

## Laissez-Faire Analogical Change

I explore the role of *Paradigm Saturation* (PS) [4] and *Inverse Paradigm Saturation* (IPS), both *type*-based frequency measures, in accounting for observed tendencies in analogical change and the organization of morphological paradigms: **1)** The observation that irregularly inflected stems tend to be among the most common in the language, **2)** that certain inflectional categories (e.g., 3rd person singular vs. 2nd person plural or present indicative vs. imperfect subjunctive) tend to be more resistant to analogical leveling and so are more likely to remain unique or irregular, and **3)** that there is an apparent complexity trade-off between the size of morphological paradigms and the level of irregularity among their forms. This approach is contrasted with *functional* or *usage-based* approaches which argue that a conserving effect of token frequency and trade-offs between the cognitive or communicative burdens incurred by learning and using morphological paradigms result in these tendencies [e.g., 2, 3, 5, 1]. Rather, this account is *laissez-faire* in that it frames these tendencies as the diachronic result of learner interaction with their input data without reference to communicative or functional optimization or goals. PS and IPS are consistent and empirically observable facts about the distribution of forms in corpora including child-directed speech and so constitute a baseline frequency effect in acquisition and change.

**Paradigm Saturation and Irregular Stems:** *Paradigm Saturation* (PS) measures the number of potential inflectional forms which are attested for a given lemma in a given corpus [4]. If most of a verb's paradigm is attested, then it has high saturation, and it has low saturation if only a few inflections are attested. PS consistently follows long-tailed Zipfian distributions (Fig. 1) such that a few verbs have relatively high PS and most have low PS in languages with even moderately large paradigms. This is a mathematical fact of corpora regardless of size or genre. Eg, in Spanish calculated over the Universal Dependency Treebank (UDT; [1]), the highest PS non-auxiliary verb is only at 43.28% while the mean is 4.91%, or just over one form attested per verb. In a similarly sized collection of child-directed speech from CHILDES [7], max and mean PS are only slightly higher at 44.83% and 7.59% respectively. PS has clear implications for learning [4, 6]: as relevant here, if a form is not attested in the input, it must be inferred by some kind of productive or regular pattern, and if a form is really irregular, it must be memorized. But to be memorized, it needs to appear in the input. Since most stems have very low PS, most of their forms must be inferred and so cannot be irregular. Only those stems with high PS are likely to remain irregular over time as successive generations of learners memorize them. PS correlates well with *token* frequency (Fig. 2), thus accounting for the relationship between frequency and irregularity. Furthermore, PS rank is colinear with token frequency rank, so correlations between irregularity and token frequency are not unique evidence for that position. The contribution of PS would have to be factored out of a model before token frequency effects can be taken into account.

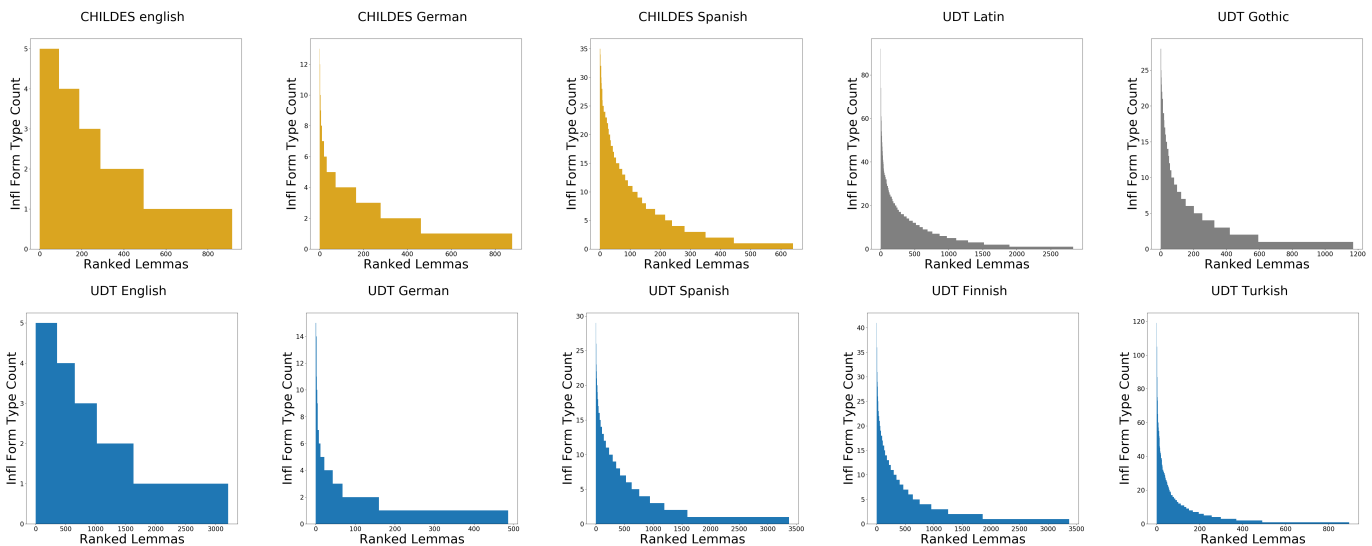


Figure 1: Zipfian CHILDES (gold; English, German, Spanish), modern UDT (blue; English, German, Spanish, Finnish, Turkish), historical UDT (grey; Latin, Gothic) verb paradigm saturation distrs.

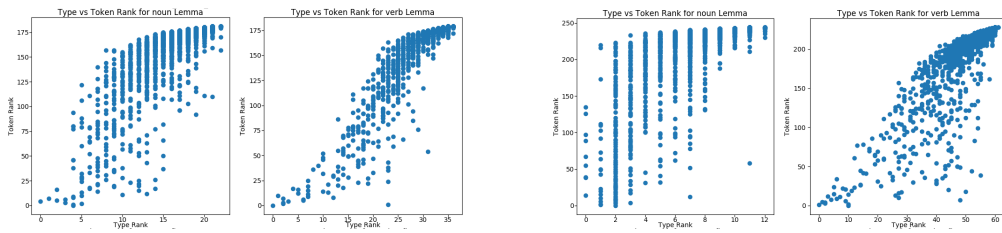


Figure 2: UDT Finnish noun and verb PS rank vs token freq rank (RIGHT) UDT Latin ” ” ” ”

**Inverse Paradigm Saturation and Irregular Categories:** *Inverse Paradigm Saturation* (IPS) measures the number of lemmas inflected into a given category in a corpus out of all those that could potentially support it, for example, how many verbs in a language with person/number marking are inflected as the 3rd person singular out of all attested verbs. IPS also follows long-tailed distributions such that only a few inflectional categories have high type-frequency and the vast majority have very low type-frequency. Eg, for Spanish, more than half of verbs are attested in the 3sg pres. indic., while in a million words of Spanish CHILDES, the 2pl impf. subjunctive is only attested with two verbs. Since 3sg and 3pl tend to have much higher IPS than other person/number combinations, and present and simple past indicative (or the equivalents) tend to have much higher IPS than other tense/mood combinations, many more inflected forms are attested to the learner, so fewer need to be inferred. They are more likely to be the basis for analogical extension than subject to analogical leveling. As a historical example, in Proto-Germanic, the *\*ē*-grade extended analogically from Class V strong verb past default stems to Class IV past stems, but the past.3sg and pptc stems did not analogize. Using Gothic as a proxy, it turns out that past.3sg and the pptc have the by far the highest IPS (Fig. 3), so analogy should have been less likely to occur because more young learners would have heard the “correct” past.3sg and pptc forms directly but would have had to infer the past default. IPS is extremely colinear with token frequency ranks of inflectional categories, so once again, simply finding correlations between token frequency and resistance to analogical leveling cannot be interpreted as evidence for a causal role for token frequency (Fig. 4).

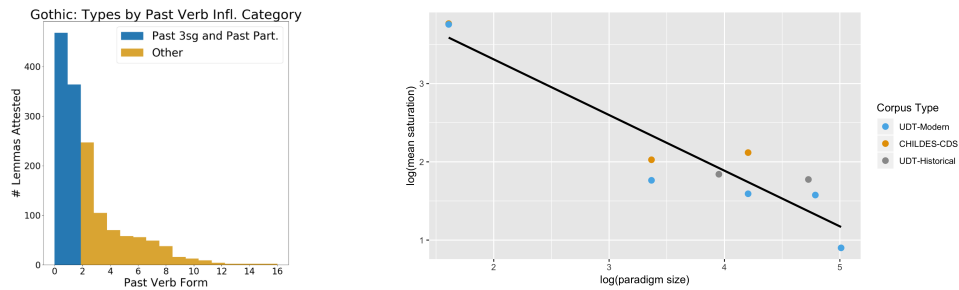


Figure 3: (LEFT) UDT Gothic IPS. (RIGHT) Paradigm size vs. mean PS on a log-log scale for Fig. 1 languages

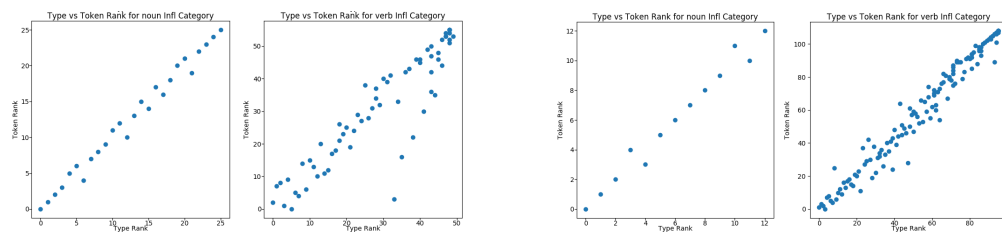


Figure 4: (LEFT) UDT Finnish noun and verb category IPS rank vs token freq rank (RIGHT) UDT Latin ” ” ” ”

**Type-based Corpus Attestation and Trade-Offs in Complexity:** Paradigm and inverse paradigm saturation conspire to yield an apparent trade-off between paradigm size and overall irregularity (*E-complexity* and *I-complexity* following [1]). Essentially, the larger that a language’s paradigm is, the less chance there is for any given verb to be inflected into any given category since each category has a more specific meaning. Average PS (Fig. 3) and IPS are systematically lower for languages with larger paradigms which means that more inflected forms must be inferred by the learner, that is, they are created by productive rule and so are not irregular. This indicates that the input itself can largely account for the observed trade-off without assumptions about the ease or difficulty of acquiring or processing large and irregular paradigms.

[1] F. Ackerman & R. Malouf. *Language*, 2013. [2] J. Bybee, 1985. [3] J. Bybee & S. Thompson. *Annual Meeting of the Berkeley Linguistics Society*, 1997. [4] E. Chan, PhD thesis, 2008. [5] Ö. Dahl, 2004. [6] C. Lignos & C. Yang. *Cambridge handbook of morphology*, 2018. [7] B. MacWhinney, 2000. [8] J. Nivre et al., 2018.