Are syntactic categories ISL-2 inferrable? A corpus study

Logan Swanson and **Kenneth Hanson** and **Thomas Graf**
Stony Brook University
logan.swanson@stonybrook.edu
mail@kennethhanson.net
mail@thomasgraf.net

Abstract

We use the MG treebank of Torr (2017) to investigate the conjecture in Graf (2020) that category systems are ISL-2 inferrable. A category system is ISL-2 inferrable iff the category feature of every lexical item can be jointly inferred from phonological exponents of both the item itself and either its selecting head or the arguments it selects. If correct, this conjecture would greatly limit the overgeneration problem posed by subcategorization mechanisms (Kobele, 2011; Graf, 2011, 2017). We find that the conjecture is largely borne out in this data set. However, we also observe that it holds even for features that are not expected to be inferrable in this manner, and we demonstrate that inferrability can arise from the assumption that certain distributional properties of the lexicon are Zipfian in nature. We conclude that category systems in natural languages may well be ISL-2 inferrable, but that this could be due to extragrammatical factors.

1 Introduction

A good model of language should be sufficiently expressive to account for observed linguistic variation while still being restrictive enough to rule out highly unnatural patterns. Graf (2017) highlights a major overgeneration problem with syntactic *subcategorization* mechanisms. Subcategorization is needed to capture basic facts such as *devour* being a verb that takes a DP subject and a DP object. But without meaningful restrictions on the inventory of syntactic categories, subcategorization can be used to enforce *any* constraint definable in monadic second-order logic (MSO).

MSO has been used extensively in model-theoretic syntax (see Rogers 1998, Rogers 2003, Morawietz 2003, Tiede and Kepser 2009, Graf 2013, and references therein) due to its ability to succinctly capture even the most byzantine proposals from the syntactic literature. However, it can also express highly unnatural constraints such as "a reflexive must c-command a verb of motion unless there are at least three CP nodes in the same tree that each properly dominate an odd number of nodes". Extending a well-known translation mechanism from MSO constraints to bottom-up tree automata (Thatcher and Wright, 1968; Doner, 1970), the states of these automata can be compiled into a grammar's category system to implicitly enforce MSO constraints via subcategorization (Graf, 2011; Kobele, 2011). Graf (2017) argues that linguists' restrictions on category systems are not tight enough to rein in subcategorization, and as a result current theories of syntax are much less restrictive than they appear.

Graf (2020) shows that many undesirable kinds of overgeneration, e.g. modulo counting, can be ruled out if category features are required to be inferrable by *input strictly 2-local* (ISL-2) functions. Intuitively, the category feature of a lexical item l is *ISL-2 inferrable* iff it can be predicted from the phonological content of l itself and its local tree context. Graf (2020) conjectures that all natural languages have category systems that are ISL-2 inferrable. If true, this would explain how subcategorization can be ubiquitous in syntax without giving rise to unnatural MSO constraints.

In order to assess the viability of ISL-2 inferrability as a linguistic universal, we test whether it holds for MGBank (Torr, 2017), a treebank of English sentences with structures very similar to those assumed by Graf (2020). We find that the category features for a large majority of lexical items can indeed be predicted from strictly local tree contexts. When a category feature is not ISL-2 inferrable, that is usually due to empty heads, i.e. lexical items that lack phonological exponents and hence provide no information for ISL-2 inferrability (an edge case already mentioned in Graf 2020). However, we also find a similarly high degree of inferrability for movement features, which operate over long

distances and would not be expected to be ISL inferrable by this conjecture. Probing further, we show that ISL-2 inferrability can arise in language datasets following Zipfian frequency distributions. This makes it difficult to assess whether ISL-2 inferrability is a guiding principle of the grammar, as conjectured by Graf (2017), or rather an artifact of other features of human language.

This paper is organized as follows. Section 2 introduces the necessary background on the Minimalist grammar formalism (Sec. 2.1), the overgeneration problem (Sec. 2.2), and ISL-2 inferrability (Sec. 2.3). Section 3 describes the data and methodology used. Section 4 displays our findings on the ISL-inferrability of category-system features and discusses how they may support the ISL-inferrability hypothesis. Section 5 complicates this picture by introducing theoretical limitations of ISL-2 inferrability and also demonstrates how a high degree of inferrability can arise naturally from other properties of language. Section 6 offers ideas for future research directions and concludes.

2 Background: ISL inferrability

2.1 Categories in Minimalist grammars

Following Graf (2017), the results of this paper are couched in the formal terms of *Minimalist Grammars* (MG) (Stabler, 1997, 2011) and *suregular syntax* (Graf, 2022b,a). However, the results of this study are not limited in relevance to just those formalisms and bear on syntax much more generally. ISL-2 inferrability asks whether certain kinds of information can be inferred from local tree contexts, and in MG trees the local relationships are those between heads and their arguments (specifiers and complements). The central question that Graf (2017) formally hashes out as ISL-2 inferrability over MG trees thus is much broader and extends far beyond MGs to other formalisms: to what extent can specific features of a lexical item be inferred from the phonological content of its arguments and/or its selecting head?

In MGs, every lexical item consists of a *phonological exponent* that determines its pronunciation, and a string of features that determine its syntactic behavior. The feature string always contains a *category feature* (x) and may contain *selector features* (=x) that encode the item's subcategorization requirements. For example, a word like *say* would have the feature string ⟨=c =d v⟩, representing that it selects a CP complement, a DP specifier, and is a verb.

The MG feature strings may also include movement features. The *licensee feature* -m indicates that the item is a mover of type m, while a *licensor feature* +m indicates that this item furnishes a landing site that must be filled by an m-mover. Graf (2020) explicitly states that movement features are not expected to be ISL-2 inferrable. This effectively makes inferrability of movement features a "control group" for our corpus experiment, a point we will return to in Section 5.2. Until then, we omit movement features from the discussion and all examples.

MGs furnish multiple types of structural descriptions: phrase structure trees, derivation trees, and dependency trees. While a lot of early MG work focused on phrase structure trees, Kobele et al. (2007) started a shift toward derivation trees as the primary syntactic representation of MGs. Derivation trees are also used in the MG treebank (Torr, 2017) that our corpus analysis is based on. Subregular syntax, including Graf (2017), prefers dependency trees instead. But since there is a one-to-one correspondence between derivation trees and dependency trees, the choice is purely a matter of mathematical convenience and it is easy to translate between the two.[1] Graf (2017) uses dependency trees because of their close connection to head-argument relations: the mother-of relation in MG dependency trees encodes subcategorization. Every node is a (feature-annotated) lexical item, and its i-th daughter from the right is its i-th argument — the rightmost daughter is the complement, all other daughters are specifiers. Even though MGs use movement, no displacement takes place in dependency trees. Every lexical item sits in the position where it is selected, and movement is encoded purely via movement features. An example tree for the sentence *The child laughed at a bear* is given in Fig. 1.

2.2 The overgeneration problem

Although subcategorization is crucial for modeling the kinds of patterns found in syntax, it introduces

[1] As pointed out in Graf (2011, 2012), MG derivation trees are built from chunks of derivational structure called *slices*. Intuitively, the slice $\text{slice}(l)$ consists of the operations that assemble the projections of lexical item l in the phrase structure tree. A given MG derivation tree t is converted to an equivalent MG dependency tree by replacing $\text{slice}(l)$ with l for every lexical item l of t. For example, if $\text{slice}(l) = \text{Move}(\text{Merge}(x, \text{Merge}(l, y)))$, this is condensed to $l(x, y)$. One could also say that MG dependency trees are the derivation trees of a Tree Substitution Grammar that generates MG derivation trees.

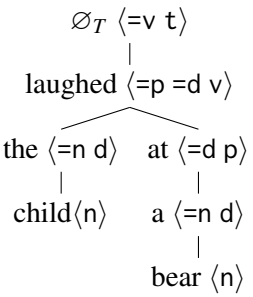

Figure 1: MG dependency tree for *The child laughed at a bear*, with empty T-head above the verb

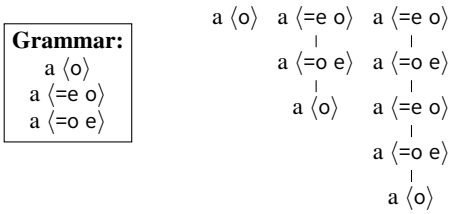

Figure 2: Smuggling in an unnatural *modulo* counting constraint via the category system. Left: Grammar which tracks o[dd] and e[ven] nodes, Right: Some trees generated by this grammar.

massive overgeneration into the formalism. As mentioned in the introduction, Graf (2011) and Kobele (2011) show that a constraint can be enforced via MG-style subcategorization iff it is definable in MSO. Figure 2 gives an example where the category system is used to track whether a subtree contains an odd (o) or an even (e) number of nodes. Graf (2017) illustrates the many ways MSO-constraints and, by extension, subcategorization undermine the restrictiveness of syntactic formalisms. A restrictive theory of syntax thus requires tight restrictions on its category system.

2.3 ISL-2 inferrability to the rescue

Graf (2020) proposes to curb the excessive power of subcategorization by requiring category features to be inferrable by input strictly 2-local (ISL-2) tree-to-tree transductions. While the definition of ISL-2 transductions in Graf (2020) is fairly technical, the general idea is simple enough (see Fig. 3 for a visualization).

Suppose we take a dependency tree t generated by some MG G and remove all feature strings from all nodes, leaving only the exponents. Is there a function f_G that correctly determines for each node n of t whether n had feature f? If the answer is positive for every node of every dependency tree of G, then f is *inferrable* for G. If f_G can do this based

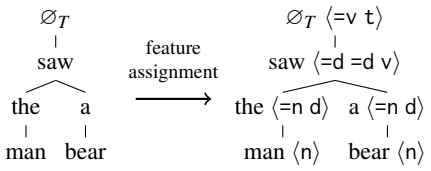

Figure 3: Feature assignment transduction

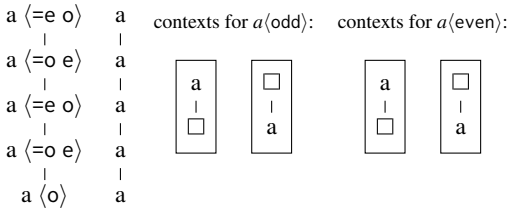

Figure 4: Category system implementing modulo counting is not ISL inferrable

solely on I) the exponent of n and II) the exponents of either IIA) n's mother and siblings (*upper context*) or IIB) the exponents of n's daughters (*lower context*), then f is *ISL-2 inferrable* for G.

Many unnatural category systems, like the *modulo* counting example in Fig. 2, are not ISL-2 inferrable. Figure 4 shows that the category features o and e are not ISL-2 inferrable because they share at least one structural context of size 2 (in fact, their contexts are exactly the same). Meanwhile, many natural patterns which require subcategorization *are* ISL-2 inferrable: Fig. 5 demonstrates how local contexts can successfully disambiguate two lexical entries for *have*. In light of this, Graf (2020) conjectures that ISL-2 inferrability (or at least ISL-k inferrability for some fixed $k \geq 2$) is a linguistic universal of category systems. Next, we will evaluate this conjecture with our corpus study.

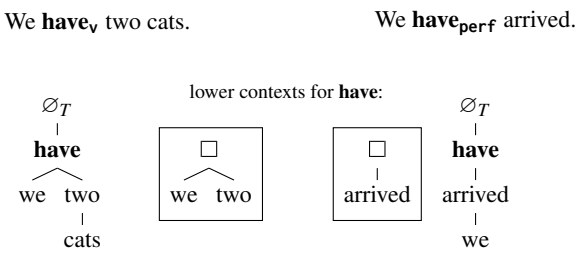

Figure 5: Example of disambiguating contexts for two lexical entries of *have*

3 Methods

3.1 Corpus: MGBank

To investigate the viability of ISL-2 inferrability as a linguistic universal, we conducted a study using data from MGBank (Torr, 2017), a database of MG derivation trees. The data in MGBank was created by automatically translating a portion of the Penn Treebank (Marcus et al., 1993) followed by a manual check for correctness. Overall, MGBank consists of 49,000 Wall Street Journal sentences, adding up to over 1 million words. While the derivation tree format in MGBank is different from the dependency tree format used here, there is a deterministic, sound, and complete translation from the former to the latter (see fn. 1). These qualities make MGBank an ideal data set for testing the conjecture that syntactic categories are ISL-2 inferrable.

3.2 Determining ISL-2 inferrability

Data from MGBank was first converted from MG derivation trees into dependency trees.[2] The MGBank annotation scheme includes some details which are not relevant to our research question and were therefore removed as part of the translation. For example, information on whether an argument should be linearized to the left or the right of the head was removed. Additionally, adjunction was converted to category-preserving selection with empty heads (the consequences of adjunction are discussed in Section 5.1). Next, a lexicon was extracted, consisting of all attested pairs of exponents and feature strings.

We then examined inferrability for category features in isolation as well as category features together with selector features. From a linguistic perspective, the category features are more important, since once these are determined, the selector features follow trivially. As a control, we also test inferrability for movement features.

In many cases, the relevant features (category / category + selector / movement) are predictable directly from the exponent itself. This means that they are ISL-1 inferrable and hence ISL-2 inferrable. For example, the category of *destruction* is always `n` irrespective of its local context. ISL-1 inferrability of feature `f` can fail only if the corpus contains two lexical items l and l' such that both have the same exponent but only of them carries `f`.

[2]Complete code for this project can be found at www.github.com/pterodactylogan/isl-k-corpus-test

$\varnothing_T$	**saw**	**the**	**man**	**a**	**bear**
$\rtimes$ / □	$\varnothing_T$ / □	saw / □ a	the / □	saw / the □	a / □
□ / saw	□ / the a	□ / man	□ / $\ltimes$	□ / bear	□ / $\ltimes$

Figure 6: Upper and lower (size 2) contexts for each lexical item in the sentence *The man saw a bear*.

Lexical Item	Contexts	Unique	Shared
a ⟨fspec1⟩	{ c1,c2,c3 }	{ c1,c2,c3 }	{ }
a ⟨fspec2⟩	{ c4,c5,c6 }	{ c4,c5,c6 }	{ }
b ⟨fspec1⟩	{ c1,c2,c3 }	{ c1 }	{ c2,c3 }
b ⟨fspec2⟩	{ c2,c3,c4 }	{ c4 }	{ c2,c3 }
c ⟨fspec1⟩	{ c1,c2,c3 }	{ }	{ c1,c2,c3 }
c ⟨fspec2⟩	{ c1,c2,c3 }	{ }	{ c1,c2,c3 }

Figure 7: Computing shared and unique contexts for each lexical item. The features for items with exponent *a* are strongly (and also weakly) inferrable, those for items with exponent *b* are weakly (but *not* strongly) inferrable, and those for items with exponent *c* are neither.

But `f` can still be ISL-2 inferrable if l and l' have distinct structural contexts.

Given a node n in tree t, its *upper context* consists of n itself, its parent, and any siblings of n, while the *lower context* consists of n itself and its children. Crucially, our contexts track only exponents, with all features omitted. Following Graf (2020), we modified each tree by inserting $\rtimes$ above the root and $\ltimes$ below each leaf so that every lexical item has an upper and a lower context in every tree. Figure 6 gives an example for the upper and lower contexts for each element in the example from Fig. 3.

The following method is used to assess ISL-2 inferrability of a given feature (or string of features) `f`: First, the set of all lexical items is extracted from the corpus together with the upper and lower contexts for each lexical item. This then allows us to assess two types of ISL-inferrability in terms of context sets. For each exponent e, let E be the set of lexical items that share the same exponent. We say that `f` is *strongly inferrable* iff it holds for every exponent e that no $l \in E$ carrying `f` ever appears in the same (upper or lower) context as some $l' \in E$ without `f`. We also say that l and l' have no *shared* contexts. When the contexts are restricted to upper and lower contexts as defined above, `f` being strongly inferrable is equivalent to it being ISL-2 inferrable. We say that `f` is *weakly inferrable* iff it

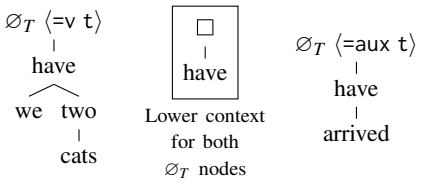

Figure 8: Inferrability is difficult with empty heads. Here, the lower context is insufficient to discriminate between the T head which selects a v complement (left) and the one which selects a aux complement (right).

holds for every exponent e and every $l \in E$ carrying f that l occurs in some (upper or lower) context that no $l' \in E$ without f occurs in. We also say that l has a *unique* context. While weak inferrability does not imply ISL-2 inferrability, it was included in this study because it might be a useful property for distributional learning algorithms. Weak and strong inferrability of each feature were then computed for each lexical item using these contexts. Figure 7 illustrates this process.

3.3 The trouble with empty heads

A possible stumbling block for ISL-inferrability comes from empty heads, which have no pronounced exponent. Empty heads introduce a lot of ambiguity, particularly when many of them are stacked together, e.g. in the functional hierarchy C-T-*v*-V commonly assumed in Minimalism. Figure 8 illustrates this issue with an empty T-head.

At the same time, these heads may actually carry prosodic information (e.g. a C-head that furnishes a wh-landing site) or contribute information that is pronounced on other heads, like tense. Arguably, this information should be taken into account for ISL-2 inferrability. In the following section, we report results with this information (empty heads have exponents such as [PAST] or [PRESENT]) and without (empty heads have the empty string as their exponent).

4 Results

4.1 Strong support for ISL-2 inferrability

We now report our findings on the inferrability of feature strings in MGBank. The full corpus contains nearly 40,000 distinct lexical items, with each lexical item including an exponent, a category feature, and zero or more selector and movement features. As mentioned above, we examined various subsets of features, and tested inferrability both with and without disambiguation of empty heads. Both of these variables affect the total number of distinct items, which we report along with results on inferrability.

For each of the feature subsets discussed in Section 3.2, the total number of *ambiguous* items was computed, that is, those that are *not* inferrable. This was done based on the criterion for ISL-1 inferrability as well as both strong and weak ISL-2 inferrability. The level of ISL-1 inferrability reflects the amount of *lexical ambiguity* in the corpus. The percentages for both weak and strong ISL-2 inferrability therefore indicate the percentage of lexically ambiguous items (rather than all items) which cannot be disambiguated using a context of size 2. Because the number of lexically ambiguous items may be much smaller than the total, taking the latter as a baseline could create a skewed view of how much work the local structural context does to disambiguate category information.

Table 1 shows the inferrability for category features and category + selector features depending on whether empty heads have as their exponent the empty string or linguistic annotations like [PAST]. These results demonstrate that ISL-inferrability holds for the vast majority of lexical items (*modulo* movement features). In fact, nearly two-thirds of category features and nearly half of category + selector feature pairs are ISL-1 inferrable. Category features alone have much less ISL-1 (lexical) ambiguity than category + selector features together, which is unsurprising as it is common for a word to correspond to multiple lexical items that differ in their subcategorization properties but still have the same category. Interestingly, more of this ambiguity can be resolved by ISL-2 contexts with category + selector pairs than category features alone. Overall, category feature assignment faces less lexical ambiguity than category + selector assignment while at the same time being harder to disambiguate via contexts.

Identifying the empty heads in the corpus has a profound effect on the inferrability of category features, nearly halving the number of ambiguous items. Even without doing this, however, category features are strongly ISL-2 inferrable for over 94% of all items, and over 75% of lexically ambiguous ones. When this is relaxed from strong to weak inferrability these numbers increase to 98% and 95% respectively. If empty heads are also identified, then category features become weakly inferrable for over 99% of all lexical items, and over 97% of lexically ambiguous ones. Our results therefore

Feature Set	Empty Heads Filled?	Total Items	ISL-1 Ambig. Subtotal	ISL-2 Ambig. Items			
				Strongly Ambig.		Weakly Ambig.	
Category Only	No	29610	8369	1762	(21.1%)	422	(5.0%)
Category Only	Yes	29685	8414	1210	(14.4%)	264	(3.1%)
Category + Selector	No	36635	18124	2861	(15.8%)	808	(4.5%)
Category + Selector	Yes	36688	18157	1571	(8.7%)	330	(1.8%)

Table 1: Count and percentage of lexical items which are ambiguous under each condition tested. Percentages of ISL-2 ambiguous items are calculated w.r.t. the number of ISL-1 ambiguous items as explained in the text.

Issue	Count	%
Wrong category	199	75.4%
Wrong category in other contexts	29	11.0%
Inconsistent category	10	3.8%
Non-alpha symbol	11	4.2%
Ambig. functional head complement	7	2.7%
Problem unclear	7	2.7%
Empty selector and complement	1	0.4%
Total	264	

Table 2: Reason for ambiguity of category features which are not weakly ISL-2 inferrable with identified empty heads.

show that both category features and selector features are largely ISL inferrable using contexts of size 2, which is in line with the conjecture that ISL inferrability is a restriction on category systems in human language.

4.2 Where ISL-2 inferrability fails

Of the subset of lexical items for which category features are not weakly ISL-2 inferrable (with identified empty heads), over 90% correspond to some error in the MGBank corpus. These included an incorrect category label on the lexical item in question, an incorrect category label on another lexical item with the same exponent, or general inconsistency in the category assigned to that form (i.e. one or the other should have been used uniformly). Table 2 summarizes the reasons why weak ISL-2 inferrability failed, with a count of the number of items affected when only category features are included and empty heads are identified. The first four reasons for ambiguity correspond to annotation problems in the corpus, while the rest reflect other reasons ISL-inferrability may have been difficult.

Looking more closely, the most common problem involved noun-noun compounds being misparsed as adjective-noun adjunction structures or vice versa. For example, “desktop computer” and “marketing director” were misparsed as adjective-noun sequences, while “imported steel” and “organized crime” were misparsed as noun-noun compounds. These errors alone accounted for nearly one third of the weakly ambiguous items. Similarly, the first word in a multi-word name like “Bloomfield Hills” or “West German” was occasionally misparsed as an adjective, adverb, or quantifier. In other cases, category for a given item was varied randomly between two reasonable choices. For instance, prenominal quantifiers were sometimes coded as ‘A’ and sometimes as ‘Q’. If the annotation had been consistent, the category would presumably have been recoverable. Overall, there were only a handful of items whose non-recoverability was not obviously related to annotation errors or empty heads.

Taken together, these results are promising for the ISL-inferrability conjecture: the category system used in MGBank displays a high degree of ISL-inferrability, and in cases where inferrability fails this is usually due to errors in the corpus itself.

5 Confounds and caveats

Our findings show that ISL-2 inferrability is an observable trend in MGBank. This could be taken as strong support of the conjecture of Graf (2020) that the category systems of natural languages are ISL-2 inferrable. However, there are several reasons why this might be too strong an inference.

5.1 The problem with adjunction

While ISL-2 inferrability looks like a plausible universal when considering heads and their arguments, it is much more likely to fail for adjuncts.

Consider a language like German, which makes a difference between adjectives and adverbs in

Feature Set	Empty Heads Filled?	Total Items	ISL-1 Ambig. Subtotal	ISL-2 Ambig. Items			
				Strongly Ambig.		Weakly Ambig.	
Movement Only	No	29456	7688	1407	(18.3%)	285	(3.7%)
Movement Only	Yes	29497	7708	469	(6.1%)	35	(0.5%)

Table 3: Count and percentage of lexical items which are ambiguous for movement features. Percentages of ISL-2 ambiguous items are calculated w.r.t. the number of ISL-1 ambiguous items as explained in the text.

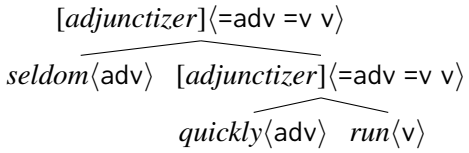

Figure 9: Adjunction as category-preserving selection. An adjunctizer head selects the adjunction site as its complement, and the adjunct itself as its specifier.

terms of distribution (and hence in terms of category features) but does not consistently mark this distinction in its morphology. Hence a form like *schnell* 'quick(ly)' could be an adverb or an adjective in a predicative construction like *Er ist schnell* 'he is fast'. In the analysis assumed by Graf (2020) and also here, adjuncts are modeled as arguments of an empty head – an *adjunctizer*. For example, an adverb adjoining to a VP would be modeled as the specifier of an empty V-head that takes a VP as its complement (Fig. 9). In such a configuration, the category of German *schnell* might not be ISL-2 inferrable. Its lower context would be $\ltimes$, and its upper context would be just the empty adjunction head and its complement, which might be yet another empty adjunction head. This context is equally compatible with *schnell* being an adjective or an adverb.

Since adjuncts are very common, even in corpora, it is suprising that we found such robust support for ISL-2 inferrability. Admittedly, over 40% of (weak and strong) ISL-2 inferrability failures for category feature in MGBank are on lexical items that are used (in at least some instance) as adjuncts, but many of those are related to coding errors. Given that there are theoretical reasons to doubt the viability of ISL-2 inferrability for a very common construction, there is reason to wonder whether the high rate of ISL-2 inferrability found in our study could be due to other confounds in the data.

5.2 Movement features as a control group

In contrast to category features, movement features represent syntactic relationships which are fundamentally non-local. There is no local way of predicting whether, say, an object is topicalized, on the basis of its arguments and seleting heads. Some features are more predictable, e.g. *which* is more likely to undergo wh-movement than remain in situ, and the C-head *do* is very likely to furnish a wh-landing site because of how *do*-support works in English. Still, theoretical considerations lead us to expect low ISL-2 inferrability scores for movement features. But, as shown in Table 3, the scores for movement features are very close to and sometimes even *better* than our findings in Section 4.1 for category (and category + selector) features.

We note that the distribution of movement features in the corpus is highly skewed, with over half of the movement-bearing lexical items being V heads with the +CASE feature. These are almost always dominated by empty transitive little-*v*, and select a DP argument — making +CASE highly inferrable. Even so, the finding is surprising from a theoretical perspective that focuses on what configurations are possible rather than which are common. In order to more accurately tease apart the factors contributing to inferrability, we turn to data simulations to provide a baseline.

5.3 Simulated data

Understanding whether ISL-inferrability is an intrinsic guiding principle of human language or simply a coincidence resulting from other properties requires setting up an appropriate baseline to test how much inferrability we might expect *without* this being an independent requirement of the system. To create such a baseline, synthetic datasets of lexical items and corresponding contexts were created programatically. These synthetic datasets are generated automatically based on I) the desired number of distinct exponents, II) the desired num-

Feature Set	Total Items	Phono. Forms	Ctxs. Per item	ISL-1 Ambig. Subtotal	ISL-2 Ambig. Items			
					Strongly Ambig.		Weakly Ambig.	
Simulation (Category)	29685	24769	11.9	6007	108	(1.8%)	53	(0.9%)
Simulation (Category & Selector)	36688	24769	9.6	13961	245	(1.8%)	137	(1.0%)
Simulation (Movement)	29497	24769	12.0	5806	100	(1.7%)	50	(0.9%)

Table 4: Count and percentage of lexical items which are ambiguous in simulated data. Metrics of total lexical items, phonological exponents, and contexts per item follow those for each category set tested (with filled empty heads). Percentages of ISL-2 ambiguous items are calculated w.r.t. the number of ISL-1 ambiguous items as explained in the text.

ber of distinct lexical items, and III) the average number of contexts in which each lexical item appears. Given these, the synthetic data is generated using the following assumptions:

1. Each exponent appears in at least one lexical item.
2. Each lexical item appears in at least one context.
3. The frequency distribution of phonological items is Zipfian, both in terms of how many lexical items each exponent appears in and in terms of how frequently they are part of contexts for other items. In other words, a few exponents appear in many lexical items while most appear in very few.
4. The frequency distribution of lexical items is Zipfian. In other words, a few lexical items appear in many contexts, while most appear in very few.

For each of the feature sets for which we examined inferrability in MGBank, corresponding synthetic datasets were created with identical values for the number of exponents, lexical items, and average contexts per lexical item. We then tested ISL-inferrability in these synthetic datasets, running three simulations for each experiment and averaging results across the simulations.

The simulated datasets show a high degree of inferrability, comparable to what we find in the actual corpus. Table 4 shows the inferrability results for simulated datasets with metrics matched to the corpus data for each feature set tested (category, category + selector, and movement). These high inferrability rates demonstrate that the simple assumption of Zipfian distributions yields datasets where inferrability arises as an emergent property, rather than being a hard constraint on feature systems.

6 Conclusion

This paper uses the MG treebank (Torr, 2017) to evaluate the conjecture of Graf (2017) that syntactic categories are ISL-2 inferrable over the kind of dependency trees used with Minimalist Grammars. Intuitively, this conjecture states that the syntactic category of a lexical item can be inferred from its own surface form and/or the surface forms of its arguments and/or the surface form of the head it is an argument of. So though the conjecture is stated in very technical terms specific to MGs and subregular syntax, its relevance — and thus the import of our findings — extends to all syntactic formalisms that assume syntactic categories and selectional restrictions. Our analysis of MGBank largely supports the conjecture in Graf (2017) that category systems are ISL-2 recoverable: ISL-2 recoverability fails only for a small number of lexical items, and many of these cases are arguably due to coding errors in the corpus.

However, we also found a high degree of ISL-2 recoverability for movement features and category features of adjuncts, which is unexpected as neither kind of feature should be reliably ISL-2 inferrable. Through simulation, we also showed that a high level of inferrability can result simply from the frequency distribution of language datasets – namely,

a Zipfian distribution.

Together, these findings indicate that human language category systems (and other syntactic features) are reliably ISL inferrable, but that this may not be due to a specific direct requirement for inferrability. In terms of Chomsky (2005), ISL-2 inferrability may be a third factor principle rather than a hard constraint of UG.

Regardless of the reason for which ISL-inferrability appears, its prevalence is a useful property of language to understand. One key benefit to identifying such properties is that they can often be leveraged for learning — just as many proposed language learning strategies leverage the Zipfian distributions that are known to be present. ISL-2 inferrability is particularly suggestive of an approach children may take in learning syntax. It offers a clear direction in which to *generalize*: two phonologically identical items in the same local context *must* also have the same category.

This work furnishes a proof-of-concept for the ISL-2 inferrability of syntactic features and suggests a method for further corpus work which might extend these results to more languages and data sources.

Acknowledgments

The work reported in this paper was supported by the National Science Foundation under Grant No. BCS-1845344 and the Institute for Advanced Computational Science at Stony Brook University. We thank the anonymous reviewers for their extensive comments and suggestions.

References

Noam Chomsky. 2005. Three factors in language design. *Linguistic Inquiry*, 36(1):1–22.

John Doner. 1970. Tree acceptors and some of their applications. *Journal of Computer and System Sciences*, 4:406–451.

Thomas Graf. 2011. Closure properties of Minimalist derivation tree languages. In *LACL 2011*, volume 6736 of *Lecture Notes in Artificial Intelligence*, pages 96–111, Heidelberg. Springer.

Thomas Graf. 2012. Locality and the complexity of Minimalist derivation tree languages. In *Formal Grammar 2010/2011*, volume 7395 of *Lecture Notes in Computer Science*, pages 208–227, Heidelberg. Springer.

Thomas Graf. 2013. *Local and Transderivational Constraints in Syntax and Semantics*. Ph.D. thesis, UCLA.

Thomas Graf. 2017. A computational guide to the dichotomy of features and constraints. *Glossa: a journal of general linguistics*, 2(1).

Thomas Graf. 2020. Curbing feature coding: Strictly local feature assignment. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 224–233.

Thomas Graf. 2022a. Diving deeper into subregular syntax. *Theoretical Linguistics*, 48:245–278.

Thomas Graf. 2022b. Subregular linguistics: bridging theoretical linguistics and formal grammar. *Theoretical Linguistics*, 48(3-4):145–184.

Gregory M. Kobele. 2011. Minimalist tree languages are closed under intersection with recognizable tree languages. In *LACL 2011*, volume 6736 of *Lecture Notes in Artificial Intelligence*, pages 129–144.

Gregory M. Kobele, Christian Retoré, and Sylvain Salvati. 2007. An automata-theoretic approach to Minimalism. In *Model Theoretic Syntax at 10*, pages 71–80.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Frank Morawietz. 2003. *Two-Step Approaches to Natural Language Formalisms*. Walter de Gruyter, Berlin.

James Rogers. 1998. *A Descriptive Approach to Language-Theoretic Complexity*. CSLI, Stanford.

James Rogers. 2003. wMSO theories as grammar formalisms. *Theoretical Computer Science*, 293:291–320.

Edward P. Stabler. 1997. Derivational Minimalism. In Christian Retoré, editor, *Logical Aspects of Computational Linguistics*, volume 1328 of *Lecture Notes in Computer Science*, pages 68–95. Springer, Berlin.

Edward P Stabler. 2011. Computational perspectives on minimalism. *Oxford Handbook of Linguistic Minimalism*.

James W. Thatcher and J. B. Wright. 1968. Generalized finite automata theory with an application to a decision problem of second-order logic. *Mathematical Systems Theory*, 2(1):57–81.

Hans-Jörg Tiede and Stephan Kepser. 2009. Monadic second-order logic and transitive closure logics over trees. *Research on Language and Computation*, 7:41–54.

John Torr. 2017. Autobank: a semi-automatic annotation tool for developing deep minimalist grammar treebanks. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 81–86.

Appendix A: Compiled Results

Category Features

	Corpus		Simulation	
Type	**Ambig. Items**	**% of Lexically Ambig. Items**	**Ambig. Items**	**% of Lexically Ambig. Items**
SL1	8414	-	6007	-
SL2 (strong)	1210	14.4%	108	1.8%
SL2 (weak)	264	3.1%	53	0.9%
Total Items: 29,685	Phono. Forms: 24,769		Contexts per Item: 11.9	

Table 5: Side-by-side comparison of inferrability results for category features (with filled empty heads) from the corpus study and the simulated data. Simulated data was generated using the number of items, exponents, and categories per item which matched the corpus (indicated at bottom of table).

Category & Selector Features

	Corpus		Simulation	
Type	**Ambig. Items**	**% of Lexically Ambig. Items**	**Ambig. Items**	**% of Lexically Ambig. Items**
SL1	18,157	-	13,961	-
SL2 (strong)	1571	8.7%	245	1.8%
SL2 (weak)	330	1.8%	137	1.0%
Total Items: 36,688	Phono. Forms: 24,769		Contexts per Item: 9.6	

Table 6: Side-by-side comparison of inferrability results for category and selector features (with filled empty heads) from the corpus study and the simulated data. Simulated data was generated using the number of items, exponents, and categories per item which matched the corpus (indicated at bottom of table).

Movement Features

	Corpus		Simulation	
Type	**Ambig. Items**	**% of Lexically Ambig. Items**	**Ambig. Items**	**% of Lexically Ambig. Items**
SL1	7708	-	5806	-
SL2 (strong)	469	6.1%	100	1.7%
SL2 (weak)	35	0.5%	50	0.9%
Total Items: 29,497	Phono. Forms: 24,769		Contexts per Item: 12	

Table 7: Side-by-side comparison of inferrability results for movement features only (with filled empty heads) from the corpus study and the simulated data. Simulated data was generated using the number of items, exponents, and categories per item which matched the corpus (indicated at bottom of table).

All Features

	Corpus		Simulation	
Type	**Ambig. Items**	**% of Lexically Ambig. Items**	**Ambig. Items**	**% of Lexically Ambig. Items**
SL1	19,493	-	15,314	-
SL2 (strong)	1832	9.4%	237	1.5%
SL2 (weak)	394	2.0%	133	0.9%
Total Items: 37,873	Phono. Forms: 24,769		Contexts per Item: 9.3	

Table 8: Side-by-side comparison of inferrability results for entire feature string (with filled empty heads) from the corpus study and the simulated data. Simulated data was generated using the number of items, exponents, and categories per item which matched the corpus (indicated at bottom of table).

Appendix B: MGBank Categories

Category	Num. Lexical Items
n	19,585
v	9,574
adj	5,122
adv	964
lv	693
q	589
p	349
D	293
c	147
t	90
part	89
prog	64
mod	64
d	63
perf	32
voice	28
intj	23
tbar	22
negs	21
punc	13
prd	12
neg	12
log	8
ln	8
adjc	2
advc	2
vbar	2
self	1
features	1
top	1
Total	37,874

Table 9: Category features present in MGBank and the number of lexical items of each category.