reports is perturbed into *Farmerca cow calveing reports*. The sequence *cow calveing* can be analyzed as an ANC. At the same time, under ERG-ABS alignment, matrix P does not require an overt marker, and verbs such as *reports* may also occur in ordinary transitive frames in the training corpus. Together, these factors make the perturbed sentence parsable.

Nevertheless, the models still maintain high accuracy in these configurations. This shows that the models distinguish well between the distribution of clausal-complement verbs and ordinary transitive verbs, and also learn the difference between the complement types they select.

A.3.2 Clausal Complement S Marker

This phenomenon tests whether the models learned the marker on embedded intransitive S in clausal complements. The GOOD sentence is a finite matrix clause whose complement is an embedded intransitive clause. In the source templates, the embedded S can be either a

simple noun or an NP with a genitive possessor.

In all target grammars, the embedded S marker is zero, the same as the finite intransitive S marker. The BAD sentence is constructed by adding *-ge* only to the head of the embedded S. An example is shown below:

(22) a. GOOD: *She knows bus stops*

b. BAD: *She knows busge stops*

Figure A.15 shows that this phenomenon is harder than the corresponding finite-clause S marker phenomenon. The mean accuracy across the 96 grammars is 75.2%, with a minimum of 29.0% and a maximum of 99.0%. The BAD parse rate ranges from 1.0% to 100%, with a mean of 37.3%. Across typological parameters, SVO grammars perform best, with a mean accuracy of 87.6%; VOS grammars are intermediate, at 74.1%; and SOV grammars are weakest, at 63.8%. The difference by alignment is also clear: ERG-ABS grammars reach 83.6% accuracy, compared with 66.7% for NOM-ACC grammars.

This phenomenon shows a strong asymmetry between BAD parse rate and model accuracy. The weakest configuration is SOV/NG + NOM-ACC, with a mean accuracy of only 35.1%, even though its mean BAD parse rate is also low, at 15.9%. For example, under this configuration, the GOOD sentence *Coach player recovers believes* is perturbed into *Coach playerge recovers believes*. The GOOD sentence roughly corresponds to "the coach believes that the player recovers." In the BAD sentence, however, *Coach playerge* forms an NG genitive NP on the surface, so the whole sequence may be interpreted as something like "the player's coach recovers believes." Although this BAD sentence cannot be parsed by the grammar because *believes* cannot take an object, this local genitive-like pattern followed by a natural finite verb may still increase the model's score for the BAD sentence.

By contrast, SOV/GN + ERG-ABS + Deranking + POSS-ACC has a BAD parse rate of 100.0%, but still reaches 82.3% accuracy. Under this configuration, the GOOD sentence *Heca guest leaveing thinks* is perturbed into *Heca guestge leaveing thinks*. The GOOD sentence roughly corresponds to "He thinks that the guest leaves." In the BAD sentence, *guestge*

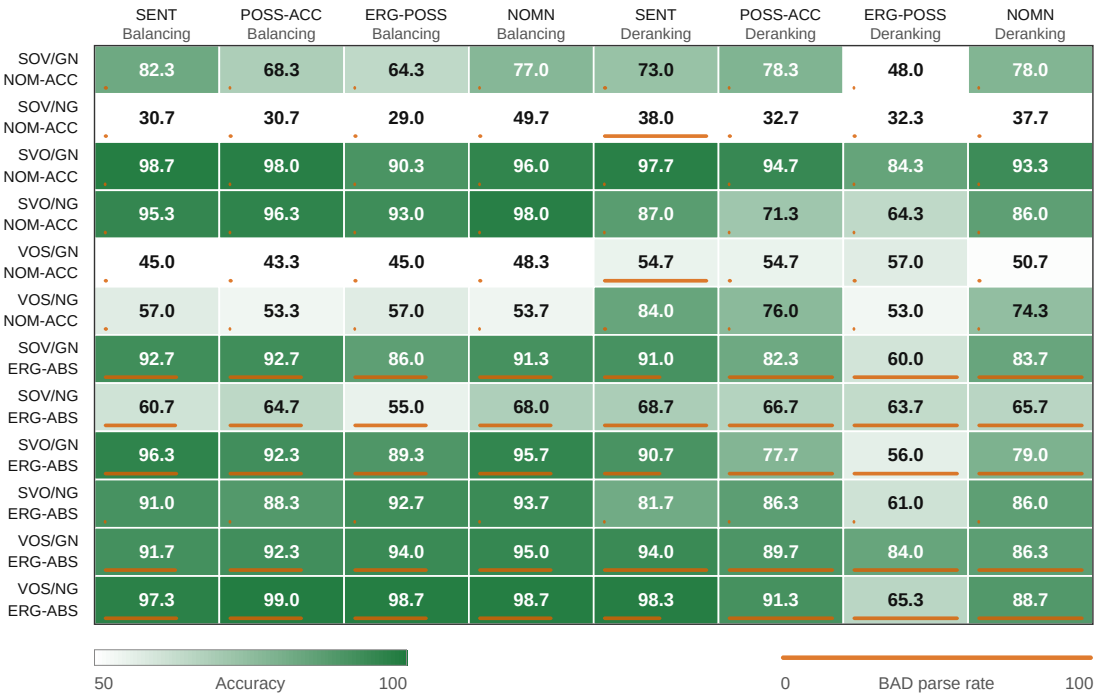

Figure A.15: Accuracy heatmap for 3.2 Clausal complement S marker.

leaveing can be analyzed as an ANC, with *guestge* as the ANC-internal S. The whole sentence is therefore similar to "He thinks [the guest's leaving]." This analysis can be parsed by the grammar, but it reanalyzes the original clausal complement as an ANC in the matrix P position. The model still prefers the GOOD sentence, suggesting that it can distinguish the complement preferences of clausal-complement verbs from those of ordinary transitive verbs.

A.3.3 Clausal Complement A Marker

This phenomenon tests whether the models learned the marker on embedded transitive A in clausal complements. The GOOD sentence is a finite matrix clause whose complement is an embedded transitive clause. In the source templates, matrix A, embedded A, and embedded P can each be either a simple NP or an NP with a genitive possessor.

In clausal complements, embedded argument marking follows the finite-clause system. Therefore, the embedded A marker is determined by alignment. In NOM-ACC, embedded

A is zero-marked; in ERG-ABS, embedded A bears *-ca*. The BAD sentence is constructed by perturbing only the head of embedded A, replacing it with the A marker used under the corresponding ANC strategy. If this replacement would not create a surface difference, I instead use another marker. Specifically, I use *-ge* for SENT, POSS-ACC, and NOMN, and *-ob* for ERG-POSS. An example is shown below:

(23) a. GOOD: *Lawyer claims witness sees attackca*

b. BAD: *Lawyer claims witnessge sees attackca*

The results are shown in Figure A.16. The mean accuracy across the 96 grammars is 83.9%, with a minimum of 13.3% and a maximum of 100%. The BAD parse rate ranges from 0% to 100%, with a mean of 30.8%.

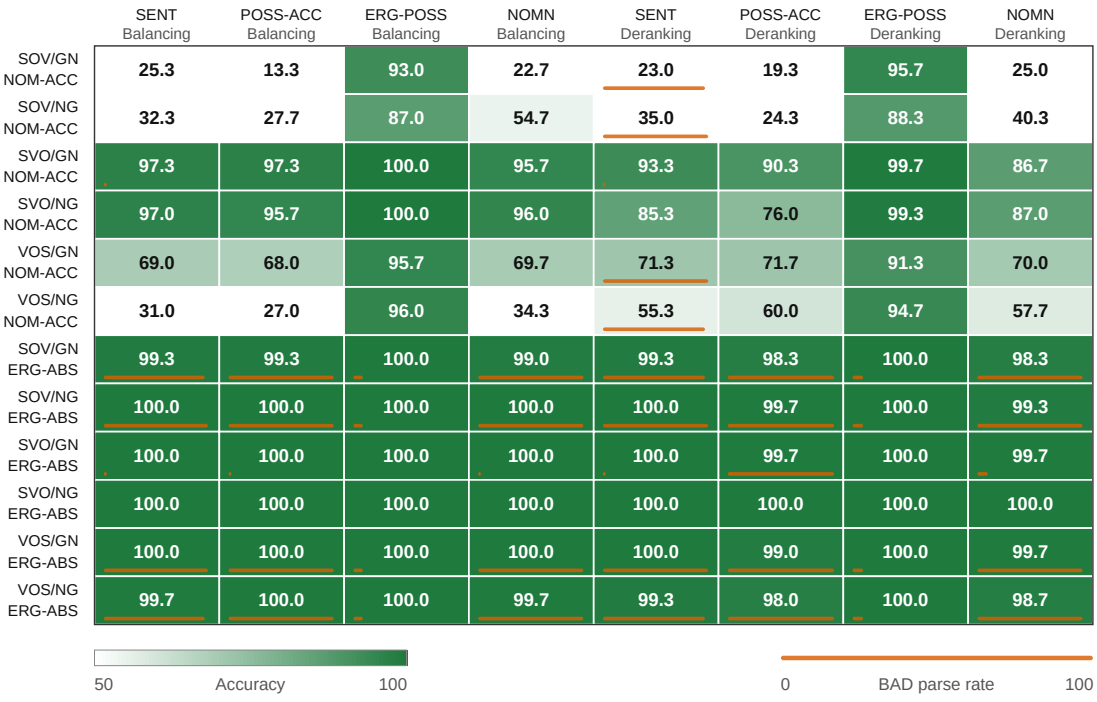

Figure A.16: Accuracy heatmap for 3.3 Clausal complement A marker.

Across typological parameters, the models learn this phenomenon well under ERG-ABS, with a mean accuracy of 99.7%, even though the mean BAD parse rate is 53.3%. By contrast,

NOM-ACC is much harder, with a mean accuracy of only 68.0%, even though the mean BAD parse rate is only 8.2%. This mismatch has the same source as in Section A.3.2.

Within NOM-ACC, two cases perform relatively well. The first is the ERG-POSS strategy, with a mean accuracy of 97.5%. The second is SVO word order, with a mean accuracy of 93.5%. Other NOM-ACC configurations are much weaker. In particular, SOV/GN + NOM-ACC has a mean accuracy of only 39.7%, while its BAD parse rate is only 12.5%.

The main difficulty in NOM-ACC comes from the zero-marking of embedded A. Under the perturbation rule, in SENT, POSS-ACC, and NOMN, the BAD sentence adds *-ge* to embedded A. Therefore, when embedded A and the following embedded P are adjacent, they can form a local genitive-like sequence on the surface. For example, under SOV/GN + NOM-ACC + Balancing + POSS-ACC, the GOOD sentence *Lawyer witness attackca sees claims* is perturbed into *Lawyer witnessge attackca sees claims*. The GOOD sentence roughly corresponds to "the lawyer claims that the witness sees the attack." In the BAD sentence, *witnessge attackca* is superficially close to an NP with a genitive possessor, roughly "the witness's attack," but this analysis cannot form a complete grammatical sentence because the head noun *attackca* should not bear the accusative marker *-ca*. Thus, although this BAD sentence cannot be parsed by the grammar, the model can still be misled by the local surface pattern.

The better performance of ERG-POSS and SVO under NOM-ACC supports this interpretation. Under ERG-POSS, the BAD marker is *-ob*, which cannot be confused with ordinary genitive marking. In SVO, the embedded order is A–V–P, so embedded A and embedded P are not adjacent.

A.3.4 Clausal Complement P Marker

This phenomenon tests whether the models learned the marker on embedded transitive P in clausal complements. The GOOD sentence is a finite matrix clause whose complement is an embedded transitive clause. In the source templates, matrix A, embedded A, and embedded P can each be either a simple NP or an NP with a genitive possessor.

In clausal complements, embedded argument marking follows the finite-clause system. Therefore, the embedded P marker is determined by alignment. In NOM-ACC, embedded P bears *-ca*; in ERG-ABS, embedded P is zero-marked. The BAD sentence is constructed by perturbing only the head of embedded P, replacing it with the P marker used under the corresponding ANC strategy. If this replacement would not create a surface difference, I instead use another marker that is ungrammatical in that position. Specifically, I use *-ge* for SENT, POSS-ACC, and ERG-POSS, and *-ob* for NOMN. An example is shown below:

(24) a. GOOD: *Guard believes visitor opens safeca*

b. BAD: *Guard believes visitor opens safeob*

Figure A.17: Accuracy heatmap for 3.4 Clausal complement P marker.

Figure A.17 shows that the models learned clausal complement P marking well overall. The mean accuracy across the 96 grammars is 97.7%, with a minimum of 79.7% and a

maximum of 100%, substantially higher than in Section A.3.3. The BAD parse rate ranges from 0% to 100%, with a mean of 15.6%.

The error pattern mirrors Clausal Complement A Marker in Section A.3.3. NOM-ACC configurations all reach 100% accuracy, while ERG-ABS is relatively weaker, with a mean accuracy of 95.4%. The weakest case is SOV/NG + ERG-ABS + Deranking + ERG-POSS, with a mean accuracy of 79.7%. For example, under SOV/NG + ERG-ABS + Deranking + ERG-POSS, the GOOD sentence *Guardca visitorca safe opening believes* is perturbed into *Guardca visitorca safege opening believes*. The reason for the errors here is the same as in the previous section. After adding *-ge*, the local sequence inside the embedded transitive clause becomes superficially similar to a possessor–head relation.

However, the models learn this phenomenon better than the previous one. A potential reason is that in Clausal Complement A Marker, the argument changed into a genitive-like form is embedded A, as in *witnessge attackca*. An animate A is semantically natural as a genitive possessor. By contrast, in the present phenomenon, the argument changed into a genitive-like form is embedded P, as in *safege opening*. Since embedded P is often an inanimate object, reinterpreting it as a genitive possessor is semantically more restricted, so the overall interference is weaker.

A.3.5 Clausal Complement–Matrix Verb Order

This phenomenon tests whether the models learned the relative order between the clausal complement as a whole and the matrix verb. The GOOD sentence is a finite matrix clause whose complement is an embedded intransitive or transitive clause. The source templates include both embedded IV and embedded TV complements. The BAD sentence is constructed by swapping the whole embedded clausal complement span with the matrix verb, while leaving the internal structure of the embedded clause unchanged. An example is shown below:

(25) a. GOOD: *Driver claims car stalls*

b. BAD: *Driver car stalls claims*

Figure A.18 shows that the models learned this phenomenon almost perfectly. The mean accuracy across the 96 grammars is 100.0%, with a minimum of 100.0% and a maximum of 100%. The BAD parse rate is also low, with a mean of 0.4% and a maximum of 4.0%.

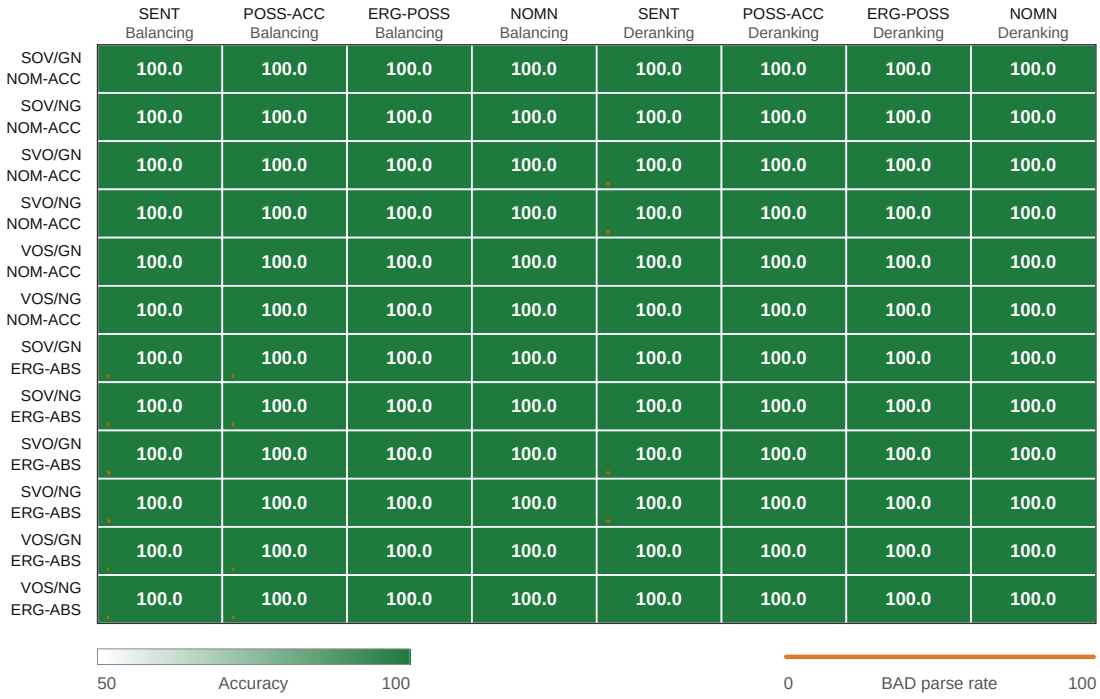

Figure A.18: Accuracy heatmap for 3.5 Clausal complement–V order.

A.3.6 Clausal Complement CV Valency

This phenomenon tests whether the models learned which matrix verbs can select a clausal complement. The GOOD sentence is a finite matrix clause whose matrix verb is a clausal-complement verb, and whose complement is an embedded intransitive or transitive clause. The source templates cover both embedded IV and embedded TV complements.

This phenomenon uses the MegaAcceptability dataset [White and Rawlins, 2020] to construct a valency pair table. The GOOD verb is a clausal-complement verb that is highly

acceptable in a clausal-complement frame in human judgments. The BAD verb is an ordinary transitive verb that is highly acceptable in a transitive frame, but has low acceptability in a clausal-complement frame. The BAD sentence is constructed by replacing the matrix verb stem. An example is shown below:

(26) a. GOOD: *Scientist speculates patient follows instructionca*

b. BAD: *Scientist names patient follows instructionca*

Figure A.19 shows that the mean accuracy across the 96 grammars is 74.9%, with a minimum of 70.3% and a maximum of 79.7%. The BAD parse rate is high, with a mean of 76.4% and a range of 75.0%–92.0%, showing that many BAD verbs also take clausal complements in the training corpus. The English baseline accuracy is 68.0%, suggesting that the synthetic training data provide a clearer valency signal than the original English corpus.

Figure A.19: Accuracy heatmap for 3.6 Clausal complement CV valency.

Compared with the other valency phenomena, 3.6 is close to Transitive V Valency in Section A.1.10 (76.2%) and higher than Intransitive V Valency in Section A.1.9 (70.4%). This suggests that the models learned the distinction between CVs and TVs reasonably well. Compared with Clausal Complement S Marker in Section A.3.2, the accuracy is also very similar: 75.2% versus 74.9%. This indirectly supports the earlier analysis of 3.2, where the models can use the distributional difference between clausal-complement verbs and ordinary transitive verbs. Overall, the models learned part of the CV/TV distributional distinction, but lexical valency remains a difficult phenomenon.

A.4 ANC Internal

A.4.1 ANC Verb Form

This phenomenon tests whether the models learned the verb form inside ANCs. The source templates are finite transitive clauses in which the ANC functions as either matrix A or matrix P. Since ANCs allow argument omission, the ANC itself covers bare IV ANCs, IV+S ANCs, bare TV ANCs, TV+A ANCs, TV+P ANCs, and TV+A+P ANCs.

In all target grammars, the ANC verb form is the nonfinite form *V-ing*. The BAD sentence is constructed by changing only the target ANC verb from *V-ing* to the finite form *V-s*. An example is shown below:

(27) a. GOOD: *Engineer checks techniciange measureingca*

b. BAD: *Engineer checks techniciange measuresca*

Figure A.20 shows that the models generally learned ANC verb form. The mean accuracy across the 96 grammars is 91.9%, with a minimum of 72.7% and a maximum of 100%. The BAD parse rate ranges from 10.0% to 41.0%, with a mean of 21.3%.

Across parameters, Deranking grammars perform better, with a mean accuracy of 94.6%, while Balancing grammars are weaker, at 89.3%. ANC strategy also has an effect. SENT is the weakest strategy, with a mean accuracy of 86.4%; POSS-ACC, ERG-POSS, and NOMN

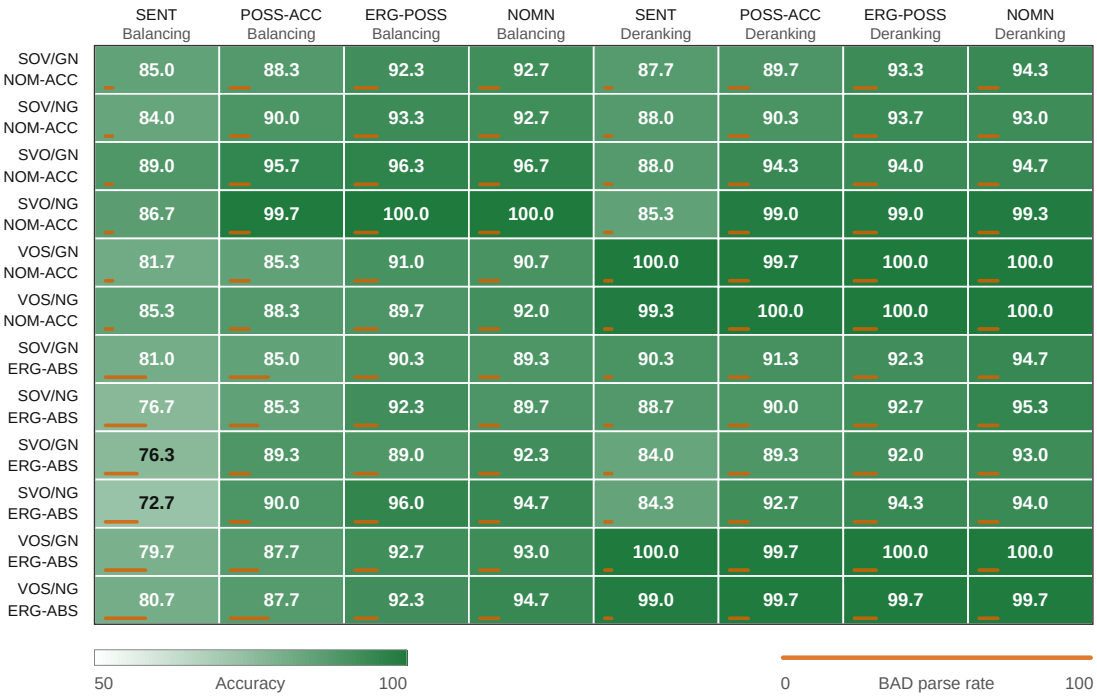

Figure A.20: Accuracy heatmap for 4.1 ANC verb form.

reach 92.0%, 94.4%, and 94.8%, respectively. The weakest group is Balancing + SENT, with a mean accuracy of 81.6%.

Overall, the models learned that verbs inside ANCs should take the nonfinite form. The main difficulty appears in the Balancing + SENT configuration. In these grammars, clausal complements use finite verb forms, and the SENT strategy makes ANCs more similar to clausal structures, so finite *V-s* can make the BAD sentence look more like a clausal complement. By contrast, Deranking grammars provide a clearer cue for distinguishing ANC verb form.

A.4.2 ANC S Marker

This phenomenon tests whether the models learned the marker on S inside intransitive ANCs. The GOOD sentence is a finite transitive clause in which matrix P is an intransitive ANC with an overt S.

In the target grammars, the ANC-internal S marker is determined by ANC strategy. Under SENT, ANC-internal S has zero marking. Under POSS-ACC, ERG-POSS, and NOMN, ANC-internal S receives the genitive-like marker *-ge*. The BAD sentence is constructed by changing only the marker on the ANC-internal S head, adding *-ge* under SENT and removing *-ge* under the other three strategies. An example is shown below:

(28) a. GOOD: *Farmer notices soilge contractingca*

b. BAD: *Farmer notices soil contractingca*

Figure A.21 shows the results for this phenomenon. The mean accuracy across the 96 grammars is 82.1%, with a minimum of 9.7% and a maximum of 100%. The BAD parse rate ranges from 1.0% to 100%, with a mean of 25.9%.

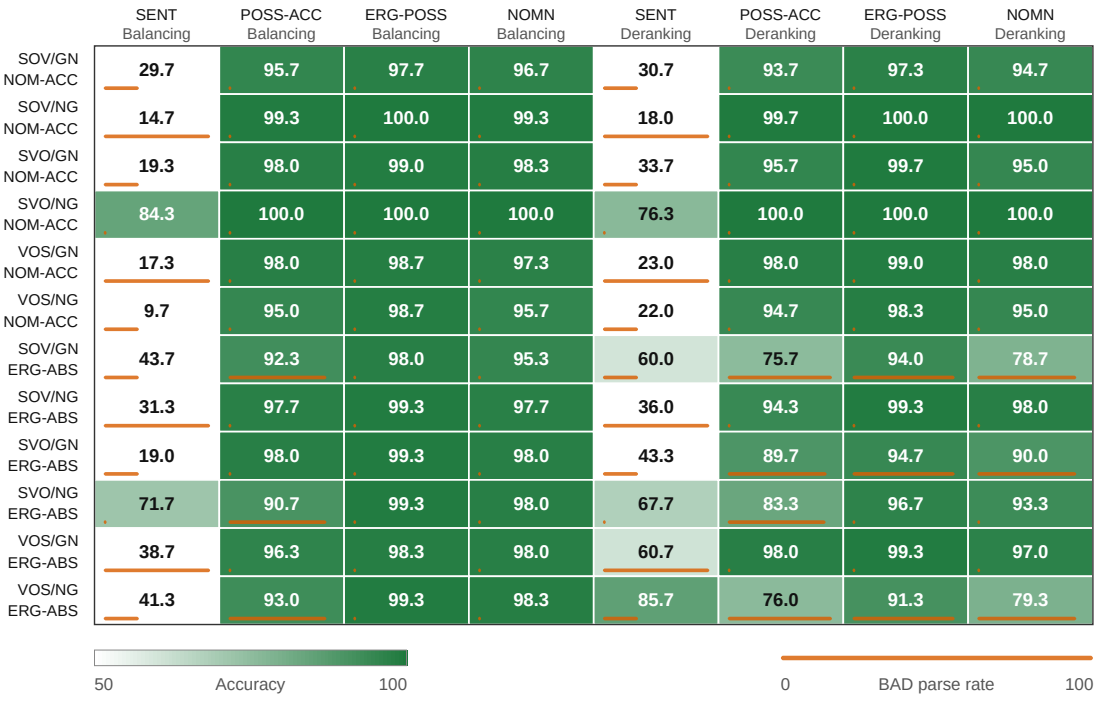

Figure A.21: Accuracy heatmap for 4.2 ANC S marker.

The main contrast comes from ANC strategy. SENT has a mean accuracy of only 40.7%, with a BAD parse rate of 50.2%. By contrast, POSS-ACC, ERG-POSS, and NOMN reach

mean accuracies of 93.9%, 98.2%, and 95.5%, with BAD parse rates of 28.1%, 12.9%, and 12.5%, respectively. The other parameters have smaller overall effects.

The low accuracy under SENT is mainly because ANC-internal S and finite-clause S have the same marking. In both environments, S has zero marking. Therefore, when the BAD sentence adds *-ge* to ANC-internal S, the surface form can instead resemble an ordinary genitive NP. For example, under SVO/GN + NOM-ACC + Balancing + SENT, the GOOD sentence *Host welcomes guest arriveingca* is perturbed into *Host welcomes guestge arriveingca*. The BAD sentence is roughly like "the host welcomes [the guest's arriving]", where *guest* is interpreted as the possessor of *arriveing*. This surface form is natural, but under the SENT strategy, ANC-internal S cannot take *-ge*, so the BAD sentence is ungrammatical in this grammar.

A.4.3 ANC A Marker

This phenomenon tests whether the models learned the marker on A inside transitive ANCs. The GOOD sentence is a finite transitive clause in which matrix P is a transitive ANC with overt A and overt P.

In the target grammars, the ANC-internal A marker is determined by ANC strategy and alignment. Under SENT, ANC-internal A takes the corresponding finite A marker. Under POSS-ACC and NOMN, ANC-internal A receives the genitive-like marker *-ge*. Under ERG-POSS, ANC-internal A receives *-ob*. The BAD sentence is constructed by changing only the marker on the ANC-internal A head. Under SENT, it is changed to *-ge*; under the other three strategies, it is changed back to the corresponding finite A marker. An example is shown below:

(29) a. GOOD: *Editor accepts translatorge renderingca titleob*

b. BAD: *Editor accepts translator renderingca titleob*

Figure A.22 shows the results for this phenomenon. The mean accuracy across the 96 grammars is 68.7%, with a minimum of 0.0% and a maximum of 100%. The BAD parse rate

ranges from 0.0% to 100%, with a mean of 20.7%.

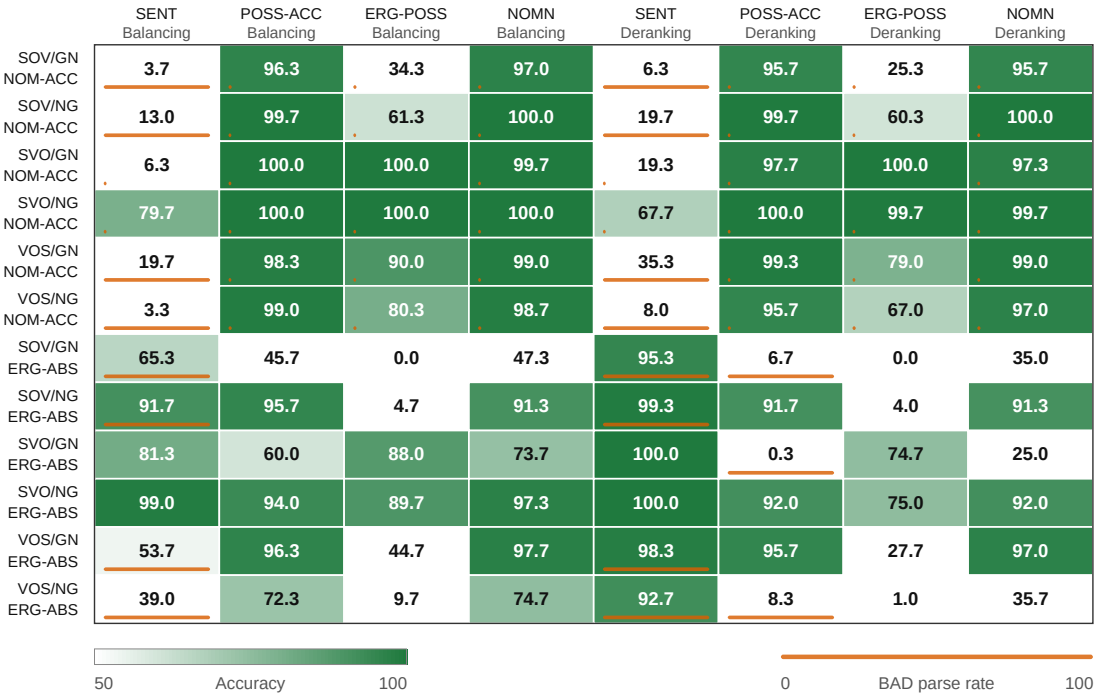

Figure A.22: Accuracy heatmap for 4.3 ANC A marker.

This phenomenon is harder than ANC S Marker. Across ANC strategies, POSS-ACC and NOMN perform relatively well, with mean accuracies of 80.8% and 85.0%, respectively. SENT and ERG-POSS are much weaker, at 54.1% and 54.8%. Alignment also has an effect, with NOM-ACC at 73.8% and ERG-ABS at 63.6%.

The errors fall into two main types. The first type appears under SENT + NOM-ACC. In these configurations, ANC-internal A has zero marking in the GOOD sentence, while the BAD sentence adds *-ge*. This makes the ANC portion of the BAD sentence resemble an ordinary genitive + *V-ing* nominal pattern, similar to ANC S Marker in Section A.4.2.

The second type appears under non-SENT strategies with ERG-ABS, especially ERG-POSS. ERG-POSS + ERG-ABS has a mean accuracy of only 34.9%, but its BAD parse rate is 0.0%. For example, under SOV/GN + ERG-ABS + Balancing + ERG-POSS, the GOOD sentence *Supervisorca traineeob tablege cleaning observes* is perturbed into *Supervi-*

sorca traineeca tablege cleaning observes. On the surface, *cleaning* can only be an ANC head in a Balancing grammar, and under ERG-POSS, ANC-internal A should take *-ob*, not *-ca*. Therefore, this type of BAD sentence cannot be parsed by the grammar and does not have an obvious alternative structural analysis. It remains unclear why the model prefers this form.

A.4.4 ANC P Marker

This phenomenon tests whether the models learned the marker on P inside transitive ANCs. The GOOD sentence is a finite transitive clause in which matrix P is a transitive ANC with overt A and overt P.

In the target grammars, the ANC-internal P marker is determined by ANC strategy and alignment. Under SENT and POSS-ACC, ANC-internal P takes the corresponding finite P marker. Under ERG-POSS, ANC-internal P receives the genitive-like marker *-ge*. Under NOMN, ANC-internal P receives *-ob*. The BAD sentence is constructed by changing only the marker on the ANC-internal P head. Under SENT and POSS-ACC, it is changed to *-ge*; under ERG-POSS and NOMN, it is changed back to the corresponding finite P marker. An example is shown below:

(30) a. GOOD: *Reporter notes scientistge measureingca sampleob*

b. BAD: *Reporter notes scientistge measureingca sampleca*

Figure A.23 shows the results for this phenomenon. The mean accuracy across the 96 grammars is 90.1%, with a minimum of 27.3% and a maximum of 100%. The BAD parse rate ranges from 0.0% to 100%, with a mean of 31.8%.

This phenomenon is clearly easier than ANC A Marker. ERG-POSS and NOMN perform best, with mean accuracies of 97.6% and 96.2%, respectively. SENT is also relatively high, at 86.0%. POSS-ACC is the weakest strategy, with a mean accuracy of 80.5%. Across alignment, NOM-ACC reaches 96.4%, compared with 83.8% for ERG-ABS.

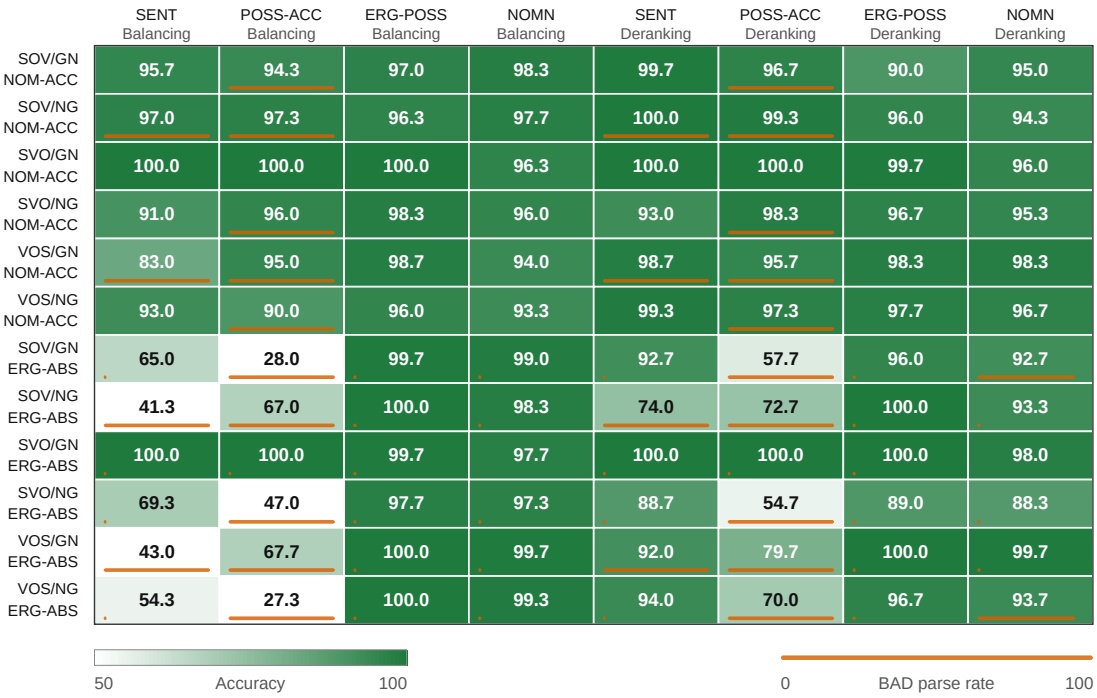

Figure A.23: Accuracy heatmap for 4.4 ANC P marker.

The main errors are concentrated in POSS-ACC + ERG-ABS. This combination has a mean accuracy of only 64.3%, with a BAD parse rate of 83.4%. In these configurations, ANC-internal P and finite P are both zero-marked, and the BAD sentence adds *-ge* to P. For example, under SOV/GN + ERG-ABS + Balancing + POSS-ACC, the GOOD sentence *Reporterca scientistge sample measureing notes* is perturbed into *Reporterca scientistge samplege measureing notes*. The GOOD sentence roughly corresponds to "the reporter notes [the scientist's measuring of the sample]." The BAD sentence is closer to "the reporter notes [the measuring of the scientist's sample]", where *scientistge samplege* forms a genitive-like nominal sequence. This BAD sentence can be parsed by the grammar, and the alternative interpretation is also semantically natural because the original A can naturally be interpreted as the possessor of the original P.

A.4.5 ANC S–V Order

This phenomenon tests whether the models learned the relative order between overt S and the ANC verb inside intransitive ANCs. The GOOD sentence is a finite transitive clause in which matrix P is an intransitive ANC with an overt S.

In the target grammars, ANC-internal S–V order is determined by the effective ANC intransitive order. As shown in Table 3.6, this order is determined jointly by clause word order, NP word order, and ANC strategy. The BAD sentence is constructed by swapping the ANC-internal S token and the ANC verb token. An example is shown below:

(31) a. GOOD: *Farmer notices soilge contractingca*

b. BAD: *Farmer notices contractingca soilge*

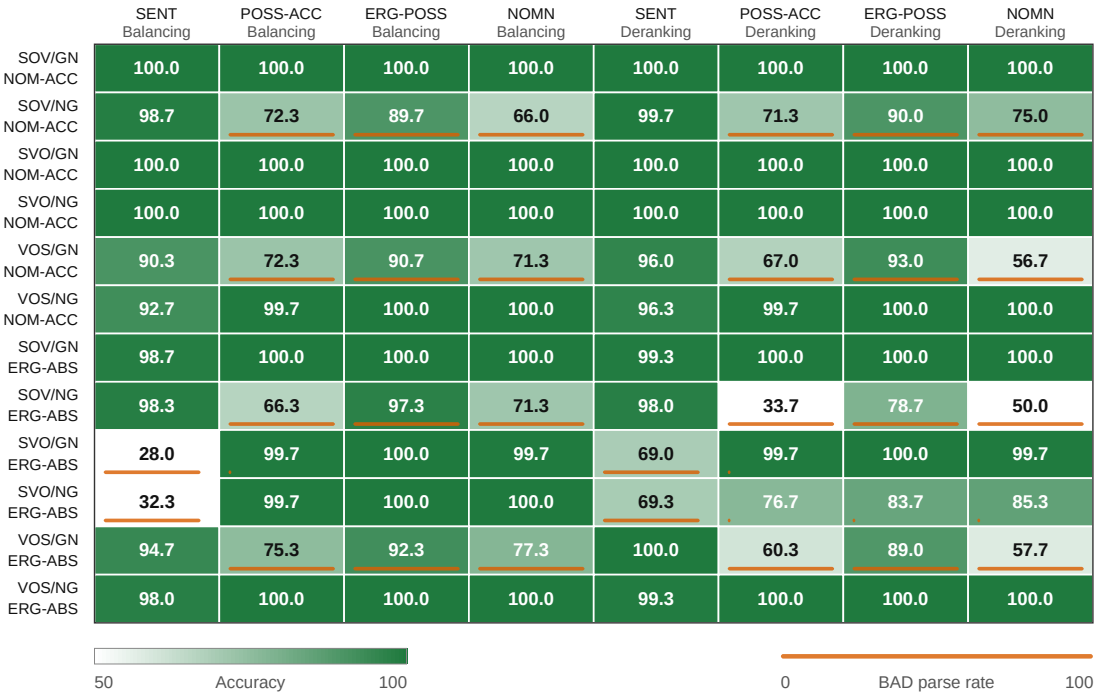

Figure A.24: Accuracy heatmap for 4.5 ANC S–V order.

Figure A.24 shows that the models generally learned this phenomenon well. The mean

accuracy across the 96 grammars is 90.3%, with a minimum of 28.0% and a maximum of 100%. The BAD parse rate ranges from 0.0% to 100%, with a mean of 28.9%.

The main problems fall into two types. The first is the combination of non-harmonic word order and non-SENT strategies, especially POSS-ACC and NOMN. In SOV/NG, the non-SENT strategies have a mean accuracy of 71.8%, with a BAD parse rate of 100%. In VOS/GN, the non-SENT strategies have a mean accuracy of 75.2%, also with a BAD parse rate of 100%. By contrast, the SENT strategy performs much better in these two profiles. SOV/NG + SENT has a mean accuracy of 98.7%, and VOS/GN + SENT has a mean accuracy of 95.2%; both have a BAD parse rate of 0%.

This error pattern follows from how ANC-internal order is defined. For non-SENT strategies, ANCs have both clausal and nominal properties, so their internal order is determined jointly by clause word order and NP word order. In the two non-harmonic profiles, SOV/NG and VOS/GN, the two orders point in different directions. In these cases, the experimental grammars align non-SENT ANCs primarily with NP word order. As a result, ANC-internal order can depart from the more familiar linear patterns of finite clauses.

The second difficulty appears under SVO + SENT, especially with ERG-ABS alignment. In the SVO + SENT + ERG-ABS configurations, the mean accuracy is only 49.7%, while the BAD parse rate is 91.0%.

This problem comes from ANC argument omission. Since arguments can be omitted inside ANCs, if an SVO-like ANC contains only one overt argument, that argument can sometimes appear on either side of the verb. For verbs with transitive/intransitive alternations, if the overt argument can naturally be interpreted either as S or as P, both the GOOD and BAD sentences may receive grammatical and semantically natural analyses. For example, the GOOD sentence *Scientistca observes cloud forming* is perturbed into *Scientistca observes forming cloud*. The GOOD sentence roughly means "the scientist observes [the cloud forming]", while the BAD sentence can be reanalyzed as "the scientist observes [the forming of the cloud]." Thus, although a high BAD parse rate does not necessarily mean that the model failed to learn the grammar, it shows that ANC-internal order is less stable

in this configuration.

A.4.6 ANC A–V Order

This phenomenon tests whether the models learned the relative order between A and the ANC verb inside transitive ANCs. The GOOD sentence is a finite transitive clause in which matrix P is a transitive ANC with overt A and overt P.

In the target grammars, ANC-internal A–V order is determined by the effective ANC transitive order. As shown in Table 3.6, this order is determined jointly by clause word order, NP word order, and ANC strategy. The BAD sentence is constructed by swapping the ANC-internal A span with the chunk containing the ANC verb and P. An example is shown below:

(32) a. GOOD: *Nurse reports patientge takingca medicineob*

b. BAD: *Nurse reports takingca medicineob patientge*

Figure A.25 shows that the models generally learned this phenomenon well. The mean accuracy across the 96 grammars is 93.1%, with a minimum of 57.7% and a maximum of 100%. The BAD parse rate ranges from 0.0% to 100%, with a mean of 16.7%.

The main source of errors is similar to ANC S–V Order in Section A.4.5: ANC-internal A–V order can conflict with finite-clause A–V order. The affected combinations are SVO/GN + ERG-POSS, VOS/GN + POSS-ACC/NOMN, and SOV/NG + POSS-ACC/NOMN, with mean accuracies of 77.8%, 64.2%/68.9%, and 80.7%/77.8%, respectively. By contrast, the remaining combinations have a mean accuracy of 98.1%.

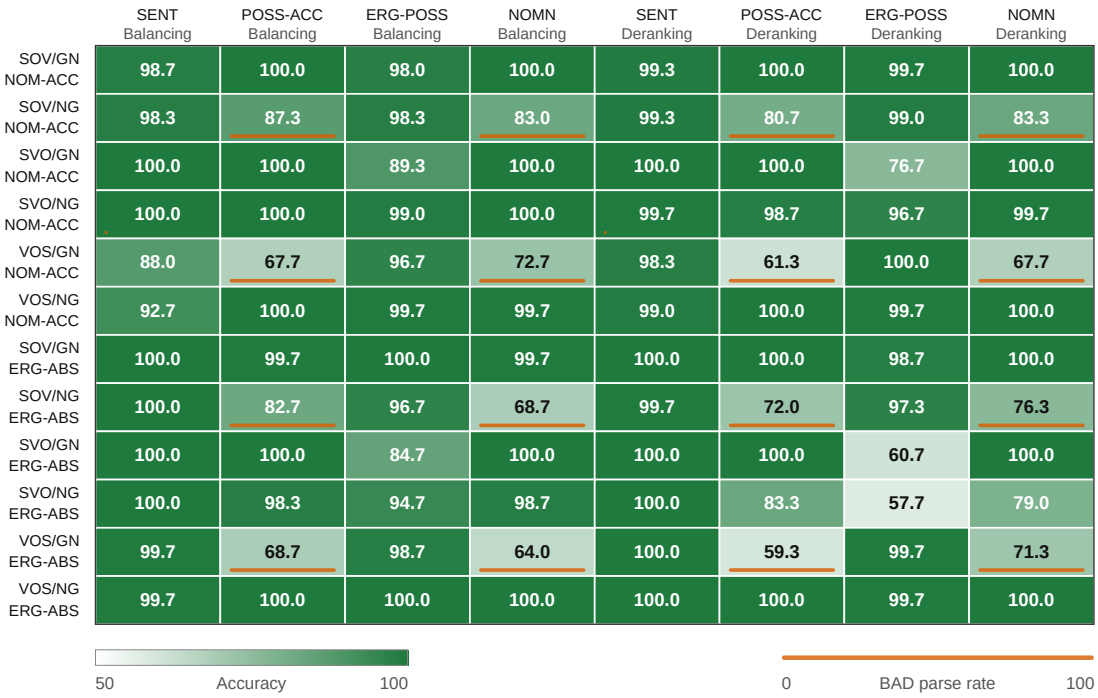

Figure A.25: Accuracy heatmap for 4.6 ANC A–V order.

A.4.7 ANC V–P Order

This phenomenon tests whether the models learned the relative order between the ANC verb and P inside transitive ANCs. The GOOD sentence is a finite transitive clause in which matrix P is a transitive ANC with overt A and overt P.

In the target grammars, ANC-internal V–P order is determined by the effective ANC transitive order, as shown in Table 3.6. The BAD sentence is constructed by swapping the ANC verb token and the ANC-internal P head. An example is shown below:

(33) a. GOOD: *Supervisor observes traineege cleaning tableob*

b. BAD: *Supervisor observes traineege tableob cleaning*

Figure A.26 shows that the models generally learned this phenomenon well. The mean accuracy across the 96 grammars is 92.9%, with a minimum of 51.3% and a maximum of 100%. The BAD parse rate ranges from 0.0% to 100%, with a mean of 13.9%.

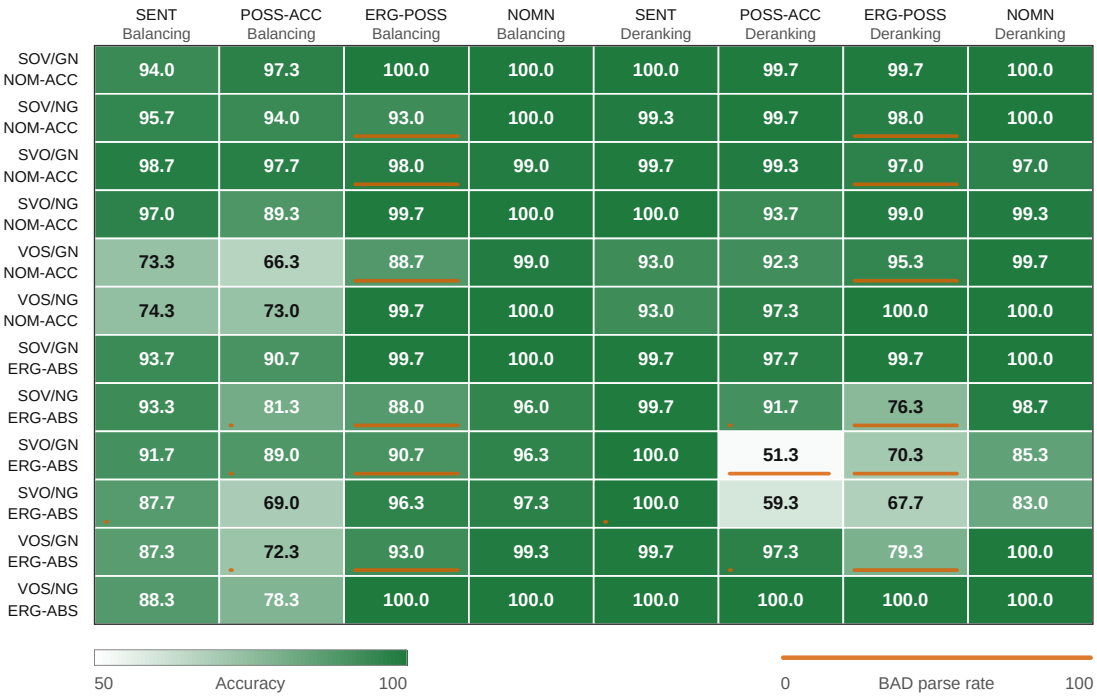

Figure A.26: Accuracy heatmap for 4.7 ANC V–P order.

The main weakness of this phenomenon appears in SVO profiles with Deranking, ERG-ABS, and non-SENT strategies. Under these configurations, POSS-ACC, ERG-POSS, and NOMN have mean accuracies of 55.3%, 69.0%, and 84.2%, respectively.

These errors mainly come from alternative parses. For example, under SVO/GN + ERG-ABS + Deranking + POSS-ACC, the GOOD sentence *Supervisorca observes traineege cleaning table* is perturbed into *Supervisorca observes traineege table cleaning*. The GOOD sentence roughly corresponds to "the supervisor observes [the trainee's cleaning of the table]." In the BAD sentence, *traineege table* can be reanalyzed as a genitive NP functioning as the overt S of an intransitive ANC, roughly corresponding to "the supervisor observes [the trainee's table cleaning]."

A.5 ANC External

A.5.1 ANC External S Marker

This phenomenon tests whether the models learned the external marker on the ANC head when the whole ANC functions as matrix intransitive S. The GOOD sentence is a finite intransitive clause in which matrix S is an ANC. The source templates cover several ANC-internal structures, including bare IV ANCs, IV+S ANCs, bare TV ANCs, TV+A ANCs, TV+P ANCs, and TV+A+P ANCs.

In all target grammars, the finite S marker is zero. Therefore, when the whole ANC functions as matrix S, the ANC head in the GOOD sentence should not receive an external overt marker. The BAD sentence is constructed by adding *-ca* or *-ge* to the ANC head. An example is shown below:

(34) a. GOOD: *Adjusting seatob succeeds*

b. BAD: *Adjustingge seatob succeeds*

Figure A.27 shows that this phenomenon is difficult for the models. The mean accuracy across the 96 grammars is 62.9%, with a minimum of 26.0% and a maximum of 99.0%. The BAD parse rate ranges from 0.0% to 55.0%, with a mean of 7.1%.

The configurations with relatively strong performance are mainly SVO/NOM-ACC/non-SENT, with a mean accuracy of 89.0%. By contrast, the remaining configurations have a mean accuracy of only 59.1%. This shows that, in most configurations, the models did not robustly learn that the ANC head should not take an external marker when the ANC functions as matrix S.

The reason for these errors is not yet fully clear. For example, under VOS/NG + NOM-ACC + Balancing + ERG-POSS, the GOOD sentence *Succeeds adjusting seatge* is perturbed into *Succeeds adjustingge seatge*. This BAD sentence contains two *-ge* markers and does not seem to have any possible alternative parse, but the model still prefers the BAD sentence.

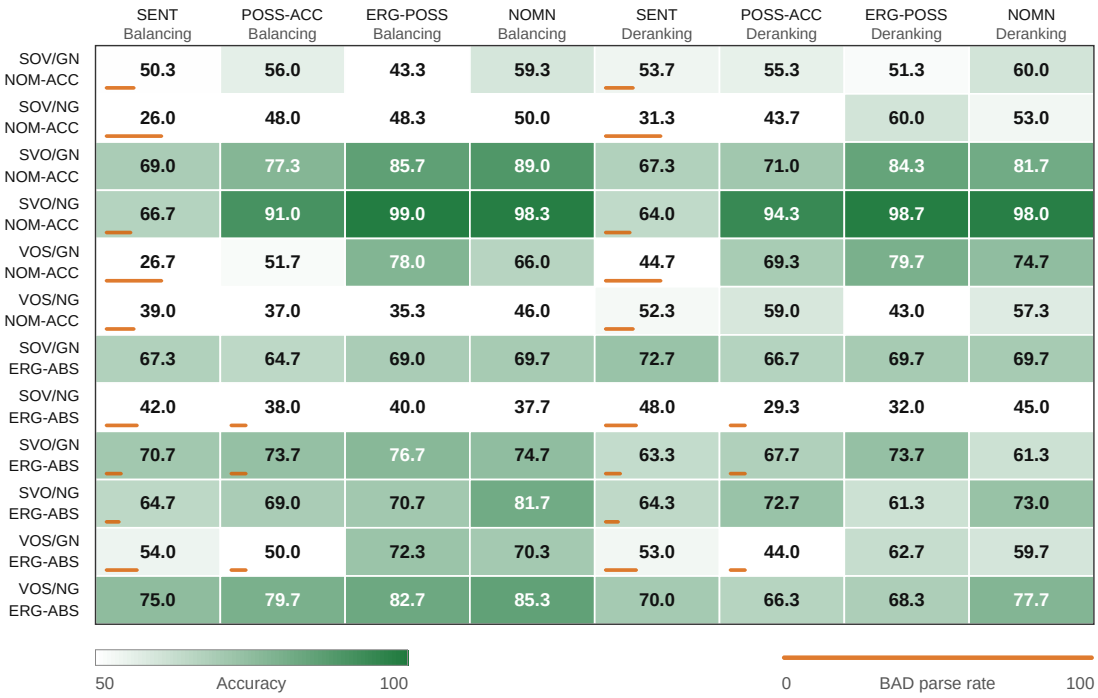

Figure A.27: Accuracy heatmap for 5.1 ANC external S marker.

A.5.2 ANC External A Marker

This phenomenon tests whether the models learned the external marker on the ANC head when the whole ANC functions as matrix transitive A. The GOOD sentence is a finite transitive clause in which matrix A is an ANC and matrix P is an ordinary NP. The source templates cover several ANC-internal structures, including bare IV ANCs, IV+S ANCs, bare TV ANCs, TV+A ANCs, TV+P ANCs, and TV+A+P ANCs.

In the target grammars, when an ANC functions as matrix A, its external A marker is the same as the finite A marker. Under NOM-ACC, A is zero-marked; under ERG-ABS, A receives *-ca*. The BAD sentence changes only the external A marker on the ANC head: under NOM-ACC, it adds *-ca* or *-ge*; under ERG-ABS, it deletes *-ca* or replaces it with *-ge*. An example is shown below:

(35) a. GOOD: *Chefge cooking delights guestca*

b. BAD: *Chefge cookingca delights guestca*

The results are shown in Figure A.28. The mean accuracy across the 96 grammars is 82.7%, with a minimum of 29.0% and a maximum of 100%. The BAD parse rate ranges from 0.0% to 50.0%, with a mean of 11.6%.

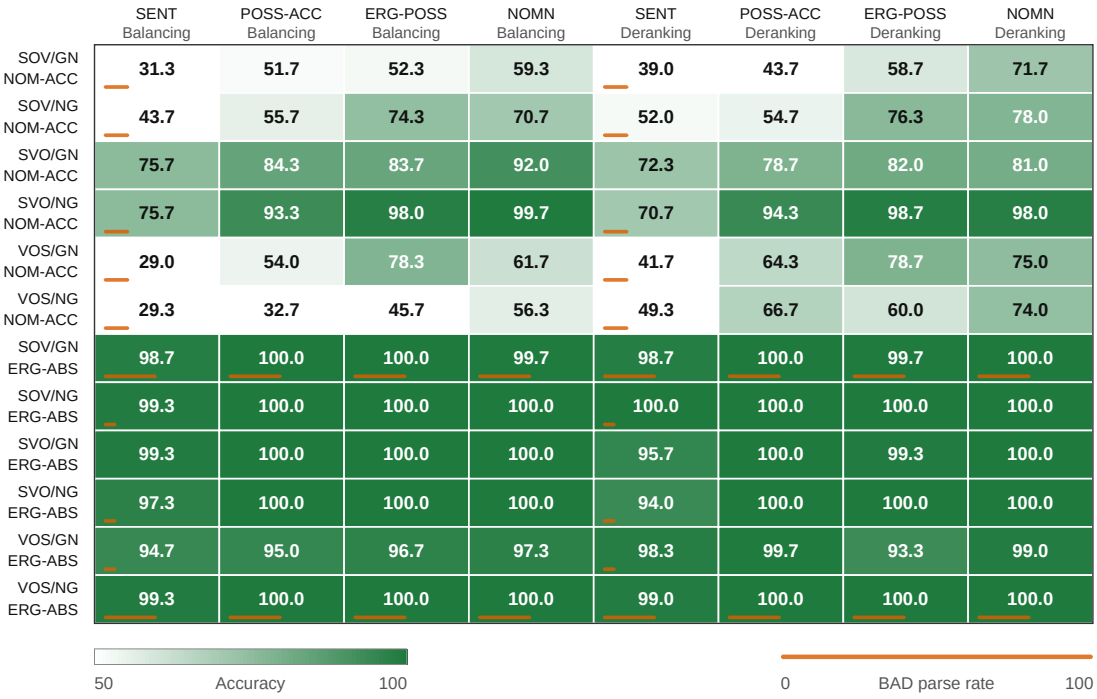

Figure A.28: Accuracy heatmap for 5.2 ANC external A marker.

This phenomenon shows the same pattern as Transitive A Marker in Section A.1.5. ERG-ABS configurations perform well, with a mean accuracy of 99.0%. By contrast, NOM-ACC configurations are much harder, with a mean accuracy of only 66.4%. This shows that the models easily learned the overt external A marker *-ca*, but did not robustly learn the zero marking of ANC external A under NOM-ACC.

A.5.3 ANC External P Marker

This phenomenon tests whether the models learned the external marker on the ANC head when the whole ANC functions as matrix transitive P. The GOOD sentence is a finite transitive clause in which matrix A is an ordinary NP and matrix P is an ANC. The source templates cover several ANC-internal structures, including bare IV ANCs, IV+S ANCs, bare TV ANCs, TV+A ANCs, TV+P ANCs, and TV+A+P ANCs.

In the target grammars, when an ANC functions as matrix P, its external P marker is the same as the finite P marker. Under NOM-ACC, P receives *-ca*; under ERG-ABS, P is zero-marked. The BAD sentence changes only the external P marker on the ANC head: under NOM-ACC, it deletes *-ca* or replaces it with *-ge*; under ERG-ABS, it adds *-ca* or *-ge*. An example is shown below:

(36) a. GOOD: *Team praises playerge improveingca*

b. BAD: *Team praises playerge improveing*

The results are shown in Figure A.29. The mean accuracy across the 96 grammars is 91.4%, with a minimum of 33.7% and a maximum of 100%. The BAD parse rate ranges from 0.0% to 99.0%, with a mean of 15.8%.

This phenomenon shows the same pattern as Transitive P Marker in Section A.1.6. NOM-ACC configurations perform well, with a mean accuracy of 98.4%. By contrast, ERG-ABS configurations are harder, with a mean accuracy of 84.4%. As in ANC External A Marker in Section A.5.2, the models learn overt marking better than zero marking here as well. Within the ERG-ABS configurations, the especially difficult case is SOV/NG + ERG-ABS, with a mean accuracy of only 55.8%. In this configuration, the ANC head is separated from the matrix verb by ANC-internal arguments, so the erroneous external P marker is farther from the main-clause verb and becomes harder for the model to identify as a matrix-level marking error.

Figure A.29: Accuracy heatmap for 5.3 ANC external P marker.

A.5.4 ANC External Genitive Marker

This phenomenon tests whether the models learned the external genitive marker on the ANC head when the whole ANC functions as a genitive possessor. The GOOD sentence is a finite clause in which the possessor inside a genitive NP is an ANC. The source templates cover four ANC-internal structures: bare IV ANCs, IV+S ANCs, bare TV ANCs, and TV+A ANCs. TV+P and TV+A+P ANCs are not included in this test because the corresponding English source sentences cannot be parsed reliably. For example, in a structure such as *the reading of the note's effect*, the possessive tends to be analyzed as modifying the internal P of the ANC rather than the whole ANC. Such expressions are also relatively unnatural in English.

In all target grammars, the genitive marker is *-ge*. Therefore, when an ANC functions as a genitive possessor, the ANC head must carry the external genitive marker *-ge*. The BAD

sentence is constructed by changing the external genitive marker on the ANC head, either deleting *-ge* or replacing it with *-ca*. An example is shown below:

(37) a. GOOD: *Selectingge size hurts visitorca*

b. BAD: *Selectingca size hurts visitorca*

The results are shown in Figure A.30. The mean accuracy across the 96 grammars is 75.9%, with a minimum of 34.0% and a maximum of 100%. The BAD parse rate is low, with a mean of 4.2%.

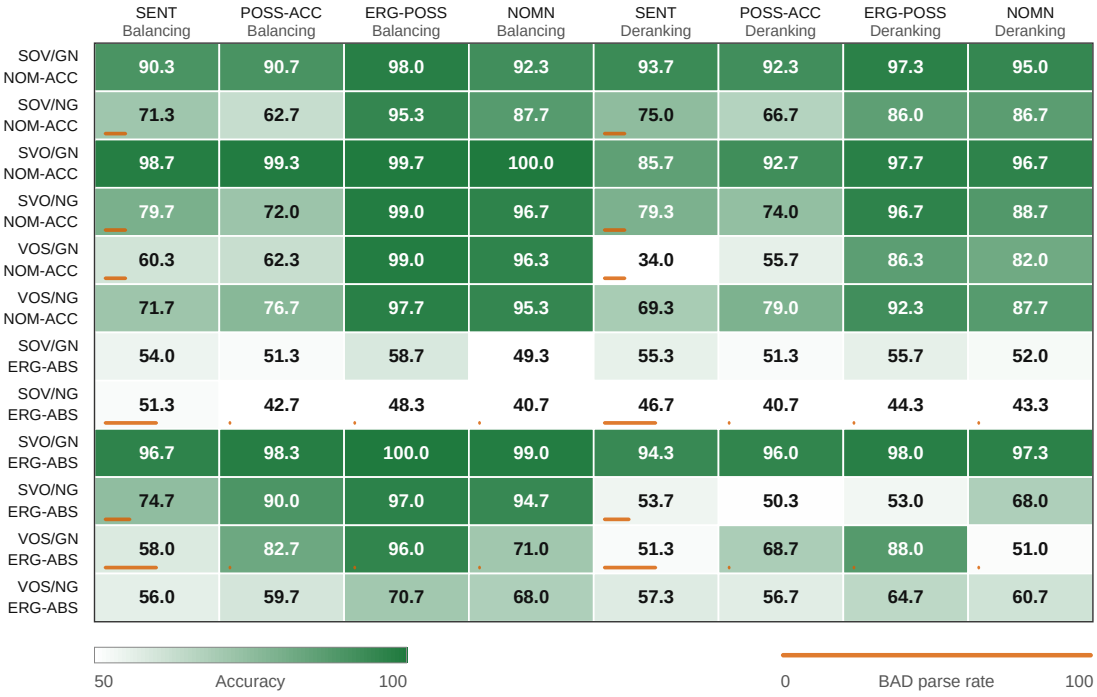

Figure A.30: Accuracy heatmap for 5.4 ANC external genitive marker.

Across parameters, Balancing and Deranking have relatively small effects, with mean accuracies of 79.2% and 72.7%, respectively. By contrast, strategy, word order, and alignment have clearer effects. ERG-POSS and NOMN perform relatively well, with mean accuracies of 84.1% and 79.2%, while SENT and POSS-ACC are weaker, at 69.1% and 71.3%. In addition, SOV + ERG-ABS configurations are relatively weak, with a mean accuracy of 49.1%.

The reason for this complex pattern is not yet clear. One possibility is distributional imbalance in the training corpus. In the classified source sentences, most ANC occurrences appear as matrix P: 72.4% are P arguments, 19.9% are A arguments, 7.4% are S arguments, and only 0.3% are genitive possessors. As a result, the models rarely see ANCs in genitive possessor environments, making generalization less stable.

A.6 ANC Argument Omission

A.6.1 ANC Omitted S

This phenomenon tests whether the models distinguish bare intransitive ANCs from finite intransitive clauses. Bare intransitive ANCs allow the S argument to be omitted, whereas finite clauses require an overt S. The GOOD sentence is a finite transitive clause in which matrix P is an intransitive ANC without an overt S.

In all target grammars, the ANC verb uses the nonfinite form *V-ing*. The BAD sentence is constructed by replacing the bare IV ANC head with the corresponding finite intransitive predicate. An example is shown below:

(38) a. GOOD: *Company questions disappearingca*

b. BAD: *Company questions disappears*

The results are shown in Figure A.31. The mean accuracy across the 96 grammars is 98.5%, with a minimum of 91.7% and a maximum of 100%. The BAD parse rate is 11.0%. Overall, the models robustly learned that bare intransitive ANCs can function as matrix arguments and that the ANC head should take the nonfinite form.

Figure A.31: Accuracy heatmap for 6.1 ANC omitted S.

A.6.2 ANC Omitted A

This phenomenon tests whether the models distinguish transitive ANCs with omitted A from ordinary finite transitive clauses. A transitive ANC with omitted A can retain only the P argument, whereas a finite transitive clause requires both A and P. The GOOD sentence is a finite transitive clause in which matrix P is a transitive ANC with overt P and omitted A.

In the target grammars, the ANC verb uses the nonfinite form *V-ing*, and ANC-internal P receives the P marker determined by the ANC strategy. The BAD sentence is constructed by replacing the ANC head with the corresponding finite transitive predicate and changing the ANC-internal P marker to the finite P marker. An example is shown below:

(39) a. GOOD: *They questions countingca ballotob*

b. BAD: *They questions counts ballotca*

The results are shown in Figure A.32. The mean accuracy across the 96 grammars is

75.8%, with a minimum of 18.3% and a maximum of 100%. The BAD parse rate ranges from 0.0% to 100%, with a mean of 16.1%.

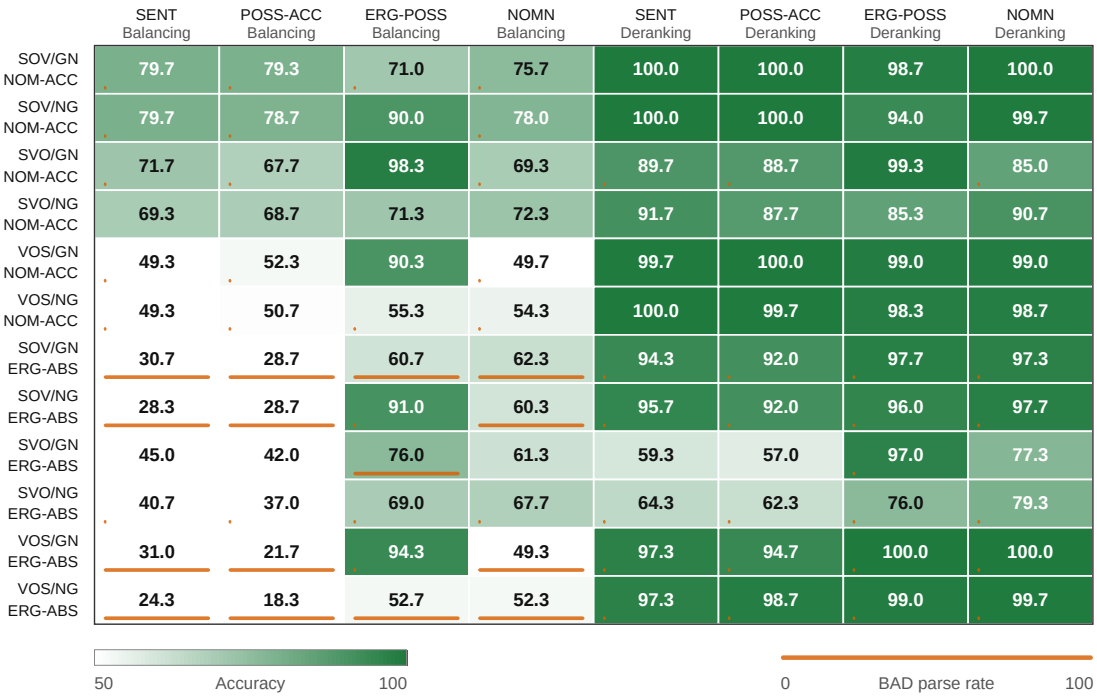

Figure A.32: Accuracy heatmap for 6.2 ANC omitted A.

Across parameters, Deranking is much easier than Balancing. Deranking grammars have a mean accuracy of 92.2%, while Balancing grammars have a mean accuracy of only 59.3%. Within Balancing grammars, ERG-ABS is especially difficult, with a mean accuracy of 48.9% and a BAD parse rate of 62.8%. By contrast, NOM-ACC + Balancing reaches 69.7% accuracy, with a BAD parse rate of only 0.7%.

The main errors come from BAD parses involving transitive/intransitive alternation. A BAD sentence missing one transitive argument can be reanalyzed as an intransitive structure with exactly one overt argument.

There are also errors that are not due to BAD parses. For example, under SVO/NG + ERG-ABS + Balancing + POSS-ACC, the accuracy is only 37.0%, but the BAD parse rate is only 1.0%. The GOOD sentence *Theyca questions counting ballot* is perturbed into

Theyca questions counts ballot. This BAD sentence usually cannot be parsed, because the finite transitive predicate *counts ballot* lacks an A argument. Still, its surface form resembles a clausal-complement structure with a matrix clausal-complement verb followed by an embedded finite predicate.

A.6.3 ANC Omitted P

This phenomenon tests whether the models distinguish transitive ANCs with omitted P from ordinary finite transitive clauses. A transitive ANC with omitted P can retain only the A argument, whereas a finite transitive clause requires both A and P. The GOOD sentence is a finite transitive clause in which matrix P is a transitive ANC with overt A and omitted P.

In the target grammars, the ANC verb uses the nonfinite form *V-ing*, and ANC-internal A receives the A marker determined by the ANC strategy. The BAD sentence is constructed by replacing the ANC head with the corresponding finite transitive predicate and changing the ANC-internal A marker to the finite A marker. An example is shown below:

(40) a. GOOD: *Officer notices driverge checkingca*

b. BAD: *Officer notices driver checks*

The results are shown in Figure A.33. The mean accuracy across the 96 grammars is 83.9%, with a minimum of 17.3% and a maximum of 100%. The BAD parse rate ranges from 0.0% to 100%, with a mean of 16.7%.

This phenomenon shows a pattern similar to ANC Omitted A in Section A.6.2. Deranking grammars still perform much better than Balancing grammars, with mean accuracies of 92.5% and 75.3%, respectively. Alignment also matters: NOM-ACC reaches 93.9%, compared with 73.8% for ERG-ABS.

Although the overall pattern is similar, accuracy is higher in ANC Omitted P: 83.9% compared with 75.8%. The likely reason is the direction of argument reanalysis. In ANC Omitted A, BAD parses often require the original P to be reanalyzed as intransitive S. In ANC Omitted P, BAD parses instead often reanalyze the original A as intransitive S. Since A

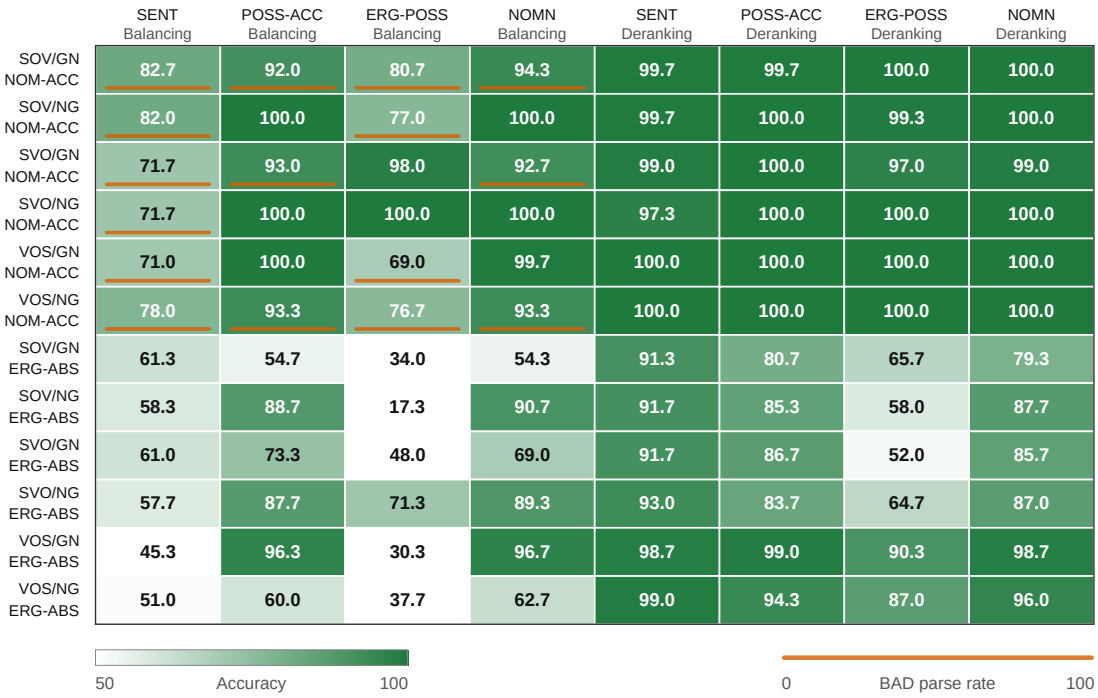

Figure A.33: Accuracy heatmap for 6.3 ANC omitted P.

arguments are usually more natural as S arguments, this reanalysis is semantically smoother, and the models perform better overall.

A.6.4 ANC Omitted A+P

This phenomenon tests whether the models distinguish bare transitive ANCs from ordinary finite transitive clauses. Transitive ANCs can omit internal arguments, whereas finite transitive clauses require both A and P. The GOOD sentence is a finite transitive clause in which matrix P is a bare transitive ANC.

In all target grammars, the ANC verb uses the nonfinite form *V-ing*. The BAD sentence is constructed by replacing the bare TV ANC head with the corresponding finite transitive predicate. An example is shown below:

(41) a. GOOD: *Manager watchs cleaningca*

b. BAD: *Manager watchs cleans*

The results are shown in Figure A.34. The mean accuracy across the 96 grammars is 90.5%, with a minimum of 73.3% and a maximum of 100%. The BAD parse rate ranges from 0.0% to 52.0%, with a mean of 25.8%.

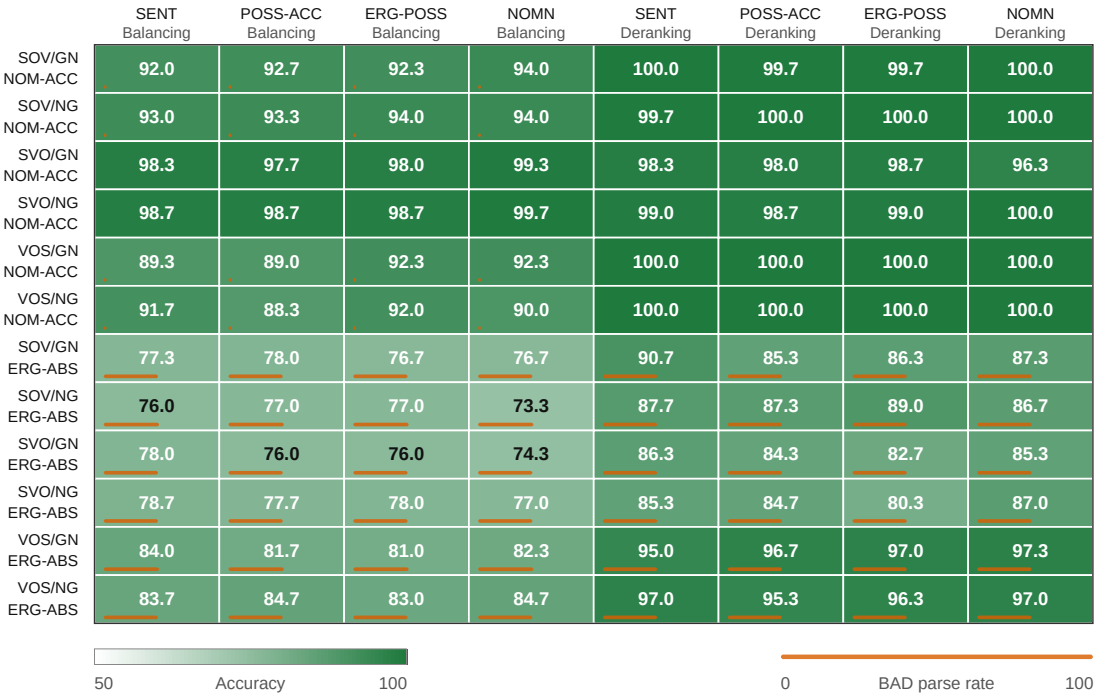

Figure A.34: Accuracy heatmap for 6.4 ANC omitted A+P.

Overall, the models learned this phenomenon well. As in the previous two omitted-argument phenomena, Deranking grammars perform better than Balancing grammars, with mean accuracies of 94.5% and 86.5%, respectively. The weakest configuration is Balancing + ERG-ABS, with a mean accuracy of 78.9%. This phenomenon is easier than ANC Omitted A in Section A.6.2 and ANC Omitted P in Section A.6.3, because the BAD sentence lacks both arguments of the finite transitive predicate, making the violation more salient.

Appendix B

MULTI-SEED VALIDATION

To check whether the evaluation results are stable across random seeds, I compare the accuracies of the three models within each phenomenon–grammar condition.

Since the evaluation measures whether a model assigns a higher probability to the grammatical member of each minimal pair, chance-level accuracy is 50%. I therefore use a softplus transformation to map accuracy to a 0–1 learning score L, so that differences near or below chance are compressed, while differences above chance are more salient. For each phenomenon–grammar condition, I take the difference between the highest and lowest learning scores across the three seeds and define it as seed disagreement D. L_s and D are defined as follows:

$$L_s = \frac{\text{softplus}\left(\frac{a_s - 50}{\tau}\right) - \text{softplus}\left(\frac{-50}{\tau}\right)}{\text{softplus}\left(\frac{50}{\tau}\right) - \text{softplus}\left(\frac{-50}{\tau}\right)}$$

$$D = \max_s L_s - \min_s L_s$$

where a_s is the accuracy percentage for seed s, $\tau = 5$, and $\text{softplus}(x) = \log(1 + e^x)$.

Overall, the three-seed results are stable. Across the 34×96 phenomenon–grammar conditions, the mean seed disagreement is 0.061, the median is 0.020, and the 90th percentile is 0.160. This indicates that most phenomenon–grammar conditions do not depend strongly on a single random seed. The few phenomena with higher disagreement are concentrated among phenomena with relatively low overall accuracy. For example, Clausal Complement S Marker in Section A.3.2 has a mean accuracy of 75.2% and a mean seed disagreement of 0.177, while ANC External Genitive Marker in Section A.5.4 has a mean accuracy of 75.9%

and a mean seed disagreement of 0.141. Thus, variation across random seeds does not change the main conclusions of the thesis.