Systematicity between Forms and Meanings across Languages Supports Efficient Communication

Doreen Osmelak[1] and **Yang Xu**[2] and **Michael Hahn**[1] and **Kate McCurdy**[1]
[1]Saarland University, Saarland Informatics Campus, Germany
[2]Department of Computer Science, Cognitive Science Program, University of Toronto, Canada
{dosmelak,mhahn,kmccurdy}@lst.uni-saarland.de, yangxu@cs.toronto.edu

Abstract

Languages vary widely in how meanings map to word forms. These mappings have been found to support efficient communication; however, this theory does not account for systematic relations *within* word forms. We examine how a restricted set of grammatical meanings (e.g. person, number) are expressed on verbs and pronouns across typologically diverse languages. Consistent with prior work, we find that verb and pronoun forms are shaped by competing communicative pressures for *simplicity* (minimizing the inventory of distinct forms) and *accuracy* (enabling recovery of intended meanings). Our proposed model uses a novel neural-network measure of complexity (inverse of simplicity) based on the *learnability* of meaning-to-form mappings. This innovation captures fine-grained regularities in linguistic form, allowing better discrimination between attested and unattested systems, and establishes a new connection from efficient communication theory to systematicity in natural language.

1 Introduction

Languages express a vast array of grammatical meanings with a limited set of forms. A fundamental goal of natural language research is to understand how these forms map onto meanings, and why this mapping differs across languages. For example, standard English uses the pronoun *you* for any second-person addressee, while other languages use distinct forms based on features like gender or number (e.g., *sen* "you (singular)" vs. *siz* "you (plural)" in Turkish; Figure 1). What drives this cross-linguistic variation in form-meaning mappings?

We posit that this variation arises because languages evolve under two competing communicative pressures. On one hand, a general bias toward *simplicity* encourages languages to minimize distinctions between forms. This criterion favors the English pronoun system, with only one second-person form, as less cognitively demanding than Turkish, which requires the speaker to distinguish addressees by number. On the other hand, a competing bias toward *accuracy* encourages languages to maintain distinct forms for different meanings, enabling precise decoding of meaning from form. For instance, a Turkish listener can instantly understand whether the speaker intends to address an individual (*sen*) or a group (*siz*), while an English listener must infer between these two meanings of the word *you*. English and Turkish pronoun systems illustrate two different design solutions to the trade-off between simplicity and accuracy (Zaslavsky et al., 2021b).

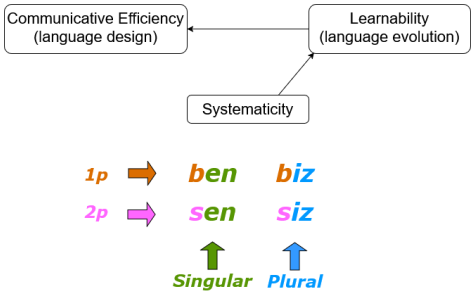


Figure 1: Turkish pronouns show *systematic* form-meaning mappings: person is consistently marked by prefixes (e.g., *s-* for second person), number by suffixes. Language evolution research demonstrates that such systematicity supports *learnability*. Our model connects these findings, proposing that learnable, systematic mappings contribute to *communicative efficiency*.

This trade-off has been explored through two distinct research traditions. An influential line of work on *efficient communication* (e.g. Kemp et al., 2018; Gibson et al., 2019; Hahn et al., 2020) analyzes how natural languages balance simplicity and accuracy in their lexicons. This research finds that lexicons achieve near-optimal simplicity-accuracy trade-offs across semantic domains including color (Zaslavsky et al., 2018), kinship (Kemp and Regier, 2012), and numerals (Xu et al., 2020). In parallel,

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29047–29083

research on *language evolution* (e.g. Smith et al., 2017; Culbertson and Schuler, 2019) investigates how artificial languages evolve in laboratory settings. Through experimental manipulation, this work shows that interacting biases toward simplicity and accuracy drive the emergence of core "design features" of natural language (Hockett, 1960; Christiansen and Chater, 2008; Smith, 2022).

One such core feature is *systematicity*,[1] in which discrete parts of linguistic forms reliably indicate specific meanings; consider the Turkish pronouns in Figure 1, where prefixes indicate person (e.g. second person pronouns begin with *s-*) and suffixes indicate number. Systematic mappings emerge in artificial languages as a design solution to the competing pressures of simplicity and accuracy (Smith et al., 2013; Kirby et al., 2015; Smith, 2022; Smith and Culbertson, 2025): decomposing complex expressions into parts yields a smaller, simpler lexicon, while systematically recombining these parts enables fine-grained distinctions and accurate decoding. The literature thus identifies systematic form-meaning mappings as a core efficiency mechanism in the evolution of artificial languages. However, this insight is not yet reflected in research on communicative efficiency in *natural* languages, which lacks a general framework for modeling the internal structure of word forms.

We propose a unified information-theoretic framework to integrate these two research traditions and assess how systematic form-meaning mapping supports efficiency in natural language. Our key contribution is a novel complexity measure based on the learnability of meaning-to-form mappings. This measure captures systematicity in linguistic forms within the theoretical framework of efficient communication.

We evaluate our framework in two domains: verbs and pronouns. Verbs offer rich variation in how structured meanings map to forms across languages, and in many languages exhibit overtly systematic form-meaning covariation through inflectional morphology (Haspelmath and Sims, 2010). Pronouns provide a natural comparative case to an earlier analysis under the well-known Information Bottleneck model (Zaslavsky et al., 2021b), allowing us to directly compare approaches.

We find that our framework successfully represents structured form-meaning mappings in both domains, and consistently discriminates attested from unattested systems. Crucially, our analysis finds that explicitly modeling systematicity yields a more precise account of efficiency in language, therefore better capturing the fine-grained regularities that characterize natural linguistic systems.

[1]Systematicity enables another core design feature: *compositionality*, the capacity to build complex expressions by combining discrete parts (Frege, 1914; McCurdy et al., 2024).

		singular (sg)	dual (du)	plural (pl)
1	m	ʔa-12u3-u	na-12u3-u	na-12u3-u
	f	ʔa-12u3-u	na-12u3-u	na-12u3-u
2	m	ta-12u3-u	ta-12u3-āni	ta-12u3-ūna
	f	ta-12u3-īna	ta-12u3-āni	ta-12u3-na
3	m	ya-12u3-u	ya-12u3-āni	ya-12u3-ūna
	f	ta-12u3-u	ta-12u3-āni	ya-12u3-na

Table 1: Basic imperfective paradigm of a stem I Classical Arabic verb for feature values gender {**m**asculine, **f**eminine}, person {1, 2, 3}, and number {sg, du, pl}. Numbers (e.g. 12u3) represent root consonants.

2 Background

2.1 Paradigms and Syncretism

A **paradigm** comprises systematic relations between word forms which are structured by grammatical categories such as person, gender, number, tense, or case. These relations can hold between distinct words, such as the Turkish pronouns shown in Figure 1, or parts of words, such as the prefixes and suffixes shown in Table 1. Specific feature combinations are realized by the form in the corresponding *cell* of the paradigm table, e.g. "1sg m" (first person singular masculine) → *ʔa- -u*.

Syncretism occurs when two or more cells of a paradigm with distinct grammatical functions share the same surface form. For instance, *ta- -u* realizes multiple feature combinations in Arabic (Table 1). Within highly structured paradigmatic meaning spaces, syncretism directly corresponds to the form-meaning partitions relevant to communicative efficiency; for instance, Zaslavsky et al. (2021b) show that pronoun syncretism patterns are optimized for efficient communication. By definition, syncretism reflects only identity relations between forms (Haspelmath and Sims, 2010). Our proposed model extends beyond syncretism to capture partial overlap in forms, such as the systematic correspondence between the grammatical second person and the prefix *ta-* in Arabic verbs (Table 1).

2.2 Systematicity in Verb and Pronoun Forms

Verbs. Verbal morphology offers a rich testbed to investigate structured forms and meanings. We consider two distinct types of verbal inflection. Concatenative inflection attaches, i.e. concatenates, affixes to a word stem; consider for example the English past tense suffix *-ed*, as in *jump, jumped*. Non-concatenative inflection requires changing the word stem; consider for example *run, ran*. We model Semitic and other Afro-Asiatic verbs, which are characterized by both kinds of morphology. On the non-concatenative side, Semitic *roots* typically comprise 3 consonants which combine templatically with vowels to produce stems (Huehnergard and Pat-El, 2019b; Weninger, 2011; Lipiński, 1997). These stems are then additionally combined with concatenative morphology in the form of suffixes and sometimes prefixes (cf. Table 1).We also model verbal morphology in Romance and Germanic languages, where inflection is mainly concatenative.

Pronouns. While inflectional morphology reflects systematic form-meaning association by definition (Haspelmath and Sims, 2010), pronouns do not generally display such rich formal structure. Systematicity, however, has also been identified at formal levels below morphology, such as specific sounds (e.g. Blasi et al., 2016; Pimentel et al., 2019; Monaghan et al., 2014). We posit that pronoun forms are also shaped by learnability, and propose modeling their efficiency with an appropriate complexity measure.

3 Model

We propose an information-theoretic framework to integrate systematicity into efficient communication theory. Our framework builds on a well-established insight: languages evolve under competing pressures for accuracy (enabling precise communication) and simplicity (minimizing cognitive demands). These competing pressures predict that linguistic systems should occupy a particular region of the accuracy-simplicity trade-off space, achieving near-optimal balance between the two.

The Information Bottleneck (IB) model (Zaslavsky et al., 2018, 2019, 2021b) provides an influential formalization of this trade-off for natural language lexicons. In this model, a speaker samples an object t based on a need distribution $p_{\text{cog}}(t)$ and communicates it by producing a signal w from an encoding distribution $q_{\text{enc}}(w|m_t)$. The listener interprets w using Bayesian inference to construct a probability distribution $q_{\text{dec}}(\hat{m}|w)$ over possible meanings. The encoding process is under pressure to minimize *complexity*, while decoding is pressured to maximize *accuracy*. The IB framework evaluates this trade-off by comparing *attested* semantic systems against *counterfactual* alternatives; if attested systems consistently achieve better complexity-accuracy trade-offs than counterfactuals, this indicates they have been shaped for communicative efficiency.

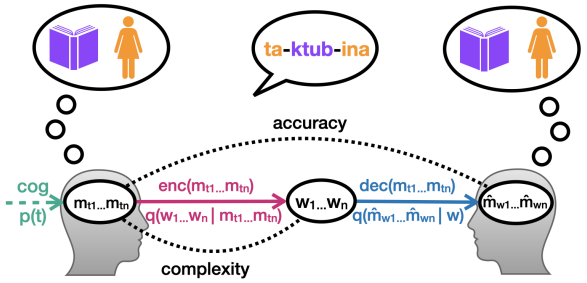


Figure 2: Communication model, adapted from Zaslavsky et al. (2018, 2021b). Our model encodes the form w as a sequence, and decodes it as an atomic unit.

The IB model, however, treats linguistic forms as atomic units with no internal structure. For example, the mappings "3fs $\rightarrow$ *ta- -u*" and "3ms $\rightarrow$ *ya- -u*" (Table 1) could be equivalently represented as "3fs $\rightarrow X$" and "3ms $\rightarrow Y$." This means the model can detect reduced complexity through *syncretism* (using identical forms for multiple meanings) but not through *systematic* form-meaning covariation *within* forms (e.g., the suffix *-u* in the third-person singular). Our framework addresses this gap while preserving the IB model's theoretical foundations and its measure of accuracy (Figure 2).

Our key contribution is a novel measure of complexity based on the learnability of meaning-to-form mappings. This measure captures systematicity in linguistic forms through a neural network encoder that learns character-level sequences, allowing us to quantify how structural regularities facilitate learning and reduce cognitive complexity.

3.1 Complexity

3.1.1 Our model

We hypothesize that paradigms with more structured mappings will be more learnable. Following the efficient communication literature, we model this with two components: a need distribution $p_{\text{cog}}(t)$, and an encoder $q_{enc}(w|m_t)$.

Need distribution. $p_{\text{cog}}(t)$ is a probabilistic weighting over targets t associated with meanings m_t, reflecting communicative need. We assume that more frequent meanings have greater need, and estimate $p_{\text{cog}}(t)$ using corpus frequency. This weights meanings according to a learner's exposure to different meanings at different frequencies.

Encoder. $q_{enc}(w|m_t)$ maps from grammatical meanings m_t to a linguistic form w. We implement this using a sequence-to-sequence neural network that takes morphosyntactic features as input (m_t) and generates the corresponding surface form as output (w; Figure 3).[2] Critically, by encoding forms as character sequences rather than atomic units, our encoder can exploit systematic regularities across forms. We expect training to converge more rapidly for more learnable paradigms. We therefore measure complexity in terms of the encoder's cross-entropy decay during learning.

Complexity measure. Let $q_{enc}^{(T)}(w|m_t)$ be the encoder after T training epochs. We define the cross-entropy training loss (**CETL**) as:

$$-\frac{\sum_{T=1}^{T_{max}} \sum_t p_{\text{cog}}(t) \log q_{\text{enc}}^{(T)}(w_t|m_t)}{T_{max}} \quad (1)$$

Here w_t is the correct form for meanings m_t, and T_{max} is the maximum number of epochs. The core of this equation is $-\log q_{enc}^{(T)}(w_t|m_t)$, which quantifies how well the encoder predicts the form after T training steps. By summing over timesteps, Equation 1 is smaller when training is faster. CETL is therefore information-theoretic, grounded in learnability, and fully general — it can be applied broadly across languages and semantic systems.

Implementation. Our neural network is a LSTM-based sequence-to-sequence (seq2seq) architecture,[3] with two stacked LSTM layers in the encoder and in the decoder. We train with batch size 1 and dropout rate 0.5. Based on preliminary experiments, we set $T_{max} = 50$ epochs[4] and measure total cross-entropy loss weighted by the need distribution. We train ten separate networks for each original paradigm, and five for each counterfactual paradigm, and average over the results.[5]

Figure 3: Example input and output used in training.

3.1.2 Computational measures of complexity

Our learnability-based approach connects to several lines of research on linguistic complexity, though CETL uniquely captures systematicity in form-meaning mappings.

Learnability-based complexity. Similar neural learnability-based approaches to complexity have been proposed, but none shares our focus on systematicity across forms. Steinert-Threlkeld, Szymanik and colleagues (Steinert-Threlkeld and Szymanik, 2019, 2020; Carcassi et al., 2021) also measure complexity via RNN learning rates within a communicative efficiency framework. Johnson and colleagues (Johnson et al., 2021, 2024) similarly measure morphological complexity via RNN classification accuracy during artificial language learning. These proposed measures, however, rely on classification tasks, and do not attend to forms' internal structure. More recently, Denić and Szymanik (2024) model morphosyntactic complexity in recursive numeral systems, finding an optimal trade-off with lexicon size. This work, like ours, addresses compositional morphological structure; however, it abstracts away from form, instead measuring complexity by average morpheme count. By contrast, CETL detects systematicity across forms (e.g. the shared prefix between "two" and "twenty") without manual annotation. Compared to other learnability-based measures, our approach is more fine-grained, using a continuous information-theoretic measure which is critically sensitive to output forms.

Information-theoretic complexity. Other information-theoretic measures of morphological complexity have been proposed, but most do not fit our specific application. Ackerman and Malouf (2013) define *enumerative complexity* (*e-complexity*; essentially the number of for-

[2]The encoded form w is domain-specific: for pronouns, w is the entire word; for verbs, w represents only inflectional markers. The model removes whitespaces and reads letter by letter, including special markers for sequence boundaries. For example, "*2p m G*" becomes "*<sos>2pmG<eos>*" as input, with output "*<sos>ta12u3uuna<eos>*".

[3]We use LSTMs in keeping with prior neural models of morphology (e.g. Wu et al., 2019; Rathi et al., 2021; Cotterell et al., 2018).

[4]This was sufficient to achieve zero loss on some Semitic verbal paradigms. While not guaranteed sufficient for all variants, crucially, we hold T_{max} constant across paradigms.

[5]All code and data developed in this project is available under https://github.com/Doosme/paradigm_efficiency_verbpron.

mal distinctions) and *integrative complexity* (*i-complexity*; the degree to which forms predict each other). They argue that natural languages vary substantially in e-complexity, but tend toward low i-complexity. Variants of this account have been operationalized with seq2seq models (Cotterell et al., 2019; Johnson et al., 2020, 2021). There is a conceptual link to our work, which uses seq2seq learners to model predictive relations between inflected forms. However, i-complexity estimates predictability based on variation *across* paradigms (e.g., over different inflection classes). Similarly, Wu et al. (2019) quantify the irregularity of inflected forms based on their predictability given the rest of the language. By contrast, we abstract over lexemes and estimate each paradigm's complexity *independently* from the rest of the language. Rathi et al. (2021) introduce *informational fusion*, which measures the extent to which a form cannot be predicted by the rest of the paradigm. This measure would be applicable in our domains but has a high computational cost, requiring re-fitting seq2seq models for each form in a paradigm.[6] We prioritize broad coverage, making CETL the most efficient approach.

IB Complexity. The Information Bottleneck model also uses an information-theoretic formulation of complexity based on need distribution and encoder. For a semantic system with meanings M and forms W, the IB model defines complexity as the mutual information between intended meanings $m_t \in M_t$ and word forms $w \in W$:

$$\sum_{\substack{t \in \mathcal{U} \\ w \in \mathcal{W}}} p_{\mathsf{cog}}(t) q_{\mathsf{enc}}(w|m_t) \log \frac{q_{\mathsf{enc}}(w|m_t)}{q_{\mathsf{enc}}(w)} \quad (2)$$

where $q_{\mathsf{enc}}(w) = \sum_{t \in \mathcal{U}} p_{\mathsf{cog}}(t) q_{\mathsf{enc}}(w|m_t)$. This computes the bits required for the encoder $q_{\mathsf{enc}}(w|m_t)$ to communicate all meanings, weighted by the need distribution. However, because the IB encoder treats forms as discrete atomic units, this measure of complexity is reduced by syncretism but not by systematic meaning-form covariation.[7] CETL addresses this gap by encoding w as a character sequence, allowing systematic regularities across forms and meanings to reduce complexity through faster learning.

[6]Similar concerns apply to related evolutionary models with iterated learning (e.g. Ackerman and Malouf, 2015; Round et al., 2025).

[7]Bruneau–Bongard et al. (2025) identify other unintuitive properties of IB complexity, such as insensitivity to synonyms.

3.2 Accuracy

We adopt the IB model's widely-used, domain-general accuracy measure. Under this framework, a listener uses Bayesian inference to construct a probability distribution $q_{\mathsf{dec}}(\hat{m}|w)$ over possible meanings given the received form w (Figure 2). Accuracy is defined as the similarity between the speaker's intended meaning distribution m_t and the listener's reconstructed distribution $\hat{m}_w$:

$$\mathsf{acc} = -\mathbb{E}_{\mathsf{dec}}[D_{KL}[\hat{m}_w || m_t]] \,, \quad (3)$$

where D_{KL} is the Kullback-Leibler (KL) divergence.[8] Following the literature, we assume that speakers always produce the correct form for a target meaning, so m_t is a one-hot vector.

To specify $\hat{m}_w$, we require the need distribution $p_{\mathsf{cog}}(t)$ and an underlying representation of the semantic domain, reflecting the listener's knowledge. Following Zaslavsky et al. (2021b), we encode grammatical meanings as categorical features (cf. Regier et al., 2015), and specify their similarity as a weighted Hamming distance $d(u,t)$[9]:

$$m_t(u) \propto \exp(-d(u,t)). \quad (4)$$

However, where Zaslavsky et al. (2021b) use binary features, we employ a general categorical encoding scheme.[10] For feature representations u, t, we define $d(u,t)$ as the number of dimensions i where $u_i \neq t_i$. This groups combinatorial meanings with shared features; for example, "first-person singular past" is closer to the meaning "first-person plural past" than "third-person singular future." Importantly, while this IB accuracy measure posits complex structure in semantic *meaning*, it still represents the given form w as a discrete atomic unit.

4 Evaluation

We formulate two complementary hypotheses about how systematicity affects communicative efficiency in natural language paradigms. To test these hypotheses, we compare attested paradigms against systematically generated counterfactuals. We use this approach to evaluate our model on verb and pronoun data, and compare to the IB framework.

[8]Zaslavsky et al. (2021b) add a constant to make this quantity nonnegative, which changes results by a constant offset.

[9]They include a free parameter γ; we set $\gamma = 1$.

[10]We compare against their feature encoding in App. B.3 and find that our results are robust to the change.

		sg	pl
1		'a-	na-
2	m	ta-	ta- -u
	f	**ta-**	ta- -u
3	m	ya-	ya- -u
	f	**ta- -i**	ya- -u

(a) A **structural** permutation over the person category (PERS) that **increases** naturalness: the form *ta-* now has only value 2 for grammatical person, whereas before it covered values 2, 3.

		sg	pl
1		'a-	na-
2	m	ta-	ta- -u
	f	**ya- -u**	ta- -u
3	m	ya-	ya- -u
	f	ta-	**ta- -i**

(b) A **structural** permutation over multiple categories (PERS, NUM) that **decreases** naturalness: the form *ya- -u* now has values 2, 3 for grammatical person, and sg, pl for number, whereas before it covered only the feature values 3, pl.

		sg	pl
1		'a-	na-
2	m	**ya- -u**	ta- -u
	f	ta- -i	ta- -u
3	m	ya-	**ta-**
	f	**ya- -u**	**ta-**

(c) A **form-only** permutation over multiple categories (PERS, NUM, GEN) which changes form realizations but does not affect paradigm structure, i.e. the distribution of syncretic forms; compare to Table 2b.

Table 2: Example permutations applied to Dialectal Arabic conjugation. Identical (i.e. syncretic) forms with naturalness affected by the permutation are underlined. Structural permutations affect syncretic patterns, but form-only permutations do not. Altered forms are highlighted in **orange** (structural) or **cyan** (form-only).

Hypothesis 1 (Efficiency). Attested paradigms achieve better complexity-accuracy trade-offs than counterfactual alternatives. This follows standard practice in the IB literature (e.g. Kemp and Regier, 2012; Zaslavsky et al., 2021b) and tests whether natural language paradigms are optimized for communicative pressures.

Hypothesis 2 (Naturalness). Among counterfactual paradigms, those with more natural syncretism patterns (where syncretic forms share similar meanings) are more learnable (Saldana et al., 2022) and thus have lower complexity. This tests whether our model captures not only form-internal, but also paradigm-level structure.

4.1 Generating counterfactual paradigms

We generate counterfactuals by permuting existing paradigms, creating alternatives that vary in two key ways: their syncretism patterns (relevant to systematicity *and* naturalness) and their specific form-meaning mappings (relevant to systematicity alone). All counterfactuals are compared to their source paradigm as baseline.

Structural permutations swap the contents of paradigm cells in ways that alter syncretism patterns. For example, we might permute the person feature while holding other distinctions constant, swapping "2sg f → *ta- -i*" with "3sg f → *ta-*" (Table 2a). We can permute single features (e.g.,$[2 \Rightarrow 3]$) or multiple features simultaneously (e.g., $[2 \Rightarrow 3]$ $[sg \Rightarrow pl]$; Table 2b). These permutations change which meanings are expressed by syncretic forms, affecting the paradigm's naturalness score (defined in 4.2).

Form-only permutations swap surface forms while maintaining the original syncretism patterns. For example, we might exchange the form realizing "[2sg m], [3sg f]" with the form realizing "3sg [m, f]," which retains the original syncretic structure of the paradigm (Table 2c). By definition, these permutations do not affect naturalness scores. Critically, they also cannot affect complexity in the IB model, which treats forms as atomic units. However, they can affect our complexity measure if the permutation disrupts systematic regularities (e.g., breaking a consistent prefix pattern). Form-only permutations thus provide a key test of whether our model captures systematicity beyond syncretism.

4.2 Measuring naturalness

Our naturalness hypothesis builds on an insight already present in the IB accuracy measure (3.2): grammatical meanings with shared features are more similar. Systematicity suggests that this principle should extend to forms. In particular, *syncretic* forms — which share all of their formal features — should also share many aspects of grammatical meaning. For example, the syncretic form *ta- -u* in Table 2 expresses meanings that all share $person = 2, number = pl$. By contrast, an "unnatural" syncretism might use the same form for different meanings with no shared features.

Saldana et al. (2022) develop this intuition as a **naturalness** gradient in syncretism patterns. Inspired by their analysis, we define the *unnaturalness score* at the paradigm level:

$$\mathsf{unnat}_\pi = \sum_{c \in \mathsf{SynClass}_\pi} \sum_{f \in \mathsf{FeatCat}_\pi} (\#c_f - 1) \quad (5)$$

where c_f are all feature values that exist in the syncretism class c for the feature category f. This

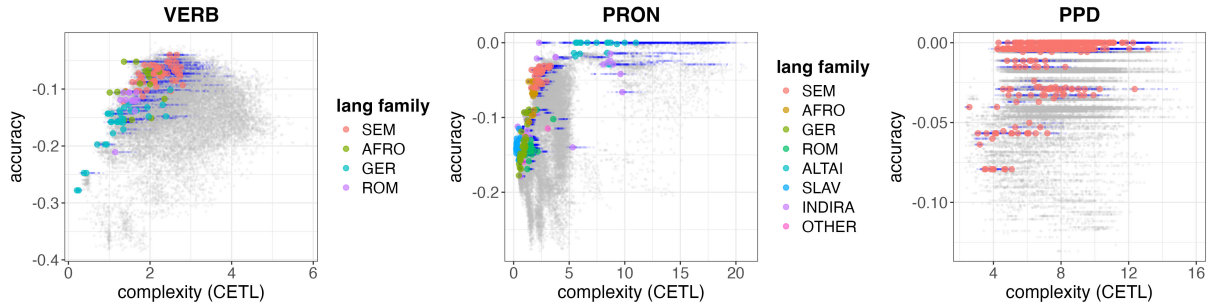


Figure 4: Complexity and accuracy in our model across typologically diverse linguistic paradigms from verbal inflection (VERB) and pronouns with detailed (PRON) and simplified (PPD) feature representation schemes. In all domains, real, attested paradigms (colored, one dot per language) are more efficient — i.e. higher in accuracy and/or lower in complexity — than nearly all counterfactual structural (gray) and form-only (blue) permutations.

	subsets	#languages	#total permutations	feature categories
PPD		561	93845	PERS, NUM
PRON	PR_SEM	45	16437	
	PR_AFRO	19	6677	
	PR_GER	30	10649	
	PR_ROM	17	3605	PERS, NUM,
	PR_BALTSLAV	20	6063	GEN, CASE
	PR_INDOIRAN	18	3616	
	PR_ALTAIC	19	3765	
	PR_OTHER	9	2264	
VERB	VERB_SEM	56	19700	
	VERB_AFRO	13	5773	PERS, NUM,
	VERB_GER	32	8875	GEN, TEN
	VERB_ROM	12	2245	

Table 3: Summary of the datasets. For details of family-specific features within each category, see Appendix A.

counts the number of feature values that differ within each syncretic form. A low unnaturalness score indicates natural syncretism patterns where syncretic forms express similar meanings.

We compute unnaturalness scores for counterfactual paradigms relative to their attested baselines. For example, the original Dialectal Arabic paradigm in Table 2 has an unnaturalness score of 4, so we subtract that from all counterfactual paradigm scores. The permutation in Table 2a *increases* naturalness, yielding an unnaturalness score of –1; Table 2b does the opposite, yielding a score of 1; and Table 2c maintains the same syncretic structure as the original paradigm, so its unnaturalness score is 0. We hypothesize that counterfactual paradigms with high unnaturalness scores will also have higher CETL complexity.

4.3 Data

Table 3 summarizes the languages and features used in our evaluation. Domains 1 and 2b use resources we collected ourselves; see Appendix D for sources and citations.

Domain 1: Verbal Morphology (VERB). We test our model on verbal inflection in Semitic, non-Semitic Afro-Asiatic, Germanic, and Romance languages, including both concatenative and non-concatenative morphology. Table 1 illustrates our approach to form and feature representation; see Appendix A for more details.

Domain 2a: Pronouns (PPD). To directly compare with Zaslavsky et al. (2021b), we use the Pronoun Paradigms Database (Greenhill, 2025), which contains hundreds of languages across families. We follow their simplified feature scheme (Table 3, top row), which uses only masculine forms and omits dual number.

Domain 2b: Detailed Pronouns (PRON). We also test on full pronoun paradigms for selected language families (Semitic, Germanic), including case and gender. Unlike Domain 2a, the feature set varies by language to capture language-specific distinctions.

Need Distribution. We estimate $p_{cog}(t)$ by combining the need distribution from Zaslavsky et al. (2021b) with additional corpus frequency estimates for gender and dual/plural distinctions, assuming uniform distributions otherwise. Results are robust to this choice (Appendix B).

5 Results

We first evaluate whether our model supports the efficiency and naturalness hypotheses (5.1), then compare its performance to the IB approach (5.2). For additional detailed results, see Appendix C.

5.1 Testing efficiency and naturalness

Efficiency. Figure 4 shows that attested verb and pronoun paradigms are more efficient than nearly all counterfactual permutations: they show lower

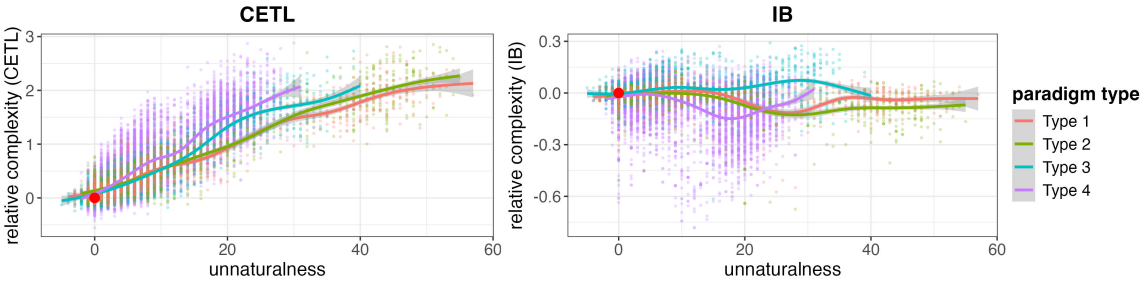


Figure 5: Complexity plotted against unnaturalness for Afro-Asiatic verbs by model, CETL (left) and IB (right). We group results by paradigm type (reflecting different grammatical categories; cf. App. A.2) and report both complexity and unnaturalness of permutations relative to original paradigms (cf. 4.2). CETL positively correlates with unnaturalness, meaning more natural paradigms have lower complexity, while IB shows no correlation.

	C_M (%)		I_M (%)		$Perf_M$		
	CETL	IB	CETL	IB	CETL	IB	support
PPD	**72.37**	4.82	2.19	**0.73**	**70.19**	4.10	44112
PRON	**90.12**	48.71	**1.48**	4.53	**88.64**	44.18	38623
VERB	**76.32**	41.03	**3.77**	4.08	**72.55**	36.95	32825

(a) Structural Permutations

	C_M (%)		I_M (%)		$Perf_M$		
	CETL	IB	CETL	IB	CETL	IB	support
PPD	**65.81**	0.00	**34.19**	100	**31.63**	-100	51061
PRON	**89.21**	0.00	**10.79**	100	**78.42**	-100	10697
VERB	**78.23**	0.00	**21.77**	100	**56.47**	-100	3781

(b) Form-only Permutations

Table 4: Performance ($Perf_M$) for structural (left) and surface permutations (right), based on percentage of correct (C_M) and incorrect (I_M) efficiency classifications, under the assumption that original paradigms should be more efficient than counterfactuals. CETL outperforms IB across the board.

CETL and higher accuracy, with statistical significance across all domains (details in Appendix C.1). This confirms Hypothesis 1: natural language paradigms occupy a near-optimal region of the complexity-accuracy trade-off space, consistent with communicative efficiency theory.

Naturalness. Figure 5 (left panel) shows an illustrative positive correlation ($\rho = 0.5745$; $p < 2.2e-16$) between CETL and unnaturalness for Afro-Asiatic verbs. Since the distribution of unnaturalness scores varies across languages as a function of paradigm size, we calculate averaged per-language correlations for each domain, finding a lower correlation of 0.36 for PPD, and strong correlations of 0.82 and 0.88 for PRON and VERB respectively. This confirms Hypothesis 2: paradigms with more natural syncretism patterns are more learnable, and thus have lower complexity.

Together, these results validate our core hypothesis: systematicity in form-meaning mappings facilitates learning and thereby reduces complexity.

5.2 Comparison to the IB Model

Our CETL measure is sensitive to surface form structure, which we expect to increase discriminative capacity relative to the IB model's complexity measure. Specifically, we expect CETL to correctly identify attested paradigms as more efficient than counterfactual permutations, and show a stronger correlation than IB to paradigm naturalness.

Evaluation. We measure each model's ability to correctly identify attested paradigms as more efficient than counterfactuals. For each counterfactual permutation, we classify each estimate as either Correct(C): worse than the attested paradigm in both accuracy and complexity, or worse in one measure while equal in the other — or Incorrect(I): better than or equal to the attested paradigm in both measures. We define the performance of a model M as $Perf_M = C_M - I_M$ (i.e., the difference between correct and incorrect identifications).

Results. Table 4b shows that CETL correctly identifies 65% to nearly 90% of form-only permutations as less efficient than attested paradigms, while the IB model cannot distinguish these permutations at all (by design, since they have identical syncretism patterns). This demonstrates that CETL captures systematicity beyond mere syncretism. Table 4a shows that CETL consistently outperforms the IB model in identifying attested paradigms as more efficient than structural permutations. Finally, Figure 5 shows that CETL (left) strongly correlates with naturalness, with correlations exceeding 0.8 for PRON and VERB, while the IB model (right) shows no correlation for either domain. This confirms that CETL successfully captures fine-grained

regularities in syncretism patterns that the IB model cannot detect. We conclude that our learnability-based complexity measure better captures systematic form-meaning mappings, enabling better discrimination between attested and counterfactual paradigms.

6 Conclusion

Languages vary widely in form-meaning mappings. We propose a unified framework that integrates systematicity between meaning and form into efficient communication theory, and evaluate our model on hundreds of languages from diverse families. Our results show that: (1) attested paradigms are more efficient than counterfactuals; (2) paradigms with more natural syncretism patterns are more learnable; and (3) our model outperforms the IB approach in discriminating attested from counterfactual paradigms. This work connects efficient communication theory with language evolution research on systematicity, showing how communicative pressures shape form-meaning mappings across the world's languages.

Limitations

Our analysis focuses exclusively on discrete, paradigmatically structured domains, specifically verbal inflection and pronouns. The applicability of CETL to other semantic domains remains uncertain. In particular, continuous domains like color present a fundamental challenge: they lack the discrete categorical structure assumed by our sequence-to-sequence architecture.

While CETL captures fine-grained systematicity that the IB measure misses, we have not established whether it is universally preferable or whether trade-offs exist. For instance, we do not replicate the fine-grained feature weight optimization of Zaslavsky et al. (2021b), leaving open whether CETL supports such analyses as effectively. The two measures may prove complementary, with different applications suited to each.

Lastly, unlike some prior work in the efficient communication literature, we do not provide an explicit Pareto frontier of maximally efficient paradigms. Doing so would require specifying all possible counterfactual paradigms, including all forms compatible with the phonology and phonotactics of each language. Our permutation-based approach generates structured counterfactuals but does not exhaustively sample the space of possibilities, limiting our ability to make strong optimality claims about attested systems.

Ethics Statement

This paper concerns foundational research on the structure of language. We do not foresee immediate ethical implications.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. We are grateful to Richard Futrell, Mora Maldonado, Carmen Saldana, and Noga Zaslavsky for helpful discussion.

References

Ernest T. Abdel-Massih. 1971. *A Reference Grammar of Tamazight (Middle Atlas Berber)*. Center for Near Eastern and North African Studies, University of Michigan.

Aynur Abish. 1998. Kazakh and Karakalpak. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 337–353. Routledge.

Farida Abu-Haidar. 1991. Christian arabic of baghdad.

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89:429–464.

Farrell Ackerman and Robert Malouf. 2015. The No Blur Principle Effects as an Emergent Property of Language Systems. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, 41(41).

Klára Agyagási. 1998. Chuvash. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 460–478. Routledge.

Rashid Al-Balushi. 2017. Omani Arabic: More than a Dialect. *Macrolinguistics*, 4:80–125.

Erik Andersson. 1994. Swedish. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 271–312. Routledge.

David L. Appleyard. 2011. Semitic-Cushitic/Omotic Relations. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 38–53. De Gruyter.

John Ole Askedal. 1994. Norwegian. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 219–270. Routledge.

Unknown Author. 1999. *Proposta per a un estandàrd oral de la llengua catalana, II – Morfologia*. Institut d'Estudis Catalans.

Elitzur A. Bar-Asher Siegal. 2013. *Introduction to the Grammar of Jewish Babylonian Aramaic*. Ugarit-Verlag.

Michael P. Barnes and Eivind Weyhe. 1994. Faroese. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 190–218. Routledge.

Daniel Birnstiel. 2019. Classical Arabic. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 367–402. Routledge.

Damián E Blasi, Søren Wichmann, Harald Hammarström, Peter F Stadler, and Morten H Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823.

Hendrik Boeschoten. 1998. Uzbek. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 388–408. Routledge.

Françoise Briquel Chatonnet and Robert Hawley. 2020. *Phoenician and Punic*, pages 297–318.

Wayles Browne. 2002. Serbo-Croat. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 306–387. Routledge.

Jeanne Bruneau–Bongard, Emmanuel Chemla, and Thomas Brochhagen. 2025. Assessing Pressures Shaping Natural Language Lexica. *Cognitive Science*, 49(12):e70145. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.70145.

Kristen Brustad and Emilie Zuniga. 2019. Levantine Arabic. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 403–432. Routledge.

Maria Bulakh. 2019. Tigrinya. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 174–201. Routledge.

Gustav Burbiel. 2018. *Tatar Grammar - A Grammar of the Contemporary Tatar Literary Language*. Institute for Bible Translation, Stockholm.

Bogdan Burtea. 2011. Mandaic. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 670–685. De Gruyter.

Aaron Michael Butts. 2019. Gəʕəz (Classical Ethiopic). In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 117–144. Routledge.

Fausto Carcassi, Shane Steinert-Threlkeld, and Jakub Szymanik. 2021. Monotone Quantifiers Emerge via Iterated Learning. *Cognitive Science*, 45(8):e13027.

George Cardona and Babu Suthar. 2003. Gujurati. In George Cardona and Dhanesh Jain, editors, *The Indo-Aryan Languages*, pages 722–765. Routledge.

James E. Cathey. 2000. *Old Saxon*. LINCOM EUROPA.

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

Eleanor Coghill. 2019. Northeastern Neo-Aramaic: the dialect of Alqosh. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 711–748. Routledge.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2018. On the diachronic stability of irregularity in inflectional morphology. In *arXiv preprint arXiv:1804.08262*.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Jennifer Culbertson and Kathryn Schuler. 2019. Artificial Language Learning in Children. *Annual Review of Linguistics*, 5(1):353–373.

Probal Dasgupta. 2003. Bangla. In George Cardona and Dhanesh Jain, editors, *The Indo-Aryan Languages*, pages 386–428. Routledge.

Anne Boyle David. 2014. *Descriptive Grammar of Pashto and its Dialects*. De Gruyter Mouton.

Milica Denić and Jakub Szymanik. 2024. Recursive Numeral Systems Optimize the Trade-off Between Lexicon Size and Average Morphosyntactic Complexity. *Cognitive Science*, 48(3):e13424. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.13424.

Bruce Donaldson. 1994. Afrikaans. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 478–504. Routledge.

Lutz Edzard. 2019. Amharic. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 202–226. Routledge.

Peter Eisenberg. 1994. German. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 349–387. Routledge.

David L. Elias. 2019. Tigre of Ginda ʕ. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 145–173. Routledge.

Jeffrey Ellis. 1953. *An Elementary Old High German Grammar: Descriptive and Comparative*. Oxford University Press.

Marcel Erdal. 2004. *A Grammar of Old Turkic*. Brill.

Jan Terje Faarlund. 1994. Old and Middle Scandinavian. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 38–71. Routledge.

Pompeu Fabra. 2006. *Gramàtica Catalana*. Institut d'Estudis Catalans.

Grazio Falzon. 1997. *Basic Maltese Grammar*.

Steven E. Fassberg. 2019. Modern Western Aramaic. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 632–652. Routledge.

Zygmunt Frajzyngier and Erin Shay. 2012a. Chadic. In Zygmunt Frajzyngier and Erin Shay, editors, *The Afroasiatic Languages*, pages 236–341. Cambridge University Press.

Zygmunt Frajzyngier and Erin Shay. 2012b. Introduction. In Zygmunt Frajzyngier and Erin Shay, editors, *The Afroasiatic Languages*, pages 1–17. Cambridge University Press.

Gottlob Frege. 1914. Letter to Jourdain. *Philosophical and mathematical correspondence*, pages 78–80. Publisher: Chicago University Press.

Victor A. Friedman. 2002. Macedonian. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 249–305. Routledge.

Dianne Friesen. 2017. *A Grammar of Moloko*. Language Science Press.

R. D. Fulk. 2012. *An Introduction to Middle English: Grammar, Texts*. Broadview Press.

I. J. Gelb. 1952. *Old Akkadian Writing and Grammar*. The University of Chicago Press.

Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How Efficiency Shapes Human Language. *Trends in cognitive sciences*, 23(5):389–407.

Jan Gonda. 1966. *A Concise Elementary Grammar of the Sanskrit Language*. Leiden.

G.C. Goswami and Jyotiprakash Tamuli. 2003. Asamiya. In George Cardona and Dhanesh Jain, editors, *The Indo-Aryan Languages*, pages 429–484. Routledge.

Gene Gragg. 2019. Semitic and Afro-Asiatic. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 22–48. Routledge.

John N. Green. 1997. Spanish. In Martin Harris and Nigel Vincent, editors, *The Romance Languages*, pages 79–130. Routledge.

Simon Greenhill. 2025. [link].

Holger Gzella. 2011. Imperial Aramaic. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 574–586. De Gruyter.

C.G. Häberl. 2019. Mandaic. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 679–710. Routledge.

Jo Ann Hackett. 2008. Phoenician and Punic. In Roger D. Woodard, editor, *The Ancient Languages of Syria-Palestine and Arabia*, pages 82–102. Cambridge University Press.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2020. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, 117(5):2347–2353.

Martin Harris. 1997. French. In Martin Harris and Nigel Vincent, editors, *The Romance Languages*, pages 209–245. Routledge.

Martin Haspelmath and Andrea D Sims. 2010. *Understanding morphology.*, 2 edition. Understanding Language Series. Routledge, London.

Rebecca Hasselbach. 2005. *Sargonic Akkadian - A Historical and Comparative Study of the Syllabic Texts*. Harrassowitz Verlag.

Rebecca Hasselbach-Andee. 2019. Akkadian. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 95–116. Routledge.

Charles F Hockett. 1960. The Origin of Speech. *Scientific American*, 203(3):88–97. Publisher: JSTOR.

Jarich Hoekstra and Peter Meijes Tiersma. 1994. Frisian. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 505–531. Routledge.

David Holton, Peter Mackridge, and Irene Philippaki-Warburton. 2004. *Greek - An Essential Grammar of the Modern Language*. Routledge.

Aaron D. Hornkohl. 2019. Pre-modern Hebrew: Biblical Hebrew. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 533–570. Routledge.

John Huehnegard. 2008. Afro-Asiatic. In Roger D. Woodard, editor, *The Ancient Languages of Syria-Palestine and Arabia*, pages 225–246. Cambridge University Press.

John Huehnergard. 2019. Proto-Semitic. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 49–79. Routledge.

John Huehnergard and Na'ama Pat-El. 2019a. Introduction to the Semnitic languages and their history. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 1–21. Routledge.

John Huehnergard and Na'ama Pat-El. 2019b. *The Semitic Languages*. Routledge.

Matthias Hüning and Ulrike Vogl. 2009. Middle Dutch – A Short Introduction. In *Of Reynaert the Fox: Text and Facing Translation of the Middle Dutch Beast Epic Van den vos Reynaerde*, pages 257–272. Amsterdam University Press.

David Huntley. 2002. Old Church Slavonic. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 125–187. Routledge.

Institute of Islamic Studies of University of Zaragoza. 2013. *A Descriptive and Comparative Grammar of Andalusi Arabic*. Brill.

Juha Janhunen. 1952. *Mongolian*. John Benjamins Publishing Company.

Otto Jastrow. 2011. Ṭuroyo and Mlaḥsô. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 697–707. De Gruyter.

Tamar Johnson, Jennifer Culbertson, Hugh Rabagliati, and Kenny Smith. 2020. Assessing integrative complexity as a predictor of morphological learning using neural networks and artificial language learning.

Tamar Johnson, Micha Elsner, and Kenny Smith. 2024. Testing the Effects of the Implicative Structure and Noun Class Size on the Learnability of Inflectional Paradigms in Adults and Artificial Neural Networks. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0).

Tamar Johnson, Kexin Gao, Kenny Smith, Hugh Rabagliati, and Jennifer Culbertson. 2021. Investigating the effects of i-complexity and e-complexity on the learnability of morphological systems. *Journal of Language Modelling*, 9(1):97–150.

Andra Kalnača and Ilze Lokmane. 2021. *Latvian Grammar*. University of Latvia Press.

Birsel Karakoç. 1998. Turkmen. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 262–286. Routledge.

Birsel Karakoç and Kenjegül Kalieva. 1998. Kirghiz. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 370–387. Routledge.

Darya Kavitskaya. 2009. *Crimean Tatar*. Lincom.

Charles Kemp and Terry Regier. 2012. Kinship Categories Across Languages Reflect General Communicative Principles. *Science*, 336(6084):1049–1054.

Charles Kemp, Yang Xu, and Terry Regier. 2018. Semantic Typology and Efficient Communication. *Annual Review of Linguistics*, 4(1):109–128.

Abdelghany A. Khalafallah. 1969. *A Descriptive Grammar Of Saidi Egyptian Arabic*. Mouton.

Lachman M. Khubchandani. 2003. Sindhi. In George Cardona and Dhanesh Jain, editors, *The Indo-Aryan Languages*, pages 683–721. Routledge.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Leonid Kogan and Maria Bulakh. 2019. Soqotri. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 280–320. Routledge.

Ekkehard König. 1994. Enlish. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 532–565. Routledge.

Maarten Kossmann. 2012. Berber. In Zygmunt Frajzyngier and Erin Shay, editors, *The Afroasiatic Languages*, pages 18–101. Cambridge University Press.

Maarten Kossmann. 2013. *A Grammatical Sketch of Ghadames Berber (Libya)*. Rüdiger Köppe Verlag.

Omkar N. Koul. 2003. Kashmiri. In George Cardona and Dhanesh Jain, editors, *The Indo-Aryan Languages*, pages 991–1051. Routledge.

Omkar N. Koul and Kashi Wali. 2006. *Modern Kashmiri Grammar*. Dunwoody Press.

N. J. C. Kouwenberg. 2017. A Grammar of Old Assyrian. In M. Weeden, editor, *Handbook of Oriental Studies - Section 1 The Near and Middle East*, volume 118, pages 479–510. Brill.

Karl Lahmer. 2018. *Grammateion*. Klett.

Thomas Leddy-Cecere and Jason Schroepfer. 2019. Egyptian Arabic. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 433–457. Routledge.

Winfred P. Lehmann. 1994. Gothic and the Reconstruction of Proto-Germanic. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 19–37. Routledge.

Winfred P. Lehmann. 2007. *A Grammar of Proto-Germanic*. The University of Texas in Austin - Linguistic Research Center.

Edward Lipiński. 1997. *Semitic Languages: Outline of a Comparative Grammar*. Peeters, Leuven.

Graham Mallinson. 1997. Rumanian. In Martin Harris and Nigel Vincent, editors, *The Romance Languages*, pages 391–419. Routledge.

Terje Mathiassen. 1996. *A short Grammar of Lithuanian*. Slavica Publishers, Inc.

Peter Mayo. 2002. Belorussian. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 887–946. Routledge.

Ernest N. McCarus. 2009. Kurdish. In Gernot Windfuhr, editor, *The Iranian Languages*, pages 587–633. Routledge.

Kate McCurdy, Paul Soulos, Paul Smolensky, Roland Fernandez, and Jianfeng Gao. 2024. Toward Compositional Behavior in Neural Models: A Survey of Current Views. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9323–9339, Miami, Florida, USA. Association for Computational Linguistics.

Ronny Meyer. 2011. Gurage. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 1220–1256. De Gruyter.

Ronny Meyer. 2019. Gurage (Muher). In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 227–256. Routledge.

Padraic Monaghan, Richard C Shillcock, Morten H Christiansen, and Simon Kirby. 2014. How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130299.

Claudia Moriena and Karen Genschow. 2004. *Große Lerngrammatik Spanisch*. Hueber.

Anne Multhoff. 2019. Ancient South Arabian. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 321–341. Routledge.

Leonard Newmark, Philip Hubbard, and Peter Prifti. 1982. *Standard Albanian - A Reference Grammar for Students*. Stanford University Press.

Morgan Nilsson. 2024. *Beginner's Somali Grammar*. University of Gothenburg - Department of Languages and Literatures.

Margaret K. Omar. 1975. *Saudi Arabic Urban Hijazi Dialect - Basic Course*. Foreign Service Institute.

Omar Othman. 2019. *Yalla Niḥki arabi Book I - A course in Colloquial Jerusalem Arabic for Beginners*. Hebrew University of Jerusalem.

Dennis Pardee. 2011. Ugaritic. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 460–472. De Gruyter.

Stephen Parkinson. 1997. Portuguese. In Martin Harris and Nigel Vincent, editors, *The Romance Languages*, pages 131–169. Routledge.

Na'ama Pat-El. 2019. Syriac. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 653–678. Routledge.

Tiago Pimentel, Arya D McCarthy, Damián Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to Form: Measuring Systematicity as Information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764.

T.M.S. Priestly. 2002. Slovene. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 388–454. Routledge.

Joseph F. Privitera. 1998. *Basic Sicilian : a brief reference grammar*. The Edwin Mellen Press.

Hamdi A. Qafisheh. 1980. *Yemeni Arabic I*. University of Arizona.

Elisabetta Ragagnin. 1998. Azeri. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 242–261. Routledge.

Neil Rathi, Michael Hahn, and Richard Futrell. 2021. An information-theoretic characterization of morphological fusion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10115–10120, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Terry Regier, Charles Kemp, and Paul Kay. 2015. Word Meanings across Languages Support Efficient Communication. In Brian MacWhinney and William O'Grady, editors, *The Handbook of Language Emergence*, 1 edition, pages 237–263. Wiley.

Caroline Jeanne Roset. 2018. *A Grammar of Darfur Arabic*. LOT.

Robert A. Rothstein. 2002. Polish. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 686–758. Routledge.

Erich Round, Louise Esher, and Sacha Beniamine. 2025. The natural stability of autonomous morphology: how an attraction–repulsion dynamic emerges from paradigm cell filling. *Morphology*, 35(1):1–49.

Aaron D. Rubin. 2014. A brief comparison of mehri and jibbali. *Proceedings of the Seminar for Arabian Studies*, Vol. 44:125–136.

Aaron D. Rubin. 2019. Mehri. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 257–279. Routledge.

Carmen Saldana, Borja Herce, and Balthasar Bickel. 2022. More or Less Unnatural: Semantic Similarity Shapes the Learnability and Cross-Linguistic Distribution of Unnatural Syncretism in Morphological Paradigms. *Open Mind*, 6:183–210.

Ernest A. Scatton. 2002. Bulgarian. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 188–248. Routledge.

Alexander M. Schenker. 2002. Proto-Slavonic. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 60–124. Routledge.

Ruth Laila Schmidt. 1999. *Urdu: An Essential Grammar*. Routledge.

Ruth Laila Schmidt. 2003. Urdu. In George Cardona and Dhanesh Jain, editors, *The Indo-Aryan Languages*, pages 415–385. Routledge.

Eckehard Schulz. 2013. *Modernes Hocharabisch - Lehrbuch mit einer Einführung in Hauptdialekte*. Edition Hamouda.

Georges De Schutter. 1994. Dutch. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 439–477. Routledge.

Alfred Senn. 1937. *Middle High German - A Grammar and Reader*. W. W. Norton & Co., Inc.

Christopher Shackle. 2003. Panjabi. In George Cardona and Dhanesh Jain, editors, *The Indo-Aryan Languages*, pages 637–682. Routledge.

George Y. Shevelov. 2002. Ukrainian. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 947–998. Routledge.

David Short. 2002a. Czech. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 455–532. Routledge.

David Short. 2002b. Slovak. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 533–592. Routledge.

Eduard Sievers. 1903. *An Old English grammar*. Ginn and Company.

Marie-Claude Simeone-Senelle. 2005. A Survey of the Dahalik language, an Afro-Semitic language spoken exclusively in Eritrea.

Marie-Claude Simeone-Senelle. 2011. Modern South Arabian. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 1073–1113. De Gruyter.

Kenny Smith. 2022. How Language Learning and Language Use Create Linguistic Structure. *Current Directions in Psychological Science*, 31(2):177–186. Publisher: SAGE Publications Inc.

Kenny Smith and Jennifer Culbertson. 2025. Communicative pressures shape language during communication (not learning): Evidence from case-marking in artificial languages. *Cognition*, 263:106164.

Kenny Smith, A. Perfors, Olga Fehér, Anna Samara, Kate Swoboda, and Elizabeth Wonnacott. 2017. Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160051.

Kenny Smith, Monica Tamariz, and Simon Kirby. 2013. Linguistic structure is an evolutionary trade-off between simplicity and expressivity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35(35).

Michael Sokoloff. 2011a. Jewish Babylonian Aramaic. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 610–619. De Gruyter.

Michael Sokoloff. 2011b. Jewish Palestinian Aramaic. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 660–670. De Gruyter.

Christian Stadel. 2019. Samaritan Aramaic. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 611–631. Routledge.

Peter Stein. 2011. Ancient South Arabian. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 1042–1072. De Gruyter.

Shane Steinert-Threlkeld and Jakub Szymanik. 2019. Learnability and semantic universals. *Semantics and Pragmatics*, 12(4):1–39.

Shane Steinert-Threlkeld and Jakub Szymanik. 2020. Ease of learning explains semantic universals. *Cognition*, 195:104076.

Gerald Stone. 2002a. Cassubian. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 759–794. Routledge.

Gerald Stone. 2002b. Sorbian. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 593–685. Routledge.

Michael P. Streck. 2011a. Babylonian and Assyrian. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 359–395. De Gruyter.

Michael P. Streck. 2011b. Eblaite and Old Akkadian. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 340–358. De Gruyter.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal Dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176, Valencia, Spain. Association for Computational Linguistics.

Shabo Talay. 2011. Arabic Dialects of Mesopotamia. In Stefan Weninger, editor, *The Semitic Languages: An International Handbook*, pages 909–920. De Gruyter.

Shabo Talay. 2017. *Slomo Surayt - Ein Einführungskurs ins Surayt-Aramäische (Turoyo)*. Bar Habraeus Verlag.

Habibullah Tegey and Barbara Robson. 1996. *A Reference Grammar of Pashto*. Center for Applied Linguistics, Washington, D.C.

Höskuldur Thráinsson. 1994. Icelandic. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 142–189. Routledge.

Alan Timberlake. 2002. Russian. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 827–886. Routledge.

Josef Tropper and Juan-Pablo Vita. 2019. Ugaritic. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 482–508. Routledge.

Mike Turner. 2019. Moroccan Arabic. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 458–481. Routledge.

Marijke J. van der Wal and Aad Quak. 1994. Old and Middle Continental West Germanic. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 72–109. Routledge.

Ans van Kemenade. 1994. Old and Middle English. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 110–141. Routledge.

Nigel Vincent. 1997a. Italian. In Martin Harris and Nigel Vincent, editors, *The Romance Languages*, pages 279–313. Routledge.

Nigel Vincent. 1997b. Latin. In Martin Harris and Nigel Vincent, editors, *The Romance Languages*, pages 26–78. Routledge.

Janet Watson, B Stalls, K al Razihi, and S Weir. 2006. Two texts from Jabal Rāziḥ, North-west Yemen.

Stefan Weninger. 2011. *The Semitic Languages: An International Handbook*. De Gruyter.

Max W. Wheeler. 1997a. Catalan. In Martin Harris and Nigel Vincent, editors, *The Romance Languages*, pages 170–208. Routledge.

Max W. Wheeler. 1997b. Occitan. In Martin Harris and Nigel Vincent, editors, *The Romance Languages*, pages 246–278. Routledge.

Aaron D. Wilson-Wright. 2019. The Canaanite Languages. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 509–532. Routledge.

Shijie Wu, Ryan Cotterell, and Timothy O'Donnell. 2019. Morphological irregularity correlates with frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5117–5126, Florence, Italy. Association for Computational Linguistics.

Yang Xu, Emmy Liu, and Terry Regier. 2020. Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, 4:57–70.

Abdurishid Yakup. 1998. Uyghur. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 409–423. Routledge.

Noga Zaslavsky, Jennifer Hu, and Roger P. Levy. 2021a. A Rate–Distortion view of human pragmatic reasoning? In *Proceedings of the Society for Computation in Linguistics 2021*, pages 347–348, Online. Association for Computational Linguistics.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression of color naming and its evolution. In *Proceedings of the National Academy of Sciences*.

Noga Zaslavsky, Mora Maldonado, and Jennifer Culbertson. 2021b. Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *Proceedings of the 43rd annual meeting of the cognitive science society*.

Noga Zaslavsky, Terry Regier, Naftali Tishby, and Charles Kemp. 2019. Semantic categories of artifacts and animals reflect efficient coding. In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society*.

Philip Zhakevich and Benjamin Kantor. 2019. Modern Hebrew. In John Huehnergard and Na'ama Pat-El, editors, *The Semitic Languages*, pages 571–610. Routledge.

Árpád Berta. 1998. Tatar and Bashkir. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 303–319. Routledge.

Éva Á. Csató and Lars Johanson. 1998. Turkish. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 195–223. Routledge.

A Representation Details

Here, we describe our approach to paradigm representation in detail, covering both form and meaning (feature) representations.

A.1 Representing Verbal Paradigms Across Language Families

Philosophy: Single Paradigms, No Variability between Verbs An important design choice in our study is to specifically focus on individual paradigms, without regard for variability across different verb classes or irregular verbs (except for the weak-strong distinction in Germanic – which translates to two fundamentally different ways of marking tense), to isolate the effect of the regularity of the paradigm-level form-meaning mapping. That is, we only include one or two (as described below) paradigms per language, deliberately excluding irregular verbs, morphological interaction with lexical stems, etc. When we include more than one paradigm for a single language, we model them fully separately. The motivation for this choice is that it allows us to *entirely control* for the effect of several proposed prominent models of complexity in morphological paradigms, including i-complexity (Ackerman and Malouf, 2013) and others (Cotterell et al., 2019; Wu et al., 2019): Our paradigms and all their counterfactual variants have the same (essentially zero) complexity under those models. We evaluate robustness to this choice in Appendix B.4.

Below, we describe in detail how we represented verbal paradigms for training the Seq2Seq models in different languages.

Afroasiatic Languages In many Afroasiatic languages (including the Semitic ones), verbs use a non-concatenative morphology based on *root consonants*. Therefore, we represent Semitic roots as *1-2-3*, where each number represents one root consonant, and fill in the vowel patterns, and suffixes. Doubled root consonants are represented as doubled numbers. For example Arabic "*(ana) a**ktab**u*" would be represented as *a12a3u* and "*(huwwa) **kuttib**a*" as *1u22i3a*.

Cushitic verbs use prefix conjugation (PC) and suffix conjugation (SC) as two separate conjugation classes. We include two different paradigms for Cushitic languages: one for PC and one for SC.

For those Afroasiatic languages in which verbs do not exhibit noncatenative morphology, we represent the stem as *123*, irregardless of it's length and consonant-verb makeup. Those languages usually have tonal elements, which we represent with *H*, *L*, *F* for a high, low or falling tone. For example Yemsa (Omotic) "*z**ag**ín*" would be represented as *123iHn*.

Semitic roots usually structurally form several different derived verbal stems conveying different derived meanings (e.g. reflexive, causative, intensive). We focus on the so called G-stem or Form-I-Stem[11], and choose only one specific voweling pattern, in case several vowel combinations exist (e.g. *-a-a-*, *-a-i-* etc. in Arabic). We disregard any derived stems as well as any further irregularities (such as weak root consonants) in the paradigms.

Germanic and Romance Languages Some Germanic verbs display a form of non-concatenative morphology (subsection 2.1) called *ablaut*, where the main vowel of the verb, the *stem vowel*, is changed in certain forms of the conjugation. Therefore, we represent all Germanic stems with their stem vowel as *1v2*, where *v* represents the stem vowel[12]. For example, the German verb forms "*(ich) empf**i**ng*" and "*(du) empf**ä**ngst*" would be represented as *1i2* and *1ä2st*. We include two different paradigms per Germanic language: one *strong verb* paradigm, where the past conjugation is based on ablaut, and one *weak verb* paradigm where the past conjugation is based on affixed dentals.

For Romance languages, we only use paradigms without internal changes, and represent all paradigm forms as pure suffixes. Romance languages usually have three basic conjugation classes (*-a-*, *-e-*, *-i-*). We only use one of the conjugations per language, since the overall structure of the paradigms is mostly not affected by the specific choice of the class.

We disregard any further kinds of irregular verbs, including stem vowel shift or diphthongization in the Romance paradigms.

Some languages have more than one realization for certain feature combinations, irregardless of flexion classes. Especially in cases where this effects the syncretism pattern of the paradigm (e.g. in Serbian certain short accusative forms can be syncretized with the genitive forms, or have distinct realizations) or might have an effect on the learnability of the paradigm (e.g. in Ge'ez the 3rd person plural can have a very similar form to the 3rd person singular, or have an unrelated form), we use distinct paradigms.

Representations of Forms The exact representation of the phonemes in the languages depend on the specific language. For Semitic and Afroasiatic languages we use transliterations that convey a phonemic representation of the language. Long vowels and geminated consonants are represented as double letters. Tonal elements are represented with capital letters put after the affected vowel.

For Germanic and Romance languages, we only depart from the official orthography in cases where phonemes are inherently different from writing, and might affect the syncretism patterns and/or learnability. This is the case for example in French where the verb forms "*(je) mang**e***", "*(tu) mang**es***", "*(ils) mang**ent***" are all pronounced the same. For languages marking stress orthographically (including Latin) with accents we keep this in the representation. For Germanic we chose verbal classes (i.e. which specific vowel combinations are used) such that interference between orthography and phonemes is minimal.

For pronouns we use the official orthography for languages originally written in Latin script, and transliterations representing the phonemes for languages originally written in other scripts.

[11] Often also called the fa3al-stem in Arabic linguistics, and the qal or pa'al stem in Hebrew linguistics.

[12] In some languages and conjugation classes, this vowel consists of two consecutive vowels.

A.2 Representing Features

Feature Representation As described in section 3, the speaker's mental representation of a meaning for a target referent t is a probability distribution $m_t(u)$ over $u \in \mathcal{U}$. Zaslavsky et al. (2021b) model $m_t(u)$ as assigning probability to different u depending on the overlap in feature representations to t:

$$m_t(u) \propto \exp(-\gamma \cdot d(u,t)). \quad (6)$$

To define those feature encodings, each meaning is mapped to a vector in a multi-dimensional discrete feature-space. The original conceptual space from Zaslavsky et al. (2021b) uses a custom binary encoding based on (1) the communicative roles a, o, s, (2) binary encoding of a three-way number distintion. We find this encoding not scalable to the larger meaning we require for verbs, and thus resort to a general categorical encoding to sidestep the need for arbitrary binarization. For a pair of feature representations u, t, we then define $d(u,t)$ as the number of dimensions i where $u_i \neq t_i$.

Dimensions of Paradigms Here, we describe the details of the paradigm dimensions in each of the three domains. Table 5 shows an overview of the features used for each data set.

PPD For the basic pronouns, each meaning has the form $m = \langle number, person \rangle$ where
$number \in \{singular, plural\}$, and
$person \in \{1, 2, 3, 12\}$.[13]
For the PPD experiments, all language paradigms use all possible values, irregardless of whether the distinction is actually existing in the language.

VERB For the verbs, we define
$m = \langle number, person, gender, tense \rangle$ where
$number \in$ {singular, dual, plural},
$person \in$ {1, 2, 3},
$gender \in$ {masculine, feminine}, and
$tense \in$ {perfective, imperfective, subjunctive, jussive, converb}.
For the VERB experiment each language uses the number, person, and gender values that exist in the language family, irregardless of weather the specific language utilized them (e.g. the German paradigm uses dual, even though dual does not exist in modern Germanic languages anymore, but was existing in Proto-Germanic). Tense values depend on the specific language.

[13] Here, 12 represents inclusive plural.

Table 14 lists the languages used for each paradigm type for the VERB paradigms.

PRON For detailed pronoun paradigms, we define

$$m = \langle number, person, gender, case \rangle$$

and add a couple of possible values for the features depending on whether the respective language uses them. For the $person$ we add 3 different forms of formality for the 2nd person (*2informal*, *2formal*, *2highformal*), 3 different types of remoteness for the 3rd person (*3proximal*, *3distal*, *3remote*) and a *neutral_reflexive* form as used in Spanish or Russian. Some languages might have a combination of formality and remoteness in the 3rd person, for which we add further possibilities (e.g. *3formalproximal*, *3formalremote* etc.). For the $gender$ we add a $neutral$ gender as used in Germanic and an *animate* vs. *inanimate* distinction as used in some Slavic languages. We use a wide range of cases depending on the specific language (*nominative*, *genitive*, *dative*, *accusative*, *locative*, *ablative* etc.).
Table 15 lists the languages used for each paradigm type for the PRON paradigms.

B Modeling Details

B.1 Details for Estimating Need Probabilities

Efficient communication models require a communicative need distribution, i.e. the prior term $p_{cog}(t)$ in Equation 2. For simple pronoun paradigms (PPD), we used the need distribution that Zaslavsky et al. (2021b) had estimated from various corpora. For the more complex pronoun (PRON) and verb (VERB) paradigms, we started from these distributions, treating these as marginal distributions of each combination of a person and binary number (singular vs dual/plural). We then obtained relative frequencies of dual versus plural and masculine vs feminine from the NUDAR Treebank (Taji et al., 2017)[14], a large treebank for a language (Standard Arabic) that marks all distinctions relevant to the Semitic paradigms experiment. We collapsed the frequencies of those forms that are not distinguished in the language, and assumed equal distribution for further features (such as tense and case). Resulting probabilities for VERB are shown in Table 7 and Figure 6.

[14] https://github.com/UniversalDependencies/UD_Arabic-NYUAD

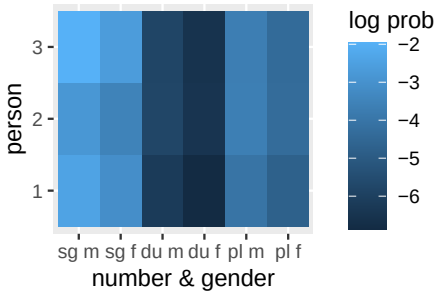


Figure 6: Estimated need probability for person-number-gender combinations , obtained by combining need probabilities estimated by Zaslavsky et al. (2021b) with corpus counts for gender and the dual/plural distinction. These collapse the inclusive/exclusive distinction; we keep that distinction (with relative frequencies from Zaslavsky et al. (2021a)) in those paradigms where it is relevant. See subsection B.1 for details.

Discussion An alternative would have been to obtain counts for all entries in the paradigm table; however, we opted against this as the written genres (e.g., newspaper text) represented in the available treebank data show overwhelming biases (e.g., towards third-person forms) unlikely to be representative of spoken language. An ideal solution would be to obtain counts for large conversational corpora of the relevant languages; however, such data is not available with suitable annotation at the required scale. We show the robustness of our results to other distributions in subsection B.2, where we evaluated with a uniform and three different randomly generated need distributions.

B.2 Robustness towards Frequency Distributions

We create a uniform distribution and 3 different random distributions of frequencies, and test them on the Classical Arabic verbal paradigm, to verify that our results are robust to the choice of frequency distributions. Figure 7 shows the efficiency plots, Figure 8 and Table 8 show the correlation results. We see a stably strong correlation for CETL across the different frequency distributions. We note that the need distribution may still play a substantial role in efficient coding (e.g. in accounting for the detailed typological patterns studied by Zaslavsky et al. (2021b)), but these analyses show that our results are robust to the choice of the need distribution.

Table 9 lists the hit and fail rates. The need distribution does have an effect on the performance of the models. For one of the random distributions the overall performance is slightly better for MI, for the other distributions, as well as the corpus-based distribution explained in the main part of this paper, CETL performs much better than MI.

B.3 Comparison using Feature Representation and Weighting from Zaslavsky et al. (2021b)

Recall from Appendix A.2 that meanings of referents t are formalized as probability distributions:

$$m_t(u) \propto \exp(-\gamma \cdot d(u,t)). \tag{7}$$

where $d(u,t)$ is a weighted Hamming distance based on feature encodings of the individual references. As described in that section, Zaslavsky et al. (2021b) defined $d(\cdot,\cdot)$ in terms of a weighted 5-bit vector feature representation with specific weights, whereas we use a more generic discrete feature vector space for our analyses that easily accommodates further features (e.g., gender, tense) even when they are not binary.

Here, in the PPD domain, we compare the CETL- and MI-based models directly using the 5-bit-vector feature representation and feature weights used by Zaslavsky et al. (2021b). We used three different values for the free parameter γ (Equation 7), and compute the efficiencies for our PPD dataset (Figure 9 and Figure 10). Overall, even in this setup, CETL provides stronger performance, even on structural permutations. This shows that, while the specific feature weights used by Zaslavsky et al. (2021b) may be important in accounting for specific typological patterns, the improved discrimination between real and counterfactual paradigms provided by the CETL-based model is general and robust to the weighting of different features.

B.4 Robustness towards Paradigm Classes

As discussed in Appendix A, our analyses do not consider variability between different verbs, in order to control for the effect of previously-proposed approaches such as i-complexity (Ackerman and Malouf, 2013). Here, we consider an alternative modeling choice, where we jointly model different paradigms in a language, and show that it leads to qualitatively equivalent results. We use two language families in which verbs tend to fall into two classes, with very distinct morphology: Germanic, where verbs have a strong and weak conjugation class, and Cushitic, where verbs have a prefix-based and a suffix-based conjugation class. We extend the meaning space by

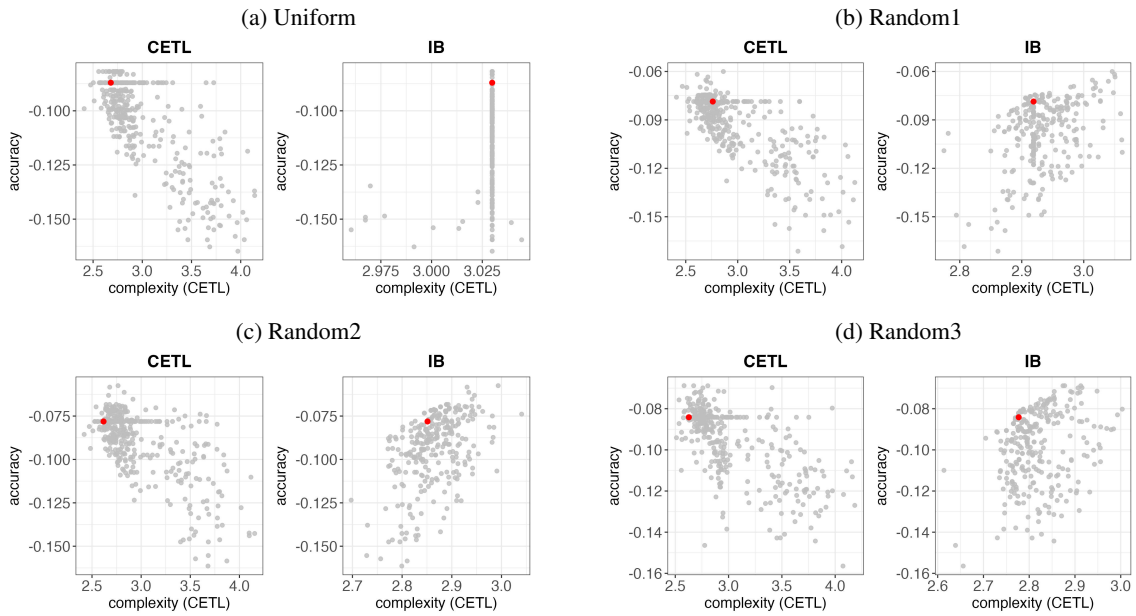


Figure 7: Results for Appendix B.2: Efficiency of permutations: Accuracy plotted (inverted) against complexity measures for a uniform distribution and three random distributions of feature frequencies.

a fifth dimension *class* for conjugation classes. $m = \langle number, person, gender, tense, class \rangle$, where $class \in \{weak, strong, pc, sc\}$. We create counterfactuals which permute inside the conjugation classes only, and ones which permute across the conjugation classes. Figure 11 shows the efficiency plots. Figure 12 and Table 10 shows the correlation results. Attested paradigms are substantially more efficient than most counterfactuals, even when permutations are applied only within conjugation classes.

We further create similar paradigms for Classical Arabic. Semitic verbs are classified into classes based on (i) whether their root consonants contain a glide (*w* or *y*; "weak verbs") or no glide ("strong verbs"), and (ii) whether they are directly inflected from a root ("Stem I"), or the result of a derivational process ("Stems II, III, . . . , X"). We create two different paradigms for Classical Arabic, one containing the 10 main stems as conjugation classes, and one containing one strong and different weak forms of Stem I as classes. Figure 13 shows the efficiency plots. Figure 14 and Table 11 shows the correlation results. For both cases, the attested paradigms are more efficient than most counterfactuals.

B.5 Relation to Informational Fusion

Here, we discuss the relation between CETL and another information-theoretic metric for morphological paradigms, Informational Fusion (Rathi et al., 2021). The Informational Fusion of a pair of features (e.g., 2nd person & plural) is defined as the cross-entropy that a seq2seq model trained to produce all paradigm cells not involving this pair experiences on predicting the cells involving this pair. Like CETL, but unlike the measures of (Ackerman and Malouf, 2013; Cotterell et al., 2019; Wu et al., 2019), Informational Fusion is defined even on a single paradigm (a single pronoun paradigm, or an affix set applicable across verbs). Informational Fusion is related to CETL in that both measures quantify the difficulty of learning a paradigm; the difference being that CETL assumes that the whole paradigm is learned progressively, whereas Informational Fusion assumes that all forms excluding a specific feature pair are learned to convergence before encountering that pair. One important difference between the two measures is that Informational Fusion requires retraining seq2seq models for each feature combination. Given the large number of languages and counterfactual paradigms, computing Informational Fusion would not have been feasible in our study.

C Additional Results

C.1 Efficiency Plots and Significance Tests

Figure 15 shows the efficiency of factual vs. counterfactual paradigms for the CETL vs. the IB model for all three domains.

We performed two-sided one-sample Student's t-tests (Student, 1908) comparing the distribution of relative CETL and relative accuracy to 0 (the relative CETL and accuracy of the baseline distribution) for each domain. Table 12 lists the results.

C.2 Naturalness Plots

The figures in this section show the correlation between complexity measure and naturalness for our CETL model vs. the original IB model for PPD (Figure 16), PRON (Figure 17 and Figure 18), and VERB (Figure 19), separated by language families and individual languages.

C.3 Correlation Values

Table 13 shows the averaged per-langauge correlation values for our CETL model vs. the IB model, separated by domains. Figure 21 shows the variance of the per-language correlations for the two models in our three domains. Figure 22, Figure 23 and Figure 24 show the variance of the per-language correlation separated by language family for PPD, PRON and VERB, respectively.

C.4 Model Comparison

Figure 20 shows the amount of languages per domain for which the CETL model outperforms the IB model. We consider CETL to outperform the IB model when it performs better by at least 5% of the permutations for a given language.

D Data Sources

D.1 Domain 1 (Verbs)

Semitic Languages Proto-Semitic based on Huehnergard (2019); Lipiński (1997).

Classical Arabic based on Birnstiel (2019) and Schulz (2013), Levantine Arabic based on Brustad and Zuniga (2019) (Beiruti and Rural), Othman (2019) (Jerusalemite). Nilo-Egyptian Arabic bases on Leddy-Cecere and Schroepfer (2019) (Egyptian), Khalafallah (1969) (Saidi), Roset (2018) (Darfur). Maghrebi Arabic bases on Turner (2019) (Moroccan), https://en.wikipedia.org/wiki/Tunisian_Arabic_morphology (Tunesian), https://en.wikipedia.org/wiki/Algerian_Arabic (Algerian), Institute of Islamic Studies of University of Zaragoza (2013) (Andalusian), Falzon (1997) (Maltese). Mesopotamia Arabic based on Abu-Haidar (1991) and Talay (2011) . Peninsular Arabic based on Omar (1975) (Hejazi), Al-Balushi (2017) (Omani), Qafisheh (1980) (Yemeni).

Hebrew based on Hornkohl (2019) (Biblical), Zhakevich and Kantor (2019) (Modern), Phoenician based on Briquel Chatonnet and Hawley (2020) and https://en.wikipedia.org/wiki/Phoenician_language (Standard), https://en.wikipedia.org/wiki/Punic_language (Punic). Aramaic based on Gzella (2011) (Imperial Aramaic), Stadel (2019) (Samaritan), Fassberg (2019) (Siryon), Sokoloff (2011b) (Jewish Palestinian), Burtea (2011) (Classical Mandaic), Häberl (2019) (Neo-Mandaic), Pat-El (2019) (Syriac), Sokoloff (2011a) (Jewish Babylonian), Coghill (2019) (Suret), Jastrow (2011) (Ṭuroyo), Talay (2017) (Surayt), Jastrow (2011) (Mlaḥsô). Ancient South Arabian based on Multhoff (2019), Stein (2011) and https://en.wikipedia.org/wiki/Sabaic. Ugaritic based on Pardee (2011), Tropper and Vita (2019)

Ge'ez based on Butts (2019), Tigre based on Elias (2019), Tigrinya based on Bulakh (2019), Amharic based on Edzard (2019), Muher based on Meyer (2019), Meyer (2011)

Sargonic based on Gelb (1952), Streck (2011b) and Hasselbach (2005), Old Assyrian based on Kouwenberg (2017) and Huehnergard and Pat-El (2019a), Old Babylonian based on Hasselbach-Andee (2019), Lipiński (1997) and Streck (2011a)

Mehri based on Rubin (2019), Soqoṭri based on Kogan and Bulakh (2019), Jibbali based on Rubin (2014), Simeone-Senelle (2011).

Non-Semitic Afroasiatic Languages Berber based on https://de.wikipedia.org/wiki/Berbersprachen#Verbalmorphologie (Tamasheq), Abdel-Massih (1971) (Tamazight), Kossmann (2012) and https://de.wikipedia.org/wiki/Siwi (Siwa), Gragg (2019) (Standard Berber), Kossmann (2013) (Ghadames).

Cushitic based on Gragg (2019) (Afar), Gragg (2019) (Beja), Nilsson (2024) (Somali).

Hausa based on Frajzyngier and Shay (2012a), Moloko based on Friesen (2017). Yemsa based on Frajzyngier and Shay (2012a).

Ancient Egyptian based on Frajzyngier and Shay (2012b), Gragg (2019) and https://en.wiktionary.org/wiki/.k#Egyptian (Early); Frajzyngier and Shay (2012b), Gragg (2019) and https://en.wiktionary.org/wiki/.k#Egyptian (Middle/Late); Frajzyngier and Shay (2012b), Gragg (2019) and

https://en.wiktionary.org/wiki/Appendix:Coptic_verbs, https://en.wiktionary.org/wiki/âšŹâš§ÏčâšŞ#Coptic (Coptic).

Germanic Languages Faroese based on Barnes and Weyhe (1994), Icelandic based on Thráinsson (1994), Swedish based on Andersson (1994), Old Norse based on Faarlund (1994), Proto-Germanic based on Lehmann (2007), Gothic based on Lehmann (1994), Standard High German based on Eisenberg (1994), Middle High German based on Senn (1937), Old High German based on Ellis (1953), Modern English based on König (1994), Middle English based on Fulk (2012), Old English based on Sievers (1903), Modern Dutch based on Schutter (1994), Middle Dutch based on Hüning and Vogl (2009), Old Dutch based on van der Wal and Quak (1994) Old Saxon based on Cathey (2000) and van der Wal and Quak (1994)

Romance Languages Catalan based on Fabra (2006), Author (1999). Corsican based on https://en.wiktionary.org/wiki/parlà#Corsican. Franco-Provençal based on https://en.wiktionary.org/wiki/chantar#Franco-Provençal. French based on Harris (1997) and https://fr.wiktionary.org/wiki/Conjugaison:français/regarder. Galician based on https://en.wiktionary.org/wiki/falar#Galician. Italian based on Vincent (1997a). Latin based on Vincent (1997b) and https://verbix.com/webverbix/go.php?&D1=9&T1=canto. Occitan based on Wheeler (1997b). Portuguese based on Parkinson (1997). Romanian based on Mallinson (1997). Spanish based on Green (1997).

D.2 Domain 2 (Pronouns)

Semitic and Afro-Asiatic Languages Berber, Chadic, Omotic and Ancient Egyptian based on Gragg (2019), Cushitic based on Appleyard (2011) (Afar, Oromo, Sidaama) and Gragg (2019) (Alaaba, Beja, Bilin, Burunge, Somali, Tsamakko).

Akkadian based on Hasselbach-Andee (2019).

Ge'ez based on Butts (2019), Tigre based on Elias (2019), Tigrinya based on Bulakh (2019), Amharic based on Edzard (2019), Muher based on Meyer (2019).

Argobba based on https://en.wikipedia.org/wiki/Argobba_language, Dahalik based on Simeone-Senelle (2005), Razihi based on Watson et al. (2006), Soddo based on https://en.wikipedia.org/wiki/Soddo_language.

Mehri based on Rubin (2019), Soqoṭri based on Kogan and Bulakh (2019).

Classical Arabic based on Birnstiel (2019), Levantine Arabic based on Brustad and Zuniga (2019), Moroccan Arabic based on Turner (2019), Egyptian Arabic based on Leddy-Cecere and Schroepfer (2019).

Ugaritic based on Tropper and Vita (2019).

Biblical Hebrew based on Hornkohl (2019), Modern Hebrew based on Zhakevich and Kantor (2019). Phoenician and Punic based on Hackett (2008), Wilson-Wright (2019), https://en.wikipedia.org/wiki/Punic_language#Personal_pronoun and https://en.wikipedia.org/wiki/Phoenician_language#Nominal_morphology.

Aramaic based on Stadel (2019) (Samaritan), Häberl (2019) (Mandaic), Pat-El (2019) (Syriac), Sokoloff (2011a) and Bar-Asher Siegal (2013) (Jewish Babylonian), Fassberg (2019) (Modern Western), Coghill (2019) (Modern North-Eastern), Jastrow (2011) (Modern Central).

Proto-Semitic based on https://en.wikipedia.org/wiki/Proto-Semitic_language#Pronouns and Huehnegard (2008).

Germanic Languages High German based on Ellis (1953) (Old), Senn (1937) (Middle), Eisenberg (1994) (Modern). English based on van Kemenade (1994) and https://en.wiktionary.org/wiki/git#Old_English (Old), https://en.wikipedia.org/wiki/Middle_English (Middle),https://en.wikipedia.org/wiki/English_personal_pronouns (Modern), https://en.wikipedia.org/wiki/English_personal_pronouns#Complete_table (Archaic). Dutch based on https://en.wiktionary.org/wiki/ik#Old_Dutch (Old), van der Wal and Quak (1994) (Middle), Schutter (1994) (Modern). Low German based on https://en.wiktionary.org/wiki/ik#Old_Saxon (Old), https://en.wiktionary.org/wiki/ik#Middle_Low_German (Middle), https://de.wikipedia.org/wiki/Niederdeutsche_Sprache#Pronomen (Modern). Bavarian based on https://en.wikipedia.org/wiki/Bavarian_language#Pronouns. Afrikaans based on Donaldson (1994). Faroese on Barnes and Weyhe (1994). Frisian based on https://en.wiktionary.org/wiki/ik#Old_Frisian (Old), Hoekstra and Tiersma (1994) (Modern). Gothic based on Lehmann (1994). Old

Norse based on Faarlund (1994). Icelandic based on https://en.wikipedia.org/wiki/Icelandic_grammar#Pronouns. Swedish based on Andersson (1994). Norwegian based on Askedal (1994). Proto-Germanic based on https://en.wiktionary.org/wiki/Reconstruction:Proto-Germanic/ek#Proto-Germanic. Old Prussian based on https://en.wikibooks.org/wiki/Prussian/Personal_Pronouns_Chart.

Romance Languages Dalmatian based on https://en.wikipedia.org/wiki/Dalmatian_grammar#Pronouns. Rumantsch based on https://de.wikipedia.org/wiki/Grammatik_des_Rumantsch_Grischun#Personalpronomen. Corsican based on https://en.wiktionary.org/wiki/eiu#Corsican. Emilian based on https://en.wiktionary.org/wiki/mè#Emilian. Franco-Provençal based on https://en.wiktionary.org/wiki/o#Franco-Provençal. French based on Harris (1997). Occitan based on Wheeler (1997b). Sicilian based on Privitera (1998) and https://it.wikipedia.org/wiki/Lingua_siciliana#Pronomi. Latin based on Vincent (1997b). Romanian based on Mallinson (1997). Catalan based on Wheeler (1997a), Author (1999). Galician based on https://en.wiktionary.org/wiki/eu#Galician. Italian based on Vincent (1997a). Portuguese based on Parkinson (1997). Spanish based on Moriena and Genschow (2004).

Balto-Slavic Languages Proto-Slavic based on Schenker (2002) and https://en.wiktionary.org/wiki/Template:sla-decl-ppron. Belorussian based on Mayo (2002). Latvian based on Kalnača and Lokmane (2021). Lithuanian based on Mathiassen (1996). Old Church Slavonic based on Huntley (2002). Russian based on Timberlake (2002). Ukrainian based on Shevelov (2002). Bulgarian based on Scatton (2002). Macedonian based on Friedman (2002). Czech based on Short (2002a). Kashubian based on Stone (2002a). Polish based on Rothstein (2002). Serbo-Croatian based on Browne (2002). Slovak based on Short (2002b). Slovene based on Priestly (2002). Sorbian based on Stone (2002b), https://en.wikibooks.org/wiki/Lower_Sorbian/Grammar/Pronouns, https://en.wiktionary.org/wiki/ja#Lower_Sorbian and https://en.wiktionary.org/wiki/ja#Upper_Sorbian.

Altaic Languages Mongol based on Janhunen (1952), https://en.wikipedia.org/wiki/Mongolian_language#Pronouns and https://en.wiktionary.org/wiki/Template:mn-personal_pronouns. Manchu based on https://en.wikipedia.org/wiki/Manchu_language#Pronouns Udihe based on http://www.tufs.ac.jp/ts/personal/kazama/shigen/18/Kazama.pdf

Southern Altai based on https://en.wikipedia.org/wiki/Altai_languages#Morphology_and_syntax. Chuvash based on Agyagási (1998). Tuvan based on https://en.wiktionary.org/wiki/Template:Tuvan_personal_pronouns. Kazakh based on Abish (1998). Azeri based on Ragagnin (1998). Bashkir based on Árpád Berta (1998) and https://en.wikipedia.org/wiki/Bashkir_language#Declension_table. Crimean based on Kavitskaya (2009). Tartar based on Árpád Berta (1998) and Burbiel (2018). Turkmen based on Karakoç (1998). Kyrgyz based on Karakoç and Kalieva (1998). Turkish based on Éva Á. Csató and Johanson (1998). Uyghur based on Yakup (1998). Uzbek based on Boeschoten (1998). Old-Turkic Erdal (2004). Proto-Turkic based on https://en.wiktionary.org/wiki/Reconstruction:Proto-Turkic/bẹ, https://en.wiktionary.org/wiki/Reconstruction:Proto-Turkic/sẹ, https://en.wiktionary.org/wiki/Reconstruction:Proto-Turkic/bíŕ, https://en.wiktionary.org/wiki/Reconstruction:Proto-Turkic/síŕ and https://en.wiktionary.org/wiki/Reconstruction:Proto-Turkic/ol.

Indo-Iranian Languages Sanskrit based on Gonda (1966). Urdu based on Schmidt (2003) and Schmidt (1999). Assamese based on Goswami and Tamuli (2003). Bengali based on Dasgupta (2003) and https://en.wikipedia.org/wiki/Bengali_grammar#Pronouns. Gujurati based on Cardona and Suthar (2003). Sindhi based on Khubchandani (2003). Kashmiri based on Koul (2003) and Koul and Wali (2006). Punjabi based on Shackle (2003).

Gilaki based on https://en.wikipedia.org/wiki/Gilaki_language#Pronouns. Pashto based on David (2014) and Tegey and Robson (1996). Persian based on https://en.wiktionary.org/wiki/Template:

prs-personal_pronouns and https://en.wiktionary.org/wiki/Template:fa-personal_pronouns. Kurdish based on https://en.wikipedia.org/wiki/Central_Kurdish#Grammar_and_Syntax and McCarus (2009). Ossetian https://en.wikipedia.org/wiki/Ossetian_language#Pronouns

Other Languages Ancient Greek based on Lahmer (2018). Modern Greek Holton et al. (2004) Albanian based on Newmark et al. (1982). Classical Armenian based on https://en.wiktionary.org/wiki/ÕěÕi#Old_Armenian, https://en.wiktionary.org/wiki/ÕďÕÿÖĆ#Old_Armenian and https://en.wiktionary.org/wiki/ÕűÕą#Old_Armenian. Eastern Armenian based on https://en.wikipedia.org/wiki/Template:Armenian_personal_pronoun_table. Circassian based on https://en.wikipedia.org/wiki/Circassian_pronouns (Adyghe) and https://en.wikipedia.org/wiki/Kabardian_grammar#Pronouns (Kabardian). Georgian based on https://en.wikibooks.org/wiki/Georgian/Pronouns. Proto-Indo-European based on https://en.wikipedia.org/wiki/Proto-Indo-European_pronouns#Personal_pronouns.

	subsets	feature sets
PRON		PERS = {1,2,3,12}; NUM = {s,p}
PR	PR_SEM	PERS = {1,2,3}; NUM = {s,d,p}; GEN = {m,f}; CASE = {isolated, suffixal}
	PR_AFRO	PERS = {1,2,3}; NUM = {s,d,p}; GEN = {m,f}; CASE = {isolated, suffixal}
	PR_GER	PERS = {1,2,3,2v,2h}; NUM = {s,d,p}; GEN = {m,f,n}; CASE = {nom, gen, dat, akk}
	PR_ROM	PERS = {1,2,3,r,2v,2h}; NUM = {s,p}; GEN = {m,f,n}; CASE = {nom, gen, dat, akk, disj, con, akk-l, dat-l}
	PR_BALTSLAV	PERS = {1,2,3,r}; NUM = {s,d,p}; GEN = {m,f,n,v,i}; CASE = {nom, gen, dat, akk, instr, loc, gen-s, dat-s, akk-s}
	PR_INDOIRAN	PERS = {1,2,3,2v,2h,2j,3e,3x,3w,12, 3ev,3eh,3ej,3xv,3xh,3xj,3wv,3wh,3wj}; NUM = {s,d,p}; GEN = {c,m,f,n}; CASE = {nom, obl, gen, akk, dat, loc, instr, abl, erg, indir, obj}
	PR_ALTAIC	PERS = {1,2,3,2v,2h,12}; NUM = {s,p}; GEN = {c}; CASE = {nom, gen, dat, akk, abl,instr, loc, alla, equa, simi, comm, prol, dir}
	PR_OTHER	
VERB	VERB_SEM	PERS = {1,2,3}; NUM = {s,d,p}; GEN = {m,f}; TEN = {perf, imperf, subj, juss, conv}
	VERB_AFRO	PERS = {1,2,3}; NUM = {s,d,p}; GEN = {m,f}; TEN = {perf, imperf, subj}
	VERB_GER	PERS = {1,2,3}; NUM = {s,d,p}; GEN = {m,f}; TEN = {pres, pret}
	VERB_ROM	PERS = {1,2,3}; NUM = {s,p}; GEN = {m,f}; TEN = {pres, imperf, perf, fut}

Table 5: Details for Appendix A.2 - Summary of the features used in each data set. For abbreviation definitions, see Table 6.

	PERS
1	1st person
2	2nd person
3	3rd person
12	1st person inclusive
2v	2nd person familiar
2h	2nd person formal
2j	2n person very formal
r	neutral-reflexive
3e	3rd person proximal
3x	3rd person distal
3w	3rd person remote

	PERS
3ev	3rd person proximal familiar
3eh	3rd person proximal formal
3ej	3rd person proximal very formal
3xv	3rd person distal familiar
3xh	3rd person distal formal
3xj	3rd person distal very formal
3wv	3rd person remote familiar
3wh	3rd person remote formal
3wj	3rd person remote very formal

	NUM
s	singular
d	dual
p	plural

	GEN
m	masculine
f	feminine
n	neutral
v	virile / animate
i	non-virile / inanimate
c	common

	TEN
perf / V	perfective
imperf / G	imperfective
subj / S	subjunctive
juss / J	jussive
conv / C	converb
pres / G	present
pret / V	preterite
fut / S	future

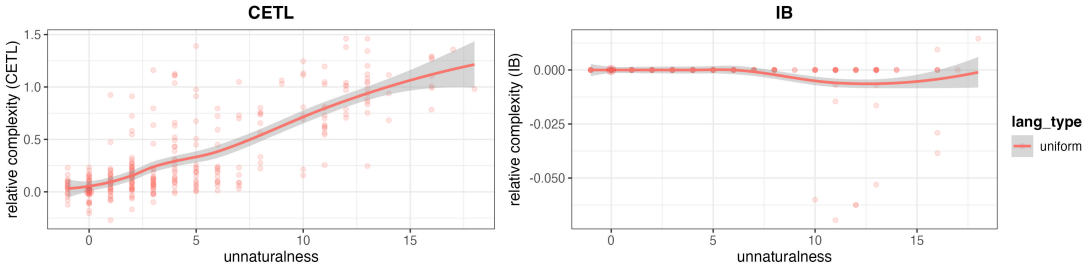
	CASE
nom	nominative
gen	genitive
dat	dative
akk	accusative
disj	disjunctive
instr	instrumental
loc	locative
abl	ablative
con	con-version

	CASE
akk-l	long accusative
dat-l	long dative
gen-s	short genitive
dat-s	short dative
akk-s	short accusative
obl	oblique
erg	ergative
indir	indirect
obj	object

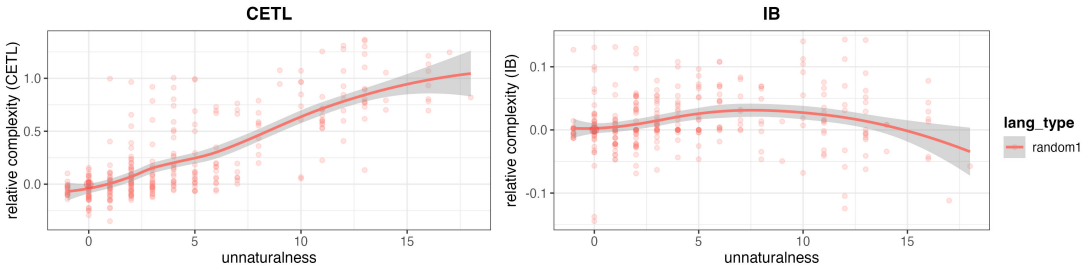
	CASE
alla	allative
equa	equative
simi	similative
comm	comitative
prol	prolative
dir	directive

Table 6: Details for Appendix A.2 - Feature Abbreviations. Note that we encode verb tense as V, G, S, J, or C, and use the case "con" corresponding with the special pronoun *con/cum/com* used in many Romance languages.

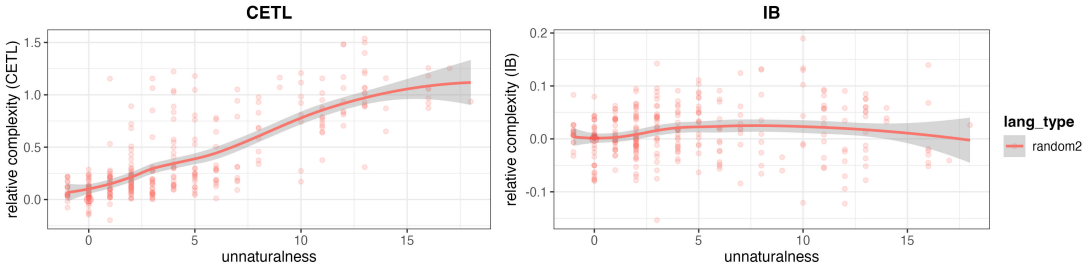
(a) Uniform



(b) Random1



(c) Random2



(d) Random3

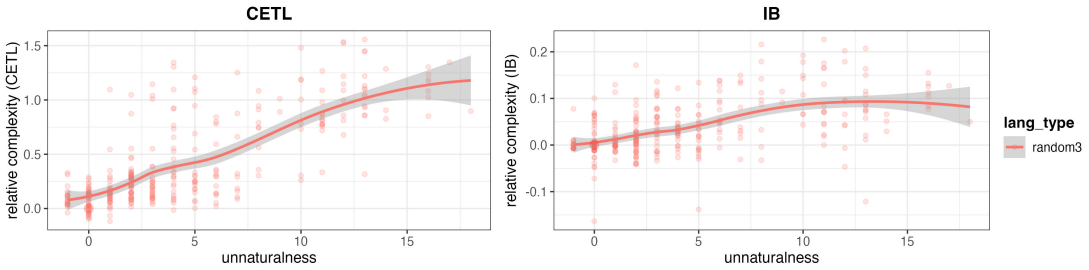


Figure 8: Results for Appendix B.2: Complexity of permutations of different naturalness: Complexity measures plotted against unnaturalness for a uniform distribution and three random distributions of feature frequencies.

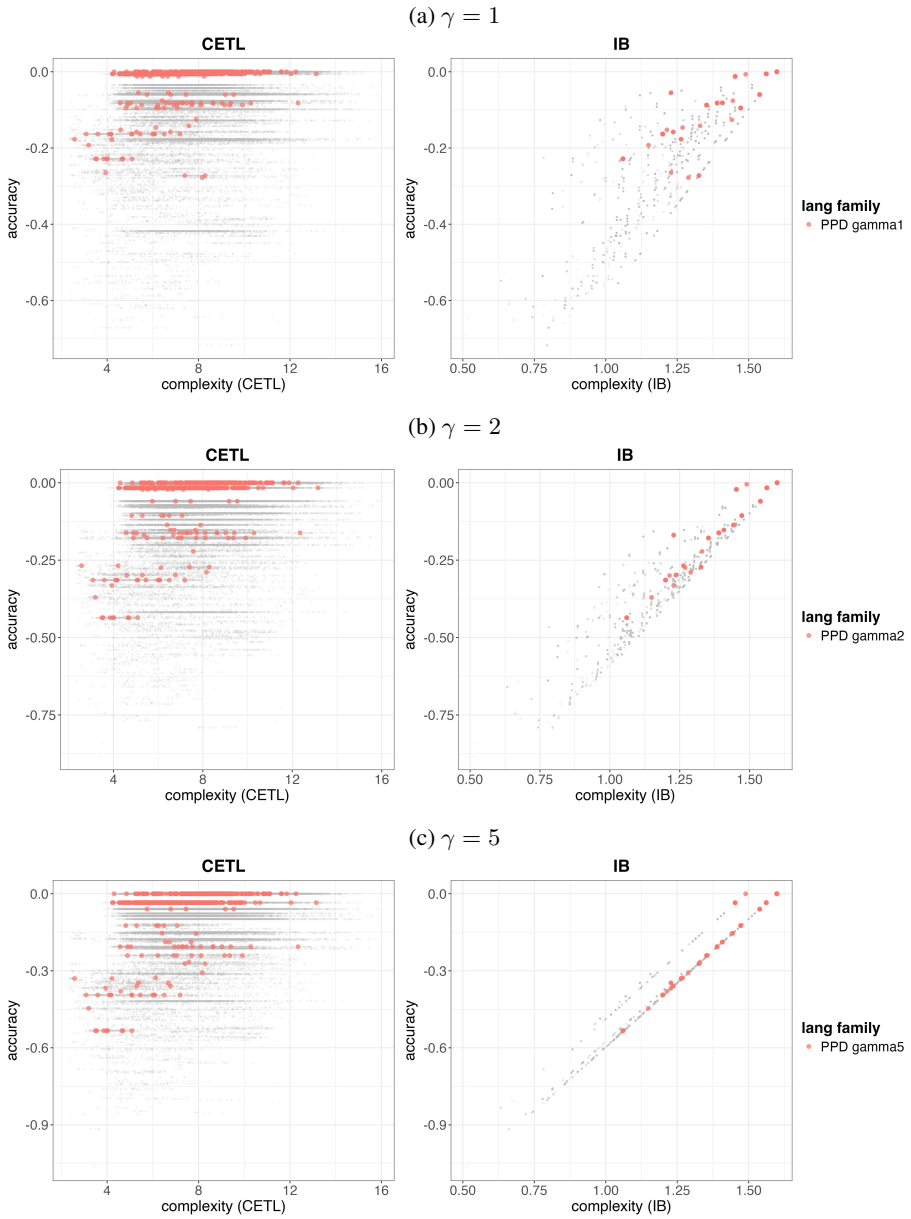
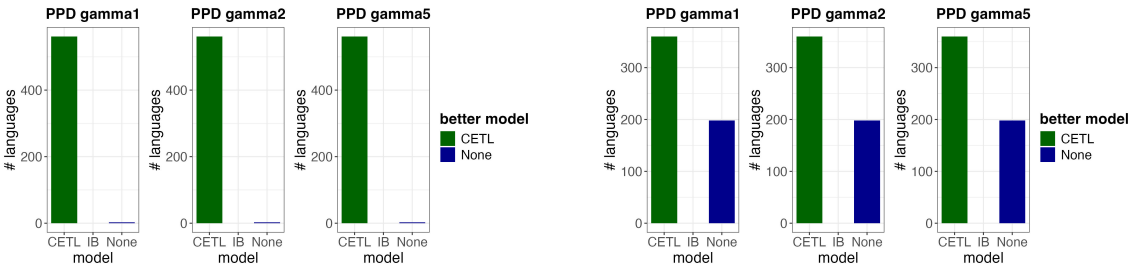


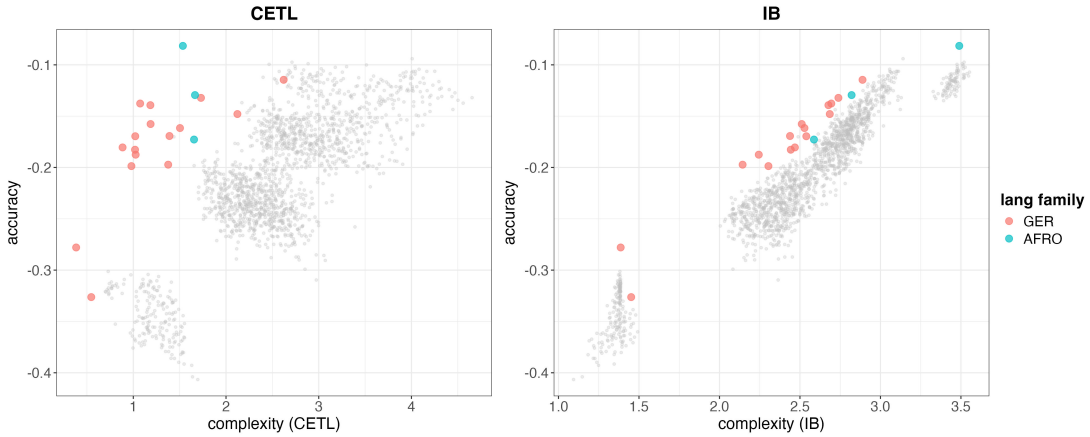
Figure 9: Results for Appendix B.3 - Efficiency of permutations. Accuracy plotted against complexity measures, using Zaslavsky et al. (2021b)'s feature representation, with $\gamma = 1$ (a), $\gamma = 2$ (b) and $\gamma = 5$ (c). Here, we used the feature representations from Zaslavsky et al. (2021b), with their feature weights. The model has a free parameter γ; we show results for three values.



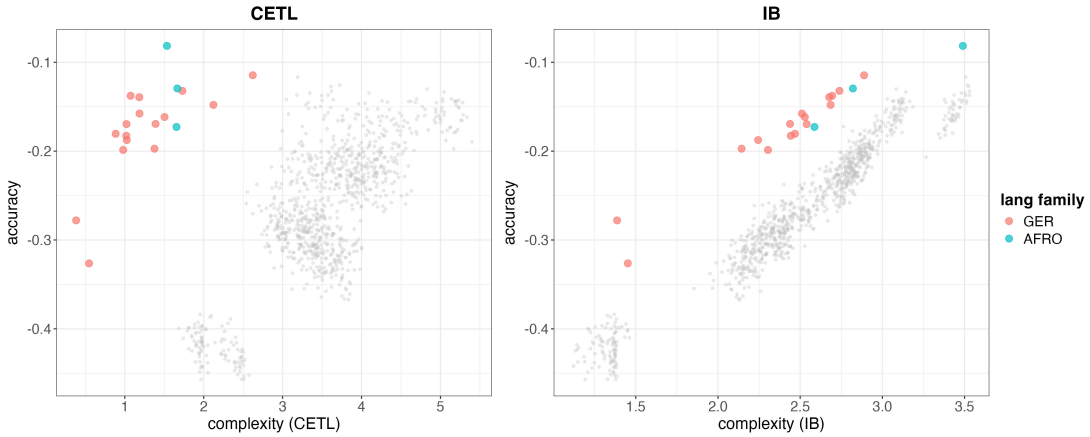
(a) Overall

(b) Structural Permutations

Figure 10: Results for Appendix B.3 - Performance of the CETL vs. the IB model using Zaslavsky et al. (2021b)'s feature representation: Amount of languages for which each model outperforms the other by at least 5% of the total permutations.

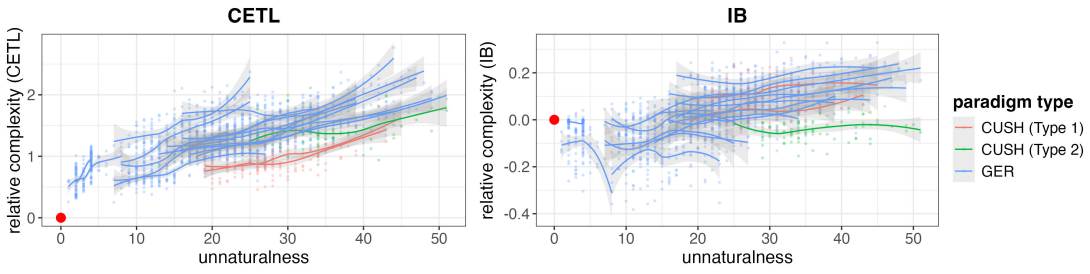


(a) Inside conjugation classes

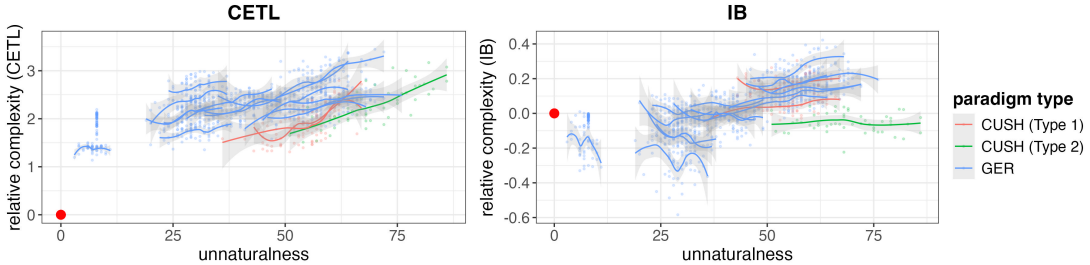


(b) Across conjugation classes

Figure 11: Results for Appendix B.4: Efficiency of permutations. Accuracy plotted against complexity measures for permutations across different conjugation classes (b) and inside permutation classes only (a). Results are shown for Germanic (red) and Cushitic (blue, a subfamily of Afro-Asiatic).

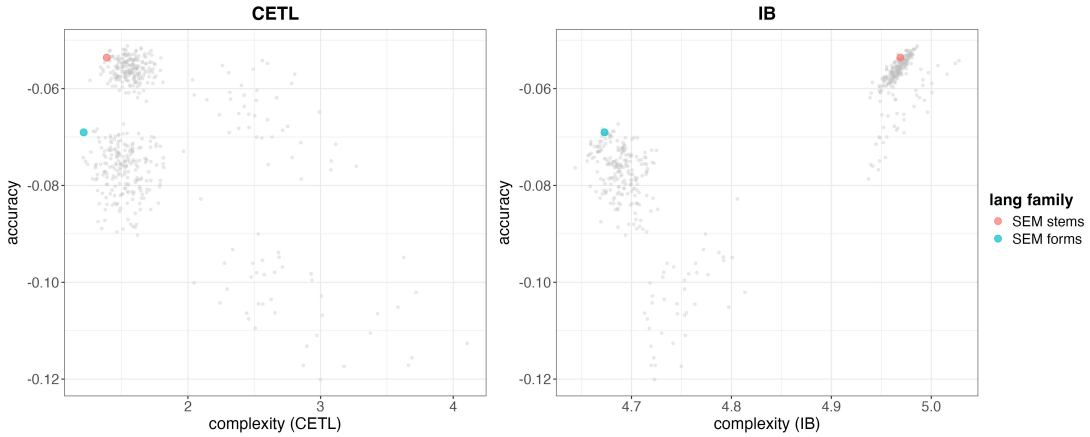


(a) Inside conjugation classes

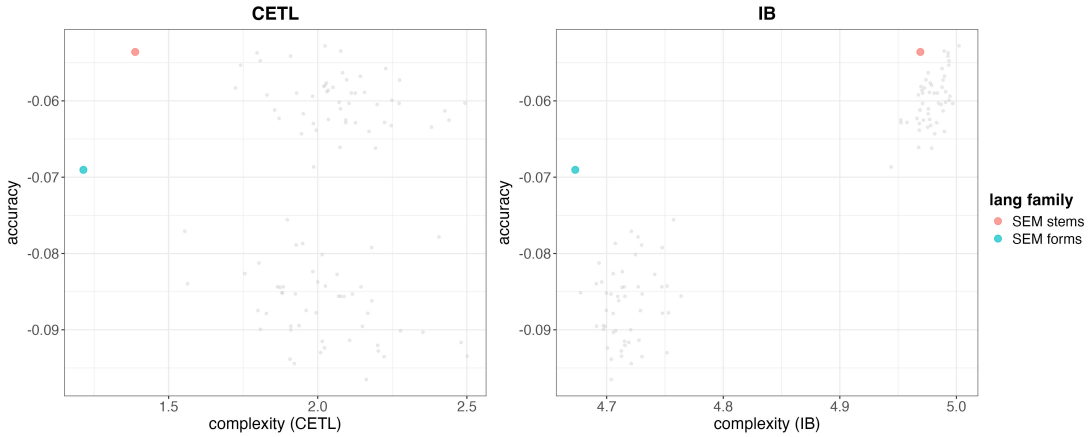


(b) Across conjugation classes

Figure 12: Results for Appendix B.4: Complexity of permutations of different naturalness: Complexity measures plotted against unnaturalness for permutations across different conjugation classes (b) and inside permutation classes only (a).

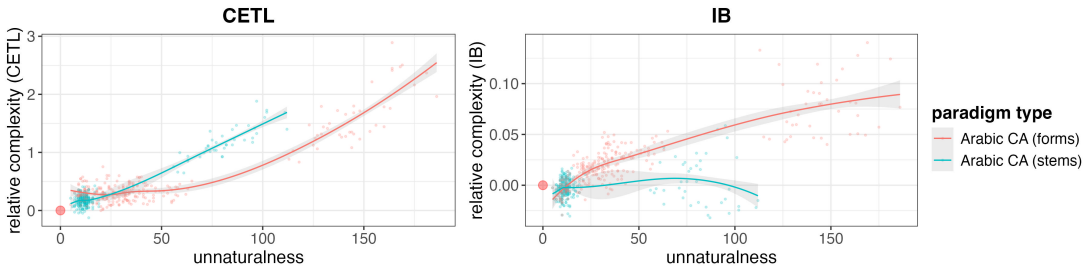


(a) Inside conjugation classes

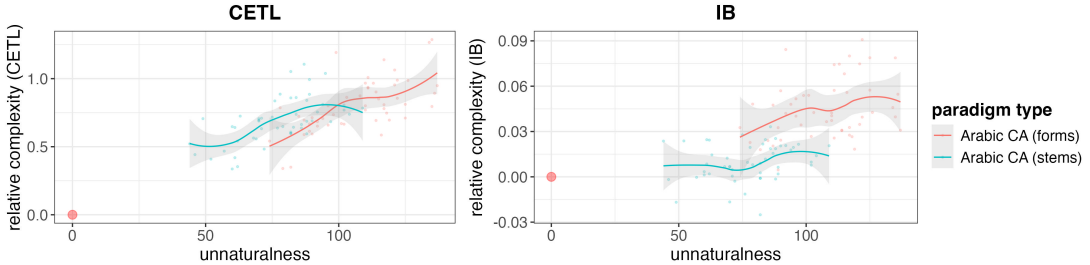


(b) Across conjugation classes

Figure 13: Results for Appendix B.4: Efficiency of permutations. Accuracy plotted against complexity measures for permutations across different conjugation classes (b) and inside permutation classes only (a). Results are shown for Arabic stems (red) and Arabic forms (blue).

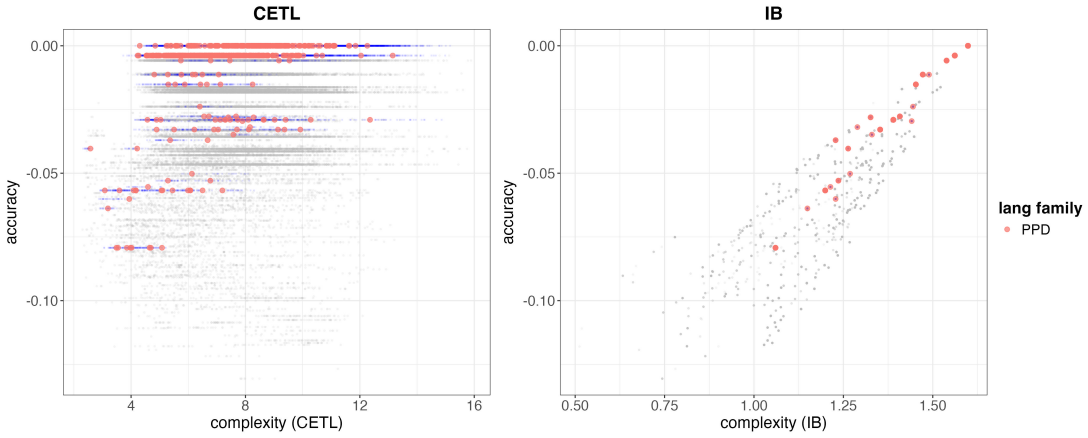


(a) Inside conjugation classes

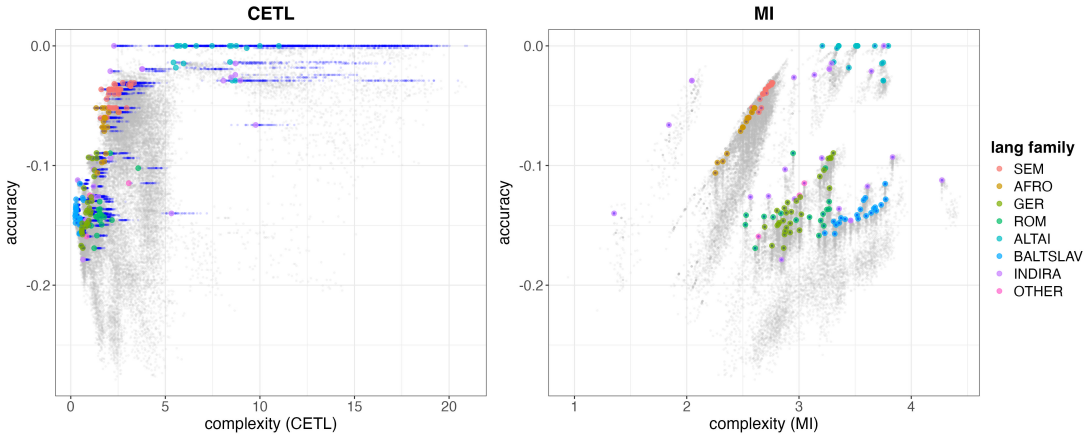


(b) Across conjugation classes

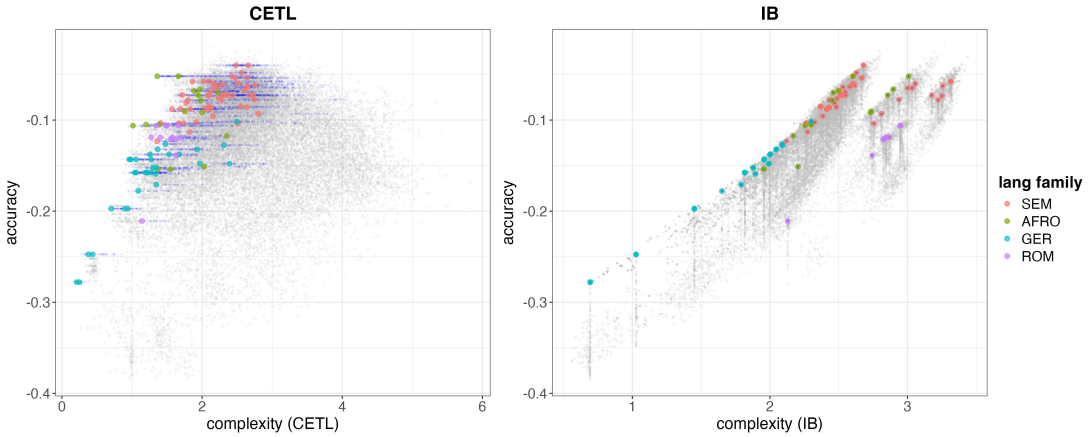
Figure 14: Results for Appendix B.4: Complexity of permutations of different naturalness: Complexity measures plotted against unnaturalness for permutations across different conjugation classes (b) and inside permutation classes only (a). Results are shown for Arabic stems (red) and Arabic forms (blue).



(a) PPD



(b) PRON



(c) VERB

Figure 15: Results for Appendix C.1 - Efficiency of permutations: Accuracy plotted against complexity measures for the CETL (left) and original IB model (right). Note that the results in the left column are also shown in Figure 4.

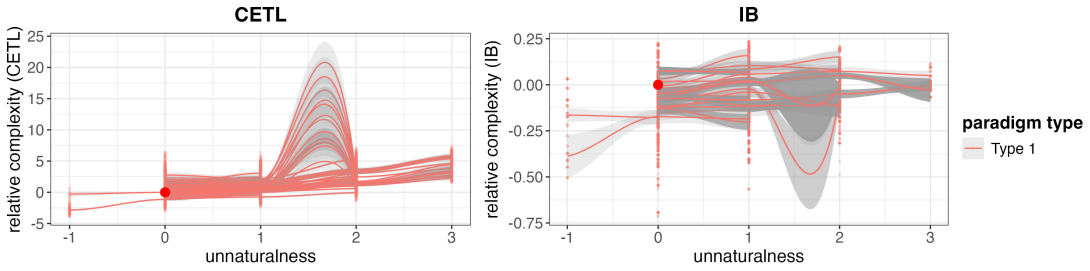


Figure 16: Results for Appendix C.2 - Complexity of Permutations of Different Naturalness on PPD: Complexity measures plotted against naturalness for the PPD data set in our CETL model (left) and the original IB model (right). Each line represents a separate language.

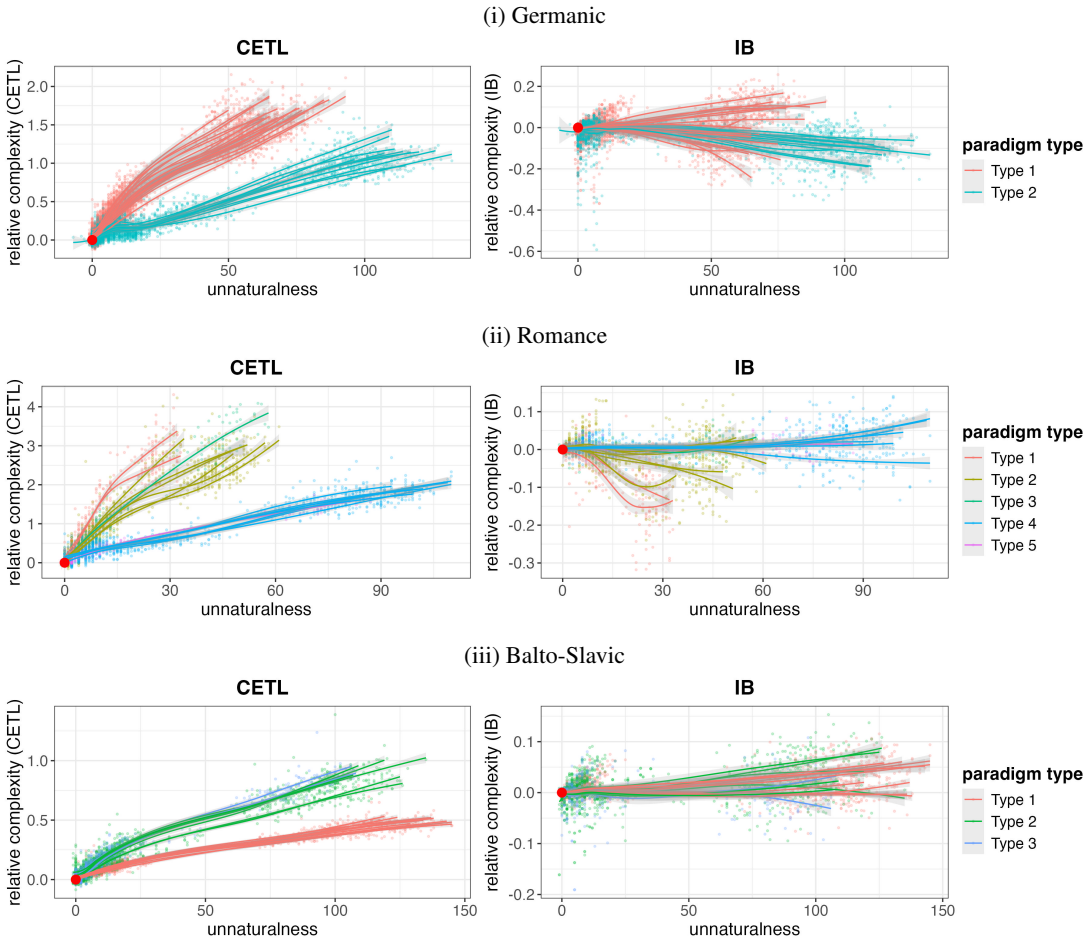


Figure 17: Results for Appendix C.2 - Complexity of Permutations of Different Naturalness on PRON (part I): Complexity measures plotted against naturalness for pronouns in our CETL model (left) and the original IB model (right). Each line represents a separate language. Each color represents a different paradigm type.

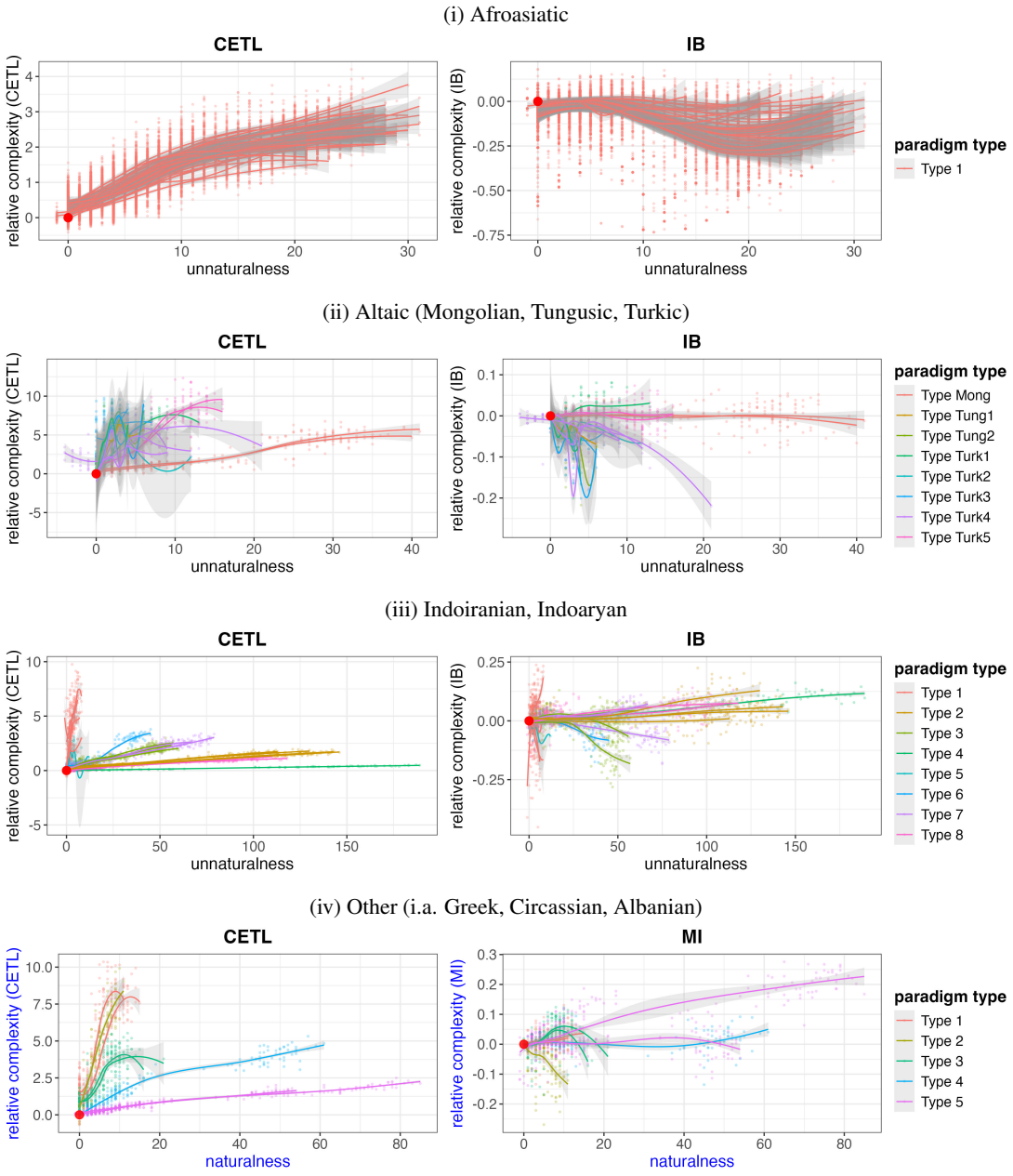
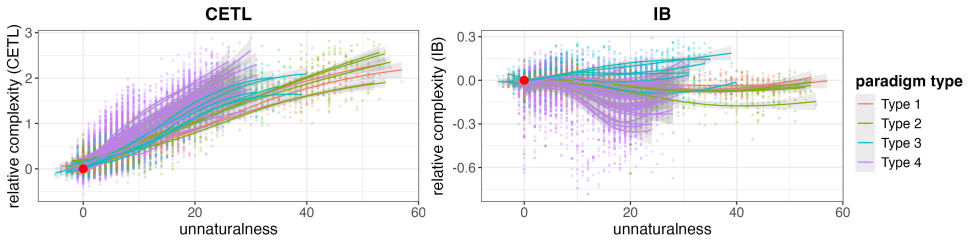
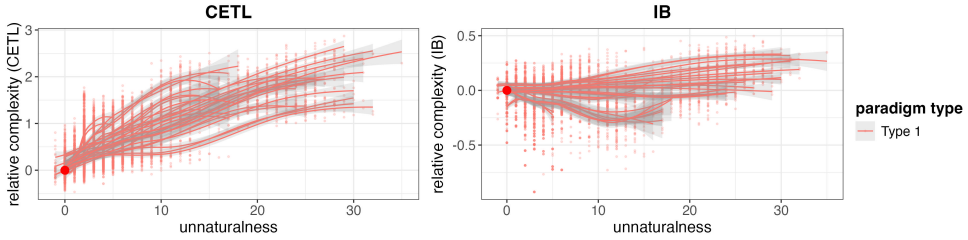


Figure 18: Results for Appendix C.2 - Complexity of Permutations of Different Naturalness on PRON (part II): Complexity measures plotted against unnaturalness for pronouns in our CETL model (left) and the original IB model (right). Each line represents a separate language. Each color represents a different paradigm type.

(i) Afroasiatic



(ii) Germanic



(iii) Romance

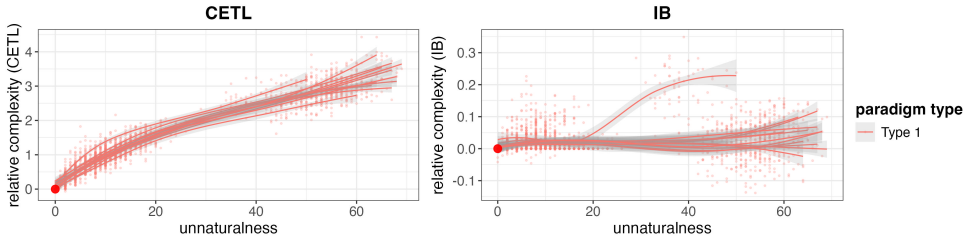
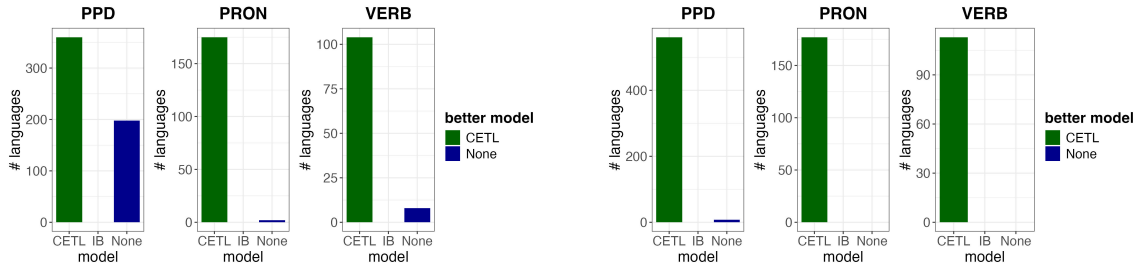


Figure 19: Results for Appendix C.2 - Complexity of Permutations of Different Naturalness on VERB: Complexity measures plotted against unnaturalness for verbs in our CETL model (left) and the original IB model (right). Each line represents a separate language. Each color represents a different paradigm type.



(a) Structural Permutations

(b) Form-Only Permutations

Figure 20: Results for Appendix C.4 - Performance of the CETL vs. the IB model using Zaslavsky et al. (2021b)'s feature representation: Amount of languages for which each model outperforms the other by at least 5% of the total permutations.

1s m G	16.1149	1p m G	3.168	1d m G	0.3769
1s f G	8.0371	1p f G	1.5789	1d f G	0.1935
2s m G	10.8485	2p m G	4.7571	2d m G	0.5704
2s f G	5.4090	2p f G	2.3734	2d f G	0.2852
3s m G	25.7411	3p m G	4.6043	3d m G	0.5501
3s f G	12.8247	3p f G	2.2919	3d f G	0.2750
1s m V	16.1149	1p m V	3.168	1d m V	0.3769
1s f V	8.0371	1p f V	1.5789	1d f V	0.1935
2s m V	10.8485	2p m V	4.7571	2d m V	0.5704
2s f V	5.4090	2p f V	2.3734	2d f V	0.2852
3s m V	25.7411	3p m V	4.6043	3d m V	0.5501
3s f V	12.8247	3p f V	2.2919	3d f V	0.2750

Table 7: Details for Appendix B.1 - Calculated Feature Probabilities. G stands for imperfective, V for perfective.

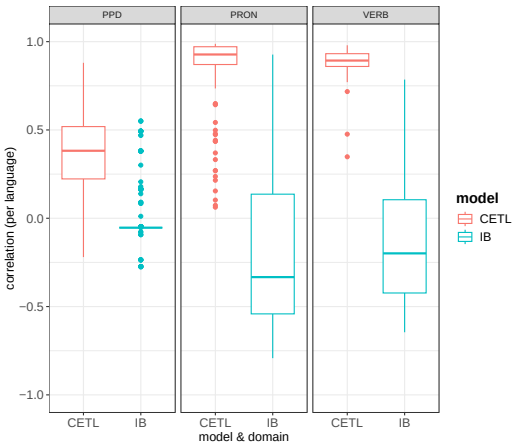


Figure 21: Results for Appendix C.3 - Distribution of correlation values for all languages in the CETL model (red) vs. the IB model (blue) across all language families and paradigm types in a domain.

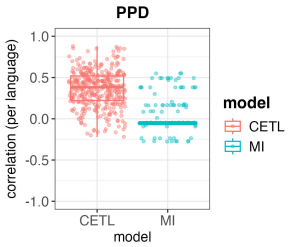


Figure 22: Results for Appendix C.3 - Distribution of correlation values for all languages in the CETL model (red) vs. the IB model (blue) in a PPD.

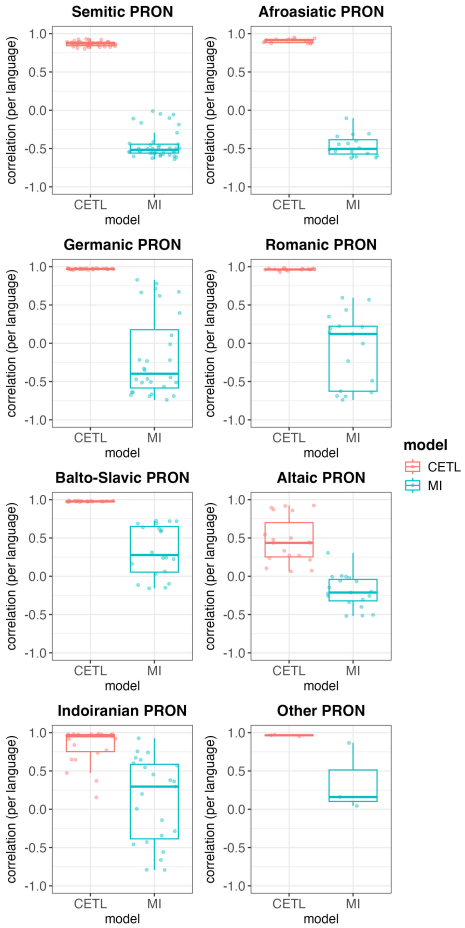


Figure 23: Results for Appendix C.3 - Distribution of correlation values for all languages in the CETL model (red) vs. the IB model (blue) in PRON, separated by language families.

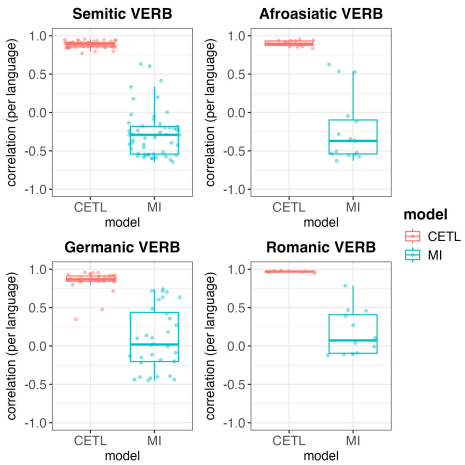


Figure 24: Results for Appendix C.3 - Distribution of correlation values for all languages in the CETL model (red) vs. the IB model (blue) in VERB, separated by language families.

	correlation (avg)		support
	CETL	IB	
uniform	0.78	-0.26	318
random1	0.76	0.13	318
random2	0.75	-0.05	318
random3	0.74	0.54	318

Table 8: Results for Appendix B.2: Correlation between complexity and unnaturalness for a uniform distribution and three random distributions.

	C_M (%)		I_M (%)		Perf_M		
Exp	CETL	IB	CETL	IB	CETL	IB	support
Unif	**82.50**	76.25	**6.56**	20.31	**75.94**	55.94	320
Rand1	**59.43**	49.06	15.09	**4.09**	44.34	**44.97**	318
Rand2	**68.77**	32.81	**1.89**	3.15	**66.88**	29.65	317
Rand3	**75.16**	47.77	2.23	**1.59**	**72.93**	46.18	314

(a) Structural Permutations

	C_M (%)		I_M (%)		Perf_M		
Exp	CETL	IB	CETL	IB	CETL	IB	support
Unif	**81.13**	0	**18.87**	100	**62.26**	-100	53
Rand1	**54.72**	0	**45.28**	100	**9.43**	-100	53
Rand2	**86.54**	0	**13.46**	100	**73.08**	-100	52
Rand3	**96.49**	0	**3.51**	100	**92.98**	-100	57

(b) Form-only Permutations

Table 9: Results for Appendix B.2: Hit and Fail rates for CETL vs. MI for structural permutations (left) and form-only permutations (right) for a uniform and three random distributions of feature frequencies.

	correlation (avg)		support
	CETL	IB	
CLASS	0.66	0.32	1620
XROSS	0.49	0.09	900

Table 10: Results for Appendix B.4 - Germanic & Cushitic: Correlation between complexity and unnaturalness for permutations inside the conjugation classes (CLASS) and across the conjugation classes (XROSS).

	correlation (avg)		support
	CETL	IB	
CLASS	0.93	0.51	500
XROSS	0.63	0.28	100

Table 11: Results for Appendix B.4 - Arabic stems and forms: Correlation between complexity and unnaturalness for permutations inside the conjugation classes (CLASS) and across the conjugation classes (XROSS).

		mean base	mean perms	t-value	p-value
PRON	cetl	7.38	7.98	-8.5632	**< 0.001**
	acc	0.0075	0.0134	-9.5378	**< 0.001**
PR	cetl	2.49	3.18	-3.6111	**< 0.001**
	acc	0.0854	0.1024	-4.0620	**< 0.001**
VERB	cetl	1.79	2.49	-12.5130	**< 0.001**
	acc	0.1092	0.1206	-2.4713	**0.015**

Table 12: Results for Appendix C.1 - Statistics for complexity and accuracy of real and counterfactual variants in the three domains.

	correlation (avg)		support
	CETL	IB	
PPD	0.36	-0.02	44112
PRON	0.82	-0.15	38623
VERB	0.88	-0.12	32825

Table 13: Results for Appendix C.3 - Correlation between complexity and unnaturalness, averaged over per-language correlations.

Family	type	description	languages
AfroSem	Type1	4 tenses (3PC, 1SC)	***Akkadian:*** Assyrian, Babylonian, Sargonic
	Type2	4 tenses (2PC, 2SC)	***Ethiopic:*** Amharic, Geez, Tigrinya, Muher
	Type3	3 tenses (2PC, 1SC)	Proto-Semitic (vers1 + vers2), ***South Arabian:*** Jibbali, Mehri, Soqotri, ***Ethiopic:*** Tigre, ***Berber:*** General Berber, ***Cushitic:*** Beja (PC + SC), ***Chadic:*** Hausa, ***Omotic:*** Yemsa,
	Type4	2 tenses (1PC, 1SC)	***Arabic:*** Classical Arabic, informal Modern Standard Arabic, Algerian Arabic, Andalusian Arabic, Dafur Arabic, Egyptian Arabic, Hejazi Arabic, Jewish Baghdadi Arabic, Muslim Iraqi Arabic, Beiruti Levantine Arabic, Rural Levantine Arabic, Jerusalemite Levantine Arabic, Moroccan Arabic, Maltese, Omani Arabic, Saidi Arabic, Tunesian Arabic, Yemenite Arabic, ***Canaanite:*** Biblical Hebrew, Modern Hebrew, Standard Phoenician, Punic Phoenician, ***Ugaritic:*** Ugaritic, ***Aramaic:*** Mlahso Central Neo-Aramaic, Surayt Central Neo-Aramaic, Turoyo Central Neo-Aramaic, Imperial Aramaic (vers1 + vers2) Jewish Babylonian, Jewish Palestinian, Classical Mandaic, Neo-Mandaic (vers1 + vers2) Alqosh North-Western Neo-Aramaic, Western Neo-Aramaic, Samaritan, Syriac (vers1 + vers2) ***Ancient South Arabian:*** Sabaic (vers1 + vers2) ***Ancient Egyptian:*** Earlier Ancient Egyptian, Middle Ancient Egyptian, Coptic, ***Berber:*** Ghadames Berber, Tamasheq Berber, Siwa Berber, ***Cushitic:*** AfarPC (PC + SC), SomaliPC (PC + SC), ***Chadic:*** Moloko,
GER	Type1		Old Dutch (strong + weak), Middle Dutch (strong + weak), Modern Dutch (strong + weak), Old English (strong + weak), Middle English (strong + weak), Modern English (strong + weak), Old High German (strong + weak), Middle High German (strong + weak), Standard High German (strong + weak), Old Norse (strong + weak), Faroese (strong + weak), Icelandic (strong + weak), Swedish (strong + weak), Gothic (strong + weak), Old Saxon (strong + weak), Proto-Germanic (strong + weak),
ROM	Type1		Spanish, Romanian, Occitan, Portuguese, Latin, Galician, Italian, Franco-Provençal, French, North Corsican, South Corsican, Catalan

Table 14: Details for Appendix A.2 - Languages belonging to each Paradigm Class for VERB. For some languages we used more than one paradigm indicated in brackets.

Family	type	distinguishing description	languages
AfroSem	Type1		*Arabic:* Moroccan Arabic, Classical Arabic, Egyptian Arabic, North Levantine Arabic, South Levantine Arabic, *Aramaic:* Babylonian Aramaic, North-East Neo-Aramaic, Baxa Western Neo-Aramaic, Jubbadin Western Neo-Aramaic, Malula Western Neo-Aramaic, Syriac (vers1 + vers2), Samaritan, Mandaic, *Canaanite:* Modern Hebrew, Biblical Hebrew (vers1 + vers2), Phoenician, Punic, *Ethiopic:* Amharic (vers1 + vers2), Geez (vers1 + vers2), Tigre, Tigrinya (vers1 + vers2 + vers3 + vers4), Muher, Dahalik, Soddo, Argobba, *Akkadian:* Akkadian, *South Arabian:* Mehri, Soqotri (vers1 + vers2), *Ancient Egyptian:* Coptic, Middle Ancient Egyptian, *Berber:* Ghadames Berber, Tashelhiyt Berber, *Cushitic:* Alaaba, Benijamer Beja (vers1 + vers2), Agaw Bilin, Burunge, Somali, Tsamakko, *Chadic:* Hdi, Mubi, Hausa, *Omotic:* Yemsa, Aari, *Ugaritic:* Ugaritic, *Proto-Semitic:* ProtoSemitic (vers1 + vers2), *Ancient South Arabian:* Razihi,
GER	Type1	no formality	Middle Dutch, Old Dutch, English, Middle English, Early Middle English, Old English, Faroese, Old Frisian, Middle High German, Low German, Middle Low German, Gothic, Icelandic, Old Norse, Norwegian-Bokmal, Norwegian-Nynorsk, Proto-Germanic, Old Prussian, Old Saxon
	Type2	formality	Swedish, Old High German, Standard High German, Bavarian (vers1 + vers2), Short Dutch, Archaic English, Dutch, Afrikaans
ROM	Type1	3 cases, reflexive	Rumantsch, Dalmatian
	Type2	5 cases, reflexive	French, Corsican, Emilian, Franco-Provençal, Occitan, Sicilian
	Type3	6 cases	Romanian
	Type4	6 cases, formality, reflexive	Castillian Spanish, Latin American Spanish,Catalan, Archaic Spanish, Portuguese, Italian, Galician
	Type5	6 cases, neutral gender, reflexive	Latin
BALTSLAV	Type1	6 cases + short vers, animation	Czech, Kashubian (vers1 + vers2), Polish, Slovak, Serbo-Croatian (vers1 + vers2), Slovene (vers1 + vers2), Lower Sorbian, Upper Sorbian
	Type2	6 cases	Belorussian, Latvian, Lithuanian, Old Church Slavonic, Proto-Slavic, Russian, Ukrainian
	Type3	4 cases + short vers	Bulgarian, Macedonian
ALTAI	Type Mong	incl 1st, formality, 9 cases	Mongol (vers1 + vers2),
	Type Tung1	incl 1st, 9 cases	Udihe
	Type Tung2	incl 1st, 5 cases	Manchu
	Type Turk1	7 cases	Chuvash, Southern Altai, Tuvan
	Type Turk2	7 cases, formality	Kazakh
	Type Turk3	6 cases	Turkmen, Tartar, Crimean, Bashkir, Azeri
	Type Turk4	6 cases, formality	Uzbek, Uyghur, Turkish, Kyrgyz
	Type Turk5	11 cases	Old Turkic, Proto-Turkic
INDOIRAN	Type1		Ossetian, Sorani Kurdish, Hewleri Kurdish, Kurmanji Kurdish
	Type2		Urdu (vers1 + vers2), Punjabi, Kashmiri
	Type3		Iranian Persian / Farsi, Afghani Persian / Dari, Pashto
	Type4		Assamese
	Type5		Bengali
	Type6		Gujurati
	Type7		Sindhi, Gilaki
	Type8		Proto-Indo-European, Sanskrit
OTHER	Type1	Armenian	Eastern Armenian, Classical Armenian
	Type2	Georgian	Georgian
	Type3	Circassian	Kabardian Circassian, Adyghe Circassian
	Type4	Albanian	Albanian
	Type5	Greek	Modern Greek, Ancient Greek

Table 15: Details for Appendix A.2 - Languages belonging to each Paradigm Class for PRON. For some languages we used more than one paradigm indicated in brackets.