

Do People Prefer “Natural” code?

Casey Casalnuovo
Department of Computer Science,
UC Davis
ccasal@ucdavis.edu

Kevin Lee
Department of Computer Science,
UC Davis
klee@ucdavis.edu

Hulin Wang
Department of Computer Science,
UC Davis
hulwang@ucdavis.edu

Prem Devanbu
Department of Computer Science,
UC Davis
ptdevanbu@ucdavis.edu

Emily Morgan
Department of Linguistics,
UC Davis
eimorgan@ucdavis.edu

ABSTRACT

Natural code is known to be very repetitive (much more so than natural language corpora); furthermore, this repetitiveness persists, even after accounting for the simpler syntax of code. However, programming languages are *very* expressive, allowing a great many different ways (all clear and unambiguous) to express even very simple computations. So why is natural code repetitive? We hypothesize that the reasons for this lie in fact that code is *bimodal*: it is executed by machines, *but also read by humans*. This bimodality, we argue, leads developers to write code in certain preferred ways that would be familiar to code readers. To test this theory, we **1**) model *familiarity* using a language model estimated over a large training corpus and **2**) run an experiment applying several *meaning preserving transformations* to Java and Python expressions in a distinct test corpus to see if forms more familiar to readers (as predicted by the language models) are in fact the ones actually written. We find that these transformations generally produce program structures that are less common in practice, supporting the theory that the high repetitiveness in code is a matter of deliberate preference. Finally, **3**) we use a human subject study to show alignment between language model score and human preference for the first time in code, providing support for using this measure to improve code.

1 INTRODUCTION AND BACKGROUND

Programming languages are highly flexible, and provide many different ways to express even the simplest computation. However, despite this flexibility, in practice code is highly repetitive, far more so than natural human language. This property of “naturalness” [32] has led to the successful application of probabilistic methods from NLP and machine learning to problems in software engineering, including code completion [31, 48, 64], finding defects [56], and recovering variable names [57, 65].

Here is a puzzle: *why is code so repetitive?* Ruling out the obvious: it’s not just syntax. There is strong evidence [14] that the simpler syntactic and lexical properties of programming language structure, *per se* are not sufficient to explain this. Indeed, Casalnuovo *et al.* [14] report that when the markers of syntax are elided similarly from Code and English, Code becomes relatively *even more* repetitive than English. They also argue that this might arise from conscious choice, citing suggestive evidence of similar, strong, repetitive structure in other corpora that (like code) might entail more effort to read & write, such as legal documents, technical manuals, and English as a second language.

However, Casalnuovo *et al.* [14] didn’t control for meaning. Given a particular meaning, are there different ways to express it? Are some forms preferred? For instance, in English, people prefer to say *bread and butter* rather than *butter and bread* [46]. Likewise, one can code a simple increment operation different ways:

$i = i+1;$ (or) $i = 1+i;$

The two forms above are trivially equivalent to a machine; if developers worked like machines, they should be indifferent to the form. However, most people would strongly prefer the first! This is noteworthy. Unlike natural languages (where the semantics are slippery, and depend in subtle ways on form) programming languages have precisely defined semantics; this means that even the simplest computations can be coded many different, but entirely equivalent ways! Programming languages thus actually provide programmers with *great choice* in forms of expression. So, the plot thickens!! *Given the many expressive choices available, why is code so repetitive?* We believe this is because coders have predictable preferences:

Even given many forms for coding a computation, developers still prefer it to be coded in more familiar forms.

We hypothesize that this preference for more familiar forms is manifest in large corpora, and is evidenced by human code readers. This preference, we believe, accounts for why code is so repetitive, despite the affordance provided by programming languages *per se* to code the same computation in many different ways.

This paper investigates this hypothesis by triangulating an observational study with a human-subject experiment. We model “familiarity” using the probability of occurrence in a large code corpus (the idea being that developers are more familiar with code they see more frequently), and check whether this measure can correctly select preferred forms in unseen code; we also use a controlled experiment to see if this measure can predict human preference.

We make the following contributions:

- Using an adaptation of the UCL-Edinburgh bimodal model [3, 7], we provide a theoretical framing of Naturalness as a preferential choice, where developers deliberately choose to express code in familiar forms, modeled via language model.
- We use transformations to generate alternate meaning equivalent forms of code in a *held-out* test corpus.
- We evaluate the alternative forms using a language model, estimated on a *training* corpus, and find evidence that more familiar (less “surprising”) forms are preferentially deployed in the held-out *test* corpus.

- Using Mechanical Turk, we find that human subjects do prefer forms of code (from the test corpus) that the language model indicates are more prevalent in the training corpus.

The contributions of this paper are primarily of a scientific (rather than Engineering) nature. However, we believe our work provides a new theoretical framework to contemplate a phenomenon of deep current interest to software engineering researchers, as well as a novel experimental approach, and strong connections to well-established theories of human cognition of language. Furthermore, to our knowledge, this is the first study to demonstrate that language model probability, used to guide many code applications, correlates with human preferences for code in a controlled experiment with human subjects.

2 BACKGROUND & THEORY

Natural language (*NL*) corpora exhibit a high-level of repetitiveness, capturable *via* statistical modeling, which has been leveraged in modern tools for translation, speech recognition, text summarization, *etc.* More recently [32] it has been noted that code is also quite repetitive and predictable, and many applications have been developed (See survey by Allamanis *et al.* [3]).

Textual predictability (of code or *NL*) can be captured using *language models* (LM), which assign probabilities to a textual utterances (an *utterance* is a usable language fragment, *e.g.*, a word in English, or a token in code). Probabilities are typically assigned contextually, *e.g.*, probability of *U* occurring in context *C*: $p(U|C)$. The more likely an utterance *U* is in context *C*, the higher the probability. LMs (*e.g.*, *n*-gram models, PCFGs, RNNs) are trained on a large corpus of text, and then are evaluated by scoring against a test corpus. Their performance is measured using a normalized *cross entropy* score (in units of bits), which is estimated by taking the average *surprisal*¹ of the utterances *U* the test corpus *T*:

$$\frac{1}{|T|} \sum_{U \in T} -\log_2(p(U))$$

where surprisal is the negative log probability of an utterance *U* ∈ *T*. A good model, presented with typical text, will find it highly probable, and thus score a low surprisal for most utterances (tokens) in the text, and have overall low entropy.

Various kinds of models have been used in both natural language and in code. Using classical *n*-gram models of code, it has been found that entropy of Java is around 3 to 4 bits lower than for English. This is surprisingly less, suggesting that code is *8 to 16 times more predictable than English* [14, 32]. This gets even more noteworthy given that the *vocabulary* of code, when controlling for corpus size, is typically much larger than natural language, because programmers keep inventing new identifiers. Greater vocabulary *could* mean more word choices to spread probability mass over—but actually doesn't! Why is code so predictable? Is it specific to Java? English? The language model? Is it just because code in general has a much simpler syntax than *NL*? Or is it somehow the result of the cognitive load of reading and writing code?

A recent paper by Casalnuovo *et al.* [14] presents a detailed comparison of predictability in code and *NL*. They find that the greater

predictability of code *vs NL* is not specific to Java and English: a *circa* 2-5 bit (code-*NL*) difference persists, across different programming languages (Java, C, Clojure, Ruby, Haskell), different natural languages (German, Spanish, English) *and* different language models (*n*gram, Cache-based, and LSTM [33, 63]). This persistent, robust difference between programming and natural languages suggests that something, perhaps the simpler syntax of code, is the cause. To deal with the question of syntax, they removed keywords, operators and delimiters from code, and analogous syntactic markers (prepositions, determiners, conjunctions, pronouns *etc*)² from *NL*. After the concomitant removal of these syntax markers, in code they are left merely with identifiers, and in English with just nouns, verbs, adjectives and adverbs³. They find that after the removal of syntax markers, the difference between programming languages and english **increased** to between 4-8 bits depending on the language model used. Finally, they incorporated full parse tree structures into code (using ASTs) and English (using the Penn Tree Bank [45]) and found that code still remains more predictable than English.

Why is this? It's certainly not because code lacks expressive power! Alternative forms abound even for something very simple. Consider the wealth of equivalent alternatives for even the trivial iteration trope: `for (i = 0; i < n; i = i+1) { ... }`. One could pick a name other than `i` or `n`; flip the conditional; use a different incrementing form; start and end differently, *all without changing the meaning*. Indeed, a literal infinity of equivalent forms are possible! Still, we persist with such tropes. Why? To further clarify, we draw upon a formulation [3, 7]⁴ that software is *bimodal*: it works on a human-machine channel, and a human-human channel. It is argued that programmers must write code with full awareness of *two modes* of eventual use: first, code has formal *operational semantics*, and executes on a machine; second, code is maintained by other programmers, who must understand it; and thus code *per se* forms a vital communication channel between the developer who writes it and the maintainer who cares for it.

This *bimodality* argument implies that two *distinct* channels exist, and suggests an origin for the high repetitiveness of code. First, suppose that the human-human channel simply re-used the formal operational semantics channel. Consider a code reader *Rick*, who is examining a piece of code *C*, written in language *L* by developer *Doris*. *Rick* desires to know its computational meaning *M*. Suppose *Rick* could directly, by himself, mentally calculate the meaning of *C* via the operational semantics of *L* quickly and easily. If readers can *always* and *efficiently* do this, there is certainly no constraint on developer *Doris*. She is free to choose from any of the choices *C_i* that implement *M*, since her readers behave like infallible machines, and will reliably and efficiently find the meaning *M*, even if she chooses bizarre (but correct) ways to implement *M*. However, readers are not machines, and the human-human channel works differently. Prior work on natural language in Psycholinguistics [58, 68] indicates a robust association between greater statistical predictability (surprisal) and ease of production & comprehension. Speakers are more likely to choose more predictable utterances, even among

²In linguistics, these are called *closed category* words, since new words in this category are very very rarely coined.

³This concomitant removal of syntax markers was done differently in [55], which explains the different findings.

⁴Talk and long paper; §3 is most pertinent in the paper.

¹As a reminder, *surprisal* of an event $\epsilon \in$ an event space *E*, with probability $p(\epsilon)$, is $-\log(p(\epsilon))$; entropy is the expectation of surprisal over all $x \in E$.

meaning-equivalent options [35]. These choices are likely driven in part both by *audience design* [36] (*viz.*, choose utterances for ease of comprehension) and by ease of production [8, 24].

On the human-human channel, we might expect code to behave like natural language. Given a meaning, some implementation forms may be more familiar, such as the `for` loop above, and will be more easily recognized. Indeed, an experienced developer will know this, and would prefer to use familiar forms whenever possible, both for her own and her reader’s convenience.

We can formally state this thus. Assume C_1, C_2, \dots, C_n are viable implementation choices for a given computation M . Although these are all semantically equivalent, for human convenience, there would be a tendency to prefer one over the others. If p refers to probability of occurrence in a corpus, we should observe the following:

$$\exists C_i, i \in \{1 \dots n\} : p(C_i | M) \gg p(C_j | M), j \in \{1 \dots n\}, j \neq i$$

We are overstating things here: it’s possible that a *few of the possible* C_i are more preferred, *viz.*, the probability mass is not uniformly spread over all the choices. Another way to state this would be that the *entropy* of this conditional distribution is less than the possible maximum which would be obtained in the case of a uniform distribution among implementation choices. However, it is difficult to know in general the number of implementation choices. Given the intractability of computing such a maximum, we formulate the central question a bit more informally:

For a given computation, do developers prefer some implementations over others, and to what degree?

We now describe our experimental decisions and specifics of our approach. First, in order to model the developer “preference” above, we use *language models*. A modern language model, well-trained over a large, diverse corpus, can reliably⁵ capture the frequency of occurrence of textual elements in a corpus, thus capturing the preferences of programmers who created that corpus.

Second, we use meaning-preserving transforms to model a range of possible implementations C_i for a meaning M . While many are possible, we focus on 3 types of transforms: expression rewriting, nonessential parentheses adding and removing, and variable name shuffling (for details see Section 3.3). These transforms are performed on code fragments from a corpus “unseen” by the language model, which is then used to score the surprisal of the original and the transformed versions. These transforms generally perform changes of a scope confined enough to be reasonably captured by our 6-gram language models and the additional LSTM language models we use to validate effects seen on the ngram transformations.

Our first RQ investigates whether developer preferences for restricted forms of expression are observable in Java:

RQ1. *Using a language model trained on Java code, if we perform meaning preserving transformations on unseen Java code, to what degree does the model find the transformed code more improbable (higher surprisal)?*

Next, to ensure that this is not simply an effect of the choice of programming language, we also choose a secondary language that is different from Java, Python, and ask:

⁵The low cross-entropy that modern models provide over unseen corpora is evidence of their power.

RQ2. *Is the Language Model’s preference for the original code also observable in Python?*

Beyond simple ngram models, we would also like to explore how local style effects the consistency of choice. Prior research has shown this to be a strongly distinguishing factor of source code over natural language [31, 64], so we theorize that cache models would prefer the original code even more strongly. Additionally, we would like to consider if these preference patterns are retained in the underlying structure of the code - i.e. when identifiers and literals are abstracted. Thus we ask:

RQ3. *Do cache based models that incorporate local style discriminate the original code more strongly? Is the preference for the original code retained even when abstracting identifiers and literals?*

We expect that some transformations will disrupt the code less or even make it more probable to our models. In particular, we would like to see how the original surprisal of the code relates to the effect of the transformation. We would expect highly improbable code to have greater potential to become more “typical” after transformation. Such code will likely be associated with less restrictions on developer choice. Thus, we ask:

RQ4. *How does the original “surprisal” of the original code relate to the effect of the transformation? Do high-surprisal and low-surprisal code behave differently?*

Finally, to compare language model judgements with human preference and validate model surprisal as a measure of preference in the context of code, we run a human subject study and ask:

RQ5. *Do the preferences of Java programmers align with language model surprisal? How do these preferences vary by transformation?*

In summary: why is code so repetitive? Prior research clearly indicates it’s not just syntax. We hypothesize that coding behavior, because of the the human-human channel, is susceptible to some of the same production and/or comprehension pressures as natural language. Thus, we expect that, for convenience, developers strongly prefer certain forms of expression, despite the great variety of options provided by programming languages. Moreover, these forms are *predictable*, and can be captured via language models.

3 METHODOLOGY

We adopt a *triangulation* approach, combining a *natural experiment* on a large code corpus with a *human subject* study. A large code corpus embodies numerous choices made by programmers, and thus is a representative sample of these choices. Within this corpus, we can examine the occurrence frequency of different implementation choices C_i of the same computation M , and determine if some choices dominate. For triangulation, (Section 3.4) we use human subjects; we ask them to preferentially select from two alternative implementations C_1, C_2 with the same meaning M , chosen such that the language model assigns quite different surprisal scores to C_1 and C_2 . We test if the surprisal scores predict human preference.

3.1 Code Corpora

Our experimental dataset is chosen to help control for potential confounds, while also affording enough opportunities for transformations. Since our main focus is Java, (see RQ 1) we use a larger Java corpus and replicate with a smaller corpus in Python (RQ 2).

We cloned the top 1000 most starred projects on Github for Python & Java. We use a subsample of these projects due to computational constraints; we select the 30 projects from Java & Python with the highest count of possible transformations. These projects are then randomly divided by project into a 70-30 training/test split. In Python, due to the lack of typing and limitations of the *Abstract Syntax Tree (AST)*, we replicate only one type of transformation, swaps over relational operators. These limitations also required us to normalize the original Python files with *astor*⁶, which can slightly change parentheses. As we do not perform parentheses transformations for Python, this should only minimally impact results.

Duplication can be a potentially confounding effect in training and testing code with language models [1]. Since our focus is on programmer preferences for certain coding forms, it would be inappropriate to remove all clones. Still, to avoid large-scale duplicated code, we do a lightweight removal of fully duplicated files, with additional filtering during testing for stability. This lightweight process compares the name of every file and its parent directory (e.g. `main/ExampleFile.java`), keeps the first one seen of an equivalent set, and removes the others from our training/test data.

Table 1: Summary of Java and Python datasets.

Language	Files	Unique Files	Training Tokens
Java	204489	184093	~118.5 M
Python	27315	23105	~18.2 M

Table 1 shows the file-counts and approximate token-counts in the training set for each corpora. Duplication filtering removes 6.1% and 10.7% of files in Java and python respectively. Despite sampling the same number of projects, the Java corpus is much larger, but as Java is our main focus, and Python is simply to see if the results replicate across languages, this is arguably adequate.

Our test data was chosen to be **distinct** from the training data. In addition, we removed lines commonly associated with generated code, coming from *equals* and *hashCode* functions⁷. These lines are generated by IDEs, and arguably do not accurately represent *human written choices* (or at least, human style choices so codified that they have been automated.). We also remove from the test data identical lines of code appearing more than 100 times⁸, as these may also be at risk of copy-pasting. We believe that it would *not* be correct to simply filter out all duplicated expressions, as it is perfectly valid for developers to rewrite the same code. Since our study is largely at the expression level, it’s difficult to precisely find & account for copy-pasting; we argue our approach gives a reasonable middle-ground between removing the extreme cases while still retaining most of the natural repetition of code. Finally, we note that we *did not* remove repeated or generated code *fragments* from the *training* data to properly reflect the code that programmers would read (and learn preferences from). Our test set pruning was to avoid overly

⁶<https://pypi.org/project/astor/>

⁷We dropped lines with the string’s `hashCode` or `other` which we observed manually to be contributing to this repetitiveness. In the case of our identifier shuffling transformations which operate at the method rather than expression level (see section 3.3) we instead removed all `equals` and `hashCode` methods.

⁸A threshold of 10 gave similar results, suggesting robustness.

weighting repeated and generated code, and emphasize more the individual, independent choices made when *writing* code.

3.2 Language Models

We estimate a language model *LM* over a large corpus, and then use the $P_{LM}(C)$ from the language model as an indication of developer preference. Specifically, we use *surprisal* of *LM* with respect to a fragment *C*, (which is precisely $-\log(P_{LM}(C))$): lower surprisal indicates higher developer preference. This use of surprisal is not unprecedented; abundant experimental evidence from psycholinguistics indicates surprisal is strongly associated with cognitive effort in language comprehension [21, 25, 29, 41, 42].

We use 4 ngram language model variants to capture various aspects of possible developer preference. First, we use a basic 6-gram model with Jelinek-Mercer smoothing, using the best order and smoothing recommended from past research [31], denoted as the *global* model. To answer the two parts of RQ 3, we first use an *ngram-cache* (henceforth abbreviated as *cache* model), as originally described by Tu *et al.* [64] to capture local patterns. Then, we build an alternate training and testing corpus where we use the Pygments⁹ syntax highlighter to replace all identifiers and types with generic token types, and literals with a simplified type¹⁰. These models are implemented in the SLP-Core framework by Hellendoorn *et al.* [31]¹¹. To assess preference, we compare the average surprisal of tokens that appear *only in both the original and the transformed version of the expression*. The tokens not involved in the changed expression are not considered. Finally, we validate the robustness of our ngram results with a 1 Layer LSTM implemented in Tensorflow, trained with 10 epochs and 0.5 dropout, for the corpus results of the 4 transformations used in the human study¹².

3.3 Meaning Preserving Transformations

We choose not to use existing transformation tools, as they are either not meaning preserving (e.g. mutation testing [39, 44]), or operate at the wrong scale of code object, such as compiler optimizations. We use source level transformations that both are meaning-preserving *and* small enough in scope to be captured by language models. Our focus is primarily on transformations of source code *expressions*, which we implement via the Java and Python *AST*. For Java, we use the *AST* Parser from the Eclipse Java development tools (JDT)¹³ and the *ast* module from Python3.7¹⁴.

We implement 12 different kinds of transformations, summarized in Table 2, grouped roughly into 3 categories: 1) swapping transformations, 2) parenthesis transformations, and 3) renaming transformations. There are 6 non overlapping sub-groups: arithmetic and relational swaps, parenthesis adding and removing, and shuffling identifiers within and between types.

⁹<http://pygments.org/>

¹⁰For example, we keep 1,2,3 and replace higher numbers with labels like `<int>` and `<float>`. For strings, we keep the empty string, single character strings, and replace everything else with `<str>`.

¹¹<https://github.com/SLP-team/SLP-Core>

¹²We only considered these 4 as training and testing in context is computationally expensive, and it is just to validate the transformations we focused on in both studies.

¹³<https://www.eclipse.org/eclipse/>

¹⁴<https://docs.python.org/3/library/ast.html>

Table 2: Pseudocode examples for the transformations.

Swap	Arithmetic	*	a * b	b * a
		+	a + b	b + a
	Relational	==, !=	a != b	b != a
		<, <=, >, >=	a <= b	b >= a
Paren.	Adding		a + b * c	a + (b * c)
	Removing		a + (b * c)	a + b * c
Rename	Within Variable Types		int a int b	int b int a
	Between Variable Types		int a float b	int b float a

Swapping transformations: We have 8 kinds of transformations involving swapping and inverting operators, divided into 2 subcategories. The first subcategory swaps arithmetic operands around the commutative operators of + and *. We swapped *very* conservatively. We limit the types of the variables and literals in the expression to doubles, floats, ints, and longs. Infix expressions with more than two operands are only transformed if the data type of the operands are int or long to avoid accuracy errors due to floating point precision limitation. We also exclude expressions that contain function calls, since these could have side-effects that alter the other variables evaluated in the expression.

The second subcategory of operator swapping involves the 6 logical operators, ==, !=, <, <=, >, >=. We flip the subexpressions that make up the operands for each of these, either retaining the operator if it’s symmetric !=, ==, or inverting it if it’s asymmetric (e.g. > becomes <). While we do not limit the types in these expressions as they are commutative and don’t risk floating point issues, expressions with function calls are excluded to avoid side effects.

Parenthesis transformations: The next category involves manipulation of extraneous parentheses in source code. Programming languages have well-defined operator precedence, but programmers can still (and often do) freely *choose* to include extraneous parentheses for readability. For instance, in cases where less common operators are used (such as bit shifts), the parentheses may make comprehension easier, leading to a preferred style.

Therefore, we can transform expressions by adding or removing extraneous parentheses from expressions. The adding parentheses transformation relies on the tree structure of the AST to insert parentheses while preserving the correctness in the order of operations. Parentheses are not added to expressions whose parent is a parenthesized expression to avoid creating double parentheses. Parentheses are also never added around the entire expression.

For parentheses removal, we select each parenthesized expression. Then, each of these are passed to the NecessaryParenthesesChecker from the Eclipse JDT Language Server¹⁵ to check if removing them would violate the order of operations. Any subexpressions that pass this check are then considered candidates for removal. This method is used by the same algorithm supporting the “Clean Up” feature within the Eclipse IDE¹⁶.

Variable shuffling transformations: Finally, we consider transformations that shuffle the names of identifiers. To avoid changing

meaning, we swap only within a method, using the key bindings of the AST to maintain scoping rules. If a variable name is used for a declaration more than once in a function (e.g. multiple loops using *i* as a variable for iteration), it is excluded to avoid assigning two variables the same name within the same scope. Methods containing lambda expressions are also ignored because their variable bindings are not available in the AST.

We separately consider renaming both *within types* and *between types*. As an example, consider a function with two *int* and two *String* variables. In the *within types* case, we only consider replacing one integer’s name with the other, and the same for the Strings. In the *between types* case, all four variable names can be assigned to any of the integers or strings other than their original variable. We expect that names given to the same types will be used more similarly than names given to different types. Thus, we would expect *between types* transformations result in code relatively more improbable than those produced by *within types* transforms.

3.3.1 Transformation Selection. As expressions grow in size, the number of possible transformations grows exponentially. Generating all these transformations are neither feasible or desirable, so we select a random subsample. For the operand swapping and parentheses modification cases, we randomly sample up to *n* transformations, where *n* is the number of possible locations to transform in the expression. For variable renaming we consider only functions with up to 10 local variables that can be shuffled.

3.4 Human Subject Study

The corpus study can tell us if some forms are preferentially used in a large corpus. We triangulate with a human subject study, checking if human preferences over different implementations of the same computation align with different language model surprisal scores. While lower surprisal has been linked to easier processing in natural language [21, 25, 29, 41, 42], we don’t know of similar findings for code, despite the metric being used as a stand-in for preference or comprehensibility in guiding many tools (e.g. [2, 31, 43]).

We use Amazon’s Mechanical Turk (AMT)¹⁷ for this investigation into the alignment of surprisal with human-subject preference. AMT has been used in several other programming studies to recruit subjects [5, 15, 54]. Since anyone can sign up with AMT, we selectively filter out a sample that can reasonably represent Java programmers. First, we follow recommended guidelines [20] for avoiding bots and poorly qualified workers; we require a 99% HIT acceptance rate, 1000 or more completed HITs, and restrict workers to those in the US and Canada. We also used Unique Turker¹⁸ along with AMT’s own internal reporting to remove any repeat users. Secondly, we deploy a short qualification test which requires subjects to read some Java code and answer 3 comprehension questions; all 3 must be correctly answered. We tuned our comprehension questions with 3 pilot surveys. MTurkers were paid at minimum wage rate (\$12/hour) for tasks they completed; the qualification test was estimated to take 5 minutes, and the main survey 20 minutes. Our survey asked forced-choice preferences of this form:

Please select which of the two following code segments you prefer:

¹⁵See <https://github.com/eclipse/eclipse.jdt.ls>

¹⁶See https://bugs.eclipse.org/bugs/show_bug.cgi?id=405096

¹⁷<https://www.mturk.com/>

¹⁸<https://uniqueturker.myleott.com/>

```
outPacket = new byte[10 + length];
outPacket = new byte[length + 10];
```

The alternative segments were selected from the relational & arithmetic swaps, and parenthesis adding & removing. Using the *global ngram model*, after some filtering¹⁹ we selected the top 20 single line transformations that most *increased* and *decreased* the average **line level** surprisal over shared tokens, replacing some from the top 40 when cases were too similar, or the transformation obviously disrupted symmetry²⁰. We use line instead of expression level averages from the corpus study because the subjects judged the entire lines. From these 160 pairs, we presented 80 to each user randomizing both the questions and order of the choices.

To measure subject attention in a 20 minute survey, we included an unidentified attention check, which was a question like the others, except more obvious and incontrovertible²¹. We do not exclude those failing the attention check (it was only one question of many), instead using it as a measure for how attentive the subjects were overall. As long as failing the check is not common, we can be confident of reasonably attentive subjects.

3.5 Modeling

To compare surprisal before and after the transformation, we use paired non-parametric Wilcoxon tests [34] and associated 95% confidence intervals, which measure the expected difference in medians between the original and transformed code. We also *widen* the intervals using the conservative family-wise Bonferroni [67] adjustment, to account for the tests on each model and transformation.

To answer RQ 4, we turn to regression modeling. Recall that we theorize that expressions that are *more improbable* to the language models should be more amenable to becoming more probable after transformation, whereas *low surprisal* would be associated with stronger norms and thus more harmful transformations. We measure this effect with ordinary linear regression; we use controls for the size of the line, the type of AST node that is the parent of the expression, and a summary of the operators involved in the expression²². Our regressions are limited to single transformations for ease of interpretation, and filtered out rare parent and child types (< 100)²³. We identified influential outliers using Cook’s method [19] and removed those with values greater than $4/n$. We examined residual diagnostic plots for violations of model assumptions, and made sure multicollinearity was not a issue by checking that variance inflation (VIF) scores were < 5 [16].

For our human subject study, we use a mixed effects logistic regression. As the complexity of the random effects structure of our model caused the frequentist estimate to not converge, we estimate the model via Bayesian regression through the R package brms [13]. Our presented model used the default priors of the package, but we validated convergence and alternative priors using the guidelines

¹⁹Code with hashing and bit shift keywords, lines over 80 characters.

²⁰For example, only adding one parenthesis to `a == b || b == c`

²¹We asked if `for(int i = 0; i < length; i++) {` was preferable to `for(int i = 0; length > i; i++) {`.

²²In the case of multiple operators, we selected the most common one from the training projects to represent the expression.

²³Limitations in the python transforms prevent an accurate count of the number of transformations, and so the first filter was not applied. We examined the comparable Java models without this filter as well but found little difference in the coefficients.

included in the WAMBS checklist [22]. Further details on these models are omitted for space but can be found in the R Notebooks in our replication package (see intro of Sec. 4).

4 RESULTS

Our results from the corpus study are in § 4.1 and the human subject study § 4.2. Due to space limits, our corpus study focuses on the swap transformations, and highlights differences in the other transformations. All our data, R notebooks, and results can be anonymously accessed at <https://doi.org/10.5281/zenodo.2573389>.

4.1 Corpus Study

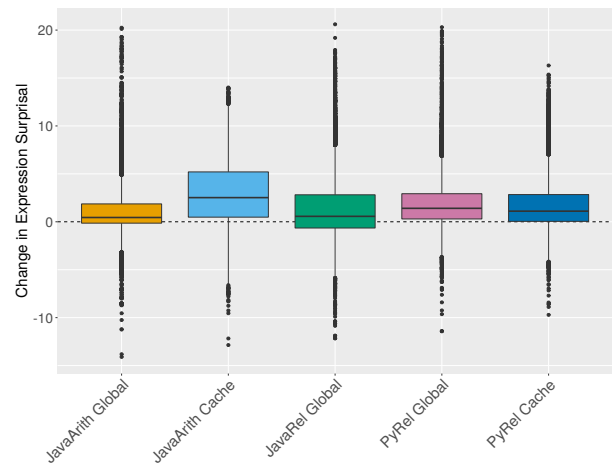


Figure 1: Average surprisal change for swaps: Java arithmetic and relational, and Python relational. Java relational cache is omitted as it is consistent with arithmetic. Positive values indicate the transformation is less predictable.

4.1.1 Swapping Expressions. We have 20,829 instances of transformable arithmetic swaps in our Java data, and 133,845/32,219 instances for relational swaps in Java and Python respectively. Figure 1 shows the difference in ngram surprisal (transformed - original) for all of the concrete swap transformations, except for the cache for the Java relational swaps, which is similar in effect to the Java arithmetic swaps. Rows 1,2, and 3 in Table 3 show the associated Wilcoxon tests and confidence intervals around the median for *all* model variants.

In general, the data supports our theory that the original code would be preferred by the language model (LM) of the swaps to varying degrees. For the Java arithmetic swaps, the global LM finds the original code 1.68 times more probable (0.75 bits of surprisal less) than the transformed version, and the cache model finds the original 8 times more probable (3 bits less) than the transformed version²⁴. The cache LM does discriminate better across all the Java transformations, both with concrete and abstracted identifiers; but *not*, however, in Python, perhaps due to a global “Pythonic” culture that transcends project boundaries. The LSTM models we ran to validate the robustness of the ngram models on Java swaps also

²⁴The increase can be measured as 2^{diff} e.g. $2^{0.75} = 1.68$.

Table 3: Two sided paired Wilcoxon signed-rank tests and 95% confidence intervals of surprisal difference *original source minus transformed source*. A 1 bit negative difference indicates the original code is twice as probable as the transformed code. Intervals are Bonferroni corrected. ◦ indicates $p > .05$, otherwise $p \ll .001$

	Global	Cache	Global Abstracted	Cache Abstracted	LSTM
Arithmetic Swap	-0.7966, -0.7049	-2.9683, -2.8249	-0.5575, -0.5056	-1.4682, -1.3902	-0.751, -0.6271
Relational Swap (Java)	-1.2168, -1.1683	-1.9637, -1.9053	-1.7201, -1.701	-1.9539, -1.9361	-1.0059, -0.9517
Relational Swap (Python)	-1.7639, -1.6841	-1.4611, -1.3783	-2.481, -2.4066	-1.5444, -1.4744	-
Add Parentheses	-0.3342, -0.3202	-0.5769, -0.5434	-0.235, -0.2238	-0.3782, -0.3641	-0.543, -0.4922
Remove Parentheses	-0.0103, 0.0344◦	-0.4388, -0.361	-0.2004, -0.1497	-0.4229, -0.3629	-0.5552, -0.4254
Variable Shuffle (Within Types)	-0.8047, -0.6761	-0.9738, -0.8411	-	-	-
Variable Shuffle (Between Types)	-1.731, -1.6121	-2.3268, -2.2314	-	-	-

indicate this effect, showing preference for the original code. Like the global ngram models, the effect is stronger for relational swaps.

Now, to answer RQ 4, we will describe one regression model in depth: the one for the global arithmetic swaps. We model *Surprisal Change* \sim *Original Surprisal* + $\log(\text{NumTokens})$ + *ParentOperator* + *Operator*, the change in surprisal as predicted by the original surprisal, controlled by both the log size of the expression, and the parent and operator types of the expression. For every bit increase²⁵ in the original expression, the change decreases by .279 bits. This effect is quite strong, explaining nearly 35% of the variance in the difference. We can conclude that less predictable expressions exhibit less strong norms, and the effect of a transformation is more variable. This negative correlation between original surprisal and the change also holds in the regressions for all the other language models on this transformation. Among the controls, longer expressions are also more likely to be amenable to transformations, and while most parent nodes in the AST are similar to the ‘==’ baseline, return statements and array accesses tend to have less strict style. Finally, swaps that occur on a ‘+’ instead of a ‘*’ are 55% less probable to the language model - likely attributable to addition being much more common than multiplication.

This effect of higher original surprisal leading to greater opportunity to make code more predictable persists across all the regressions for the Java relational swaps, and for the Python swap using the *concrete* code. However, in the abstracted Python code, this effects is reversed, but explains only a small amount of the variance of the change (<2.5%). Further study is required to understand this counterintuitive behavior in the abstracted code.

So for Java swapping transformations, RQs 1, 3, and 4 are answered affirmatively, albeit to various degrees. The models prefer the original source code regardless of model, cache models more strongly discriminate between the original and transformed code, and less probable expressions are associated with smaller increases, or even sometimes reductions in surprisal. Our results for RQ 2 provide additional support for the overall theory, but suggest complications in the details. The ability of the locality to discriminate may be language specific, and the relationship between the original expression surprisal and the change in surprisal in the abstracted models (different in direction from all other results) may suggest Python norms more closely tied to identifiers.

4.1.2 Other Transformations. For parentheses, we have 63,625 additions and 9,717 removals, with the results shown in rows 4 & 5 in Table 3. The results for surprisal change, the effect of the cache, and the regression models built to answer RQ 4 are similar to those in with the swaps, with one major exception. The difference between the original and transformed code in the *global* model is *not significant!* We delve into this unexpected result more closely using examples in Table 4. These are fairly intuitive; the biggest improvement in predictability comes from removing parentheses unnecessary to clarify the order of operations from around a literal denominator. In contrast, a large increase in difficulty of prediction occurs with rarely used bit shift operators—suggesting that developers may prefer parenthesis around rare operations to clarify order of operations. In contrast, the LSTM models agree with our theory, although the change is smaller relative to the swaps. Thus for we answer RQs 1, 3, and 4 affirmatively, except for the global ngram models of parenthesis removal. We speculate that this may be the result of less consistent style around the usage of parenthesis, similar to what Gopstein *et al.* [27, 28] found with bracket usage (see 6.1). We further discuss the influence of style guidelines in § 5.

Finally, we consider variable renaming transformations, measuring mean surprisal change across all affected expressions *within the same method*. There are 17,930 methods with shuffling within types and 48,160 with shuffling between types²⁶, with results in rows 6 and 7 of Table 3²⁷. As expected, shuffling variable names within a method increases surprisal. Variable names matter for program comprehension, and obscuring these names is one of the most common and simple forms of program obfuscation [17]. Moreover, we confirm that swapping variable names across types is more disruptive to predictability; the difference is about twice as large. Cache effects are still present, but diluted, possibly because the shuffle pulls its vocabulary from very similar contexts. As with all other Java transformations, the regression models show that variable names scoring higher in surprisal have less of a surprisal increase after renaming. So in conclusion, the renaming shuffle transformations all answer RQs 1, 3, and 4 as expected, with stronger results when shuffling between rather than within types.

²⁵Meaning the expression is twice as difficult for the model to predict.

²⁶Unconstrained shuffles are possible in more methods.

²⁷As swaps operate on concrete identifiers so the abstracted models do not apply.

Table 4: Sample transformations with the largest surprisal changes for parenthesis removal with the global model.

Original	Transformed	Surprisal Change
double seconds = time / (1000.0);	double seconds = time / 1000.0;	-13.878
return ((dividend + divisor) - 1) / divisor;	return (dividend + divisor - 1) / divisor;	-7.991
int elementHash = (int)(element ^ (element >>>32));	int elementHash = (int)(element ^ element >>>32);	11.321
c1 = (c2 >>4) & 0x0f;	c1 = c2 >>4 & 0x0f;	11.053

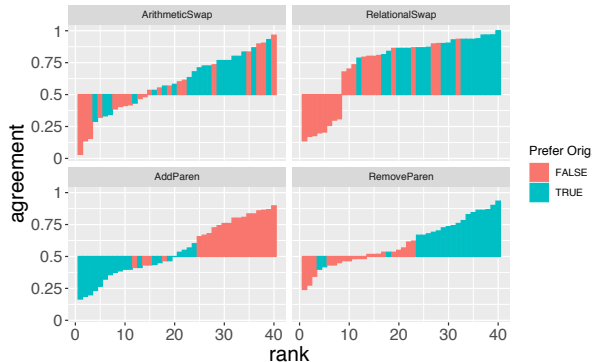


Figure 2: Break down of fraction of agreement for each question by transformation, ordered from questions with the least agreement with the model to the most. Majority vote agreements: ArithmeticSwap (65%), RelationalSwap (80%), AddParen (50%), RemoveParen (67.5%).

4.2 Surprisal and Human Preference

Our survey netted a total of 180 attempts across 3 batches, with 60 non-duplicate MTurkers fully completing the survey. Of the 60, 50 passed the attention check, though there is little difference in overall agreement with the language model between these groups. Demographically, our group had a median of 4.5 years of Java experience and 9 years of general programming experience, and were primarily developers, students, and hobbyists who coded at least a few times a week. All but one had some college education, and most used AMT for extra income.

Overall, 61.9% (65.6% with majority vote)²⁸ of the time our subjects agreed with the global ngram model. Figure 2 groups the results by each question, broken down by transformation type. Each quadrant has 40 questions arranged by rank in increasing order of fraction of human agreement with the language model, with bars pointing downwards indicating more disagreement than agreement. Red examples show where the language mode preferred the transformed code, and blue examples show where it preferred the original.

Humans overall agree with the language model swaps much more frequently than the parentheses changes. Relational swaps demonstrate have the highest agreement (indicated by values above 0.5). All of the disagreements are cases where the raters agreed with the original code (but the language model disagreed), suggesting a limitation in the language model rather than disagreement among coders. For parentheses, we also see a pattern: our group tends to

²⁸62.8 and 66.9% for those passing the attention check.

prefer variants with more parentheses (indicated by the reversed red/blue patterns in AddParen and RemoveParen Figure 2), regardless of the language model preference. Moreover, the language model poorly predicts human majority vote preference for adding parenthesis - agreeing only half the time.

Table 5: Fixed effects for bayesian mixed effects logistic regression on our human subject study. *Parenthesis removal (RMParen) does not get an independent coefficient estimate in deviation coding, so we calculate the implied coefficient.

	Estimate	Error	1-95% CI	u-95% CI
Intercept	-0.60	0.15	-0.90	-0.30
LM_Out	1.90	0.23	1.45	2.37
AddParen	-2.03	0.58	-3.19	-0.95
Arithmetic	0.49	0.25	-0.01	0.96
Relational	0.15	0.26	-0.39	0.68
RMParen*	1.39	–	–	–
LM_Out:AddParen	-0.79	0.42	-1.63	0.04
LM_Out:Arithmetic	-1.03	0.37	-1.74	-0.32
LM_Out:Relational	1.64	0.45	0.77	2.50
LM_Out:RMParen*	0.18	–	–	–

Examining the results in greater detail we present our Bayesian mixed effects logistic regression in Table 5. The model formula is: $Outcome \sim LM_Out * TransType + (1 + LM_Out * TransType / RespondId) + (1 | Question)$. The Outcome is 1 if the human subject selected the original code, 0 for the transformation. The fixed effects are a binary predictor (LM_Out), which is 1 if the language model selected the original code, and 0 otherwise, along with the type of transformation and their interaction term. We use the maximal random effects structure justified by the design [6] - a random intercept by question, a random intercept and slopes for surprisal, transformation type, and their intercept by subjects. The transformation types are deviation coded, meaning the intercept value is the grand mean over all transformations. Thus, for the coefficients for the parenthesis removal transformation, we subtract the 3 of other type coefficients and provide it in the table for convenience. Finally, Bayesian estimates do not have p-values and confidence intervals in the same way as frequentist approaches. Instead, we report the equivalent 95% credible interval, the bounds of a probability distribution that has a 95% probability containing the regression coefficient. If this range is entirely above 0, it indicates a positive effect, and vice versa for a range entirely below 0.

As this is a logistic regression, the coefficients are log odds ratios. The odds ratio of the intercept by itself, $exp(-.60) = .55$, shows that when the language model prefers the transformation, our subjects

are $1/.55 = 1.8$ times more likely to also prefer the transformed code. By contrast, when the language model prefers the original code, our subjects are $\exp(-.60 + 1.9) = 3.7$ times more likely to also prefer the original code. Importantly, the crucial predictor, `LM_out`, has a positive effect with 0 well outside its credible interval. Thus, on average, not only do humans agree with the language model more often than not, they agree with the language model almost *twice as strongly* in its judgements on the original code. This effect could pose risks to tools that use surprisal to guide transformations, as the new code may not be judged as reliably by the models. Finally, some but not all of the specific transformations also differ significantly from the grand mean in either their baseline effects or their interaction terms. In summary, we can say that the model confirms what was seen in Figure 2, that humans agree with the language model, except when it prefers the original code in an adding parenthesis transformation.

Therefore, our data supports RQ 5. In most cases, the humans tend to agree with the language model. Adding/removing parenthesis behave differently, again suggesting that human preferences are more variable for this transformation than for the swaps. In particular, perhaps a more powerful language model could better capture preferences for adding parentheses.

5 DISCUSSION

Our finding that developers prefer more predictable code forms is consistent with results from psycholinguistics, supporting that the repetitiveness of code comes from the human-human channel. The question remains: why is code *even more* repetitive than natural language? We propose some theories to be tested in future work.

One possible reason for the greater repetitiveness of code is that (unlike with code) natural language utterances are very very rarely entirely meaning equivalent—since NL has connotative as well as propositional meaning. For instance, while “bread and butter” is *prima facie* synonymous with “butter and bread”, one might infer from a description of eating “butter and bread” that butter was unusually dominant in this situation. These subtle differences of meaning may drive speakers to choose different forms in different situations. However, $i = 1+i$ and $i = i+1$ are semantically equivalent on the human-machine channel. Perhaps because developers are trained to understand the true operational semantics code, different connotations for the above two seem unlikely to evolve, even on the human-human channel. Thus, it is difficult to imagine how $i = 1+i$ might carry a subtly different and useful connotation, even on the human-human channel; and thus it is difficult to find a situation (analogous to “butter and bread”) where a developer might consciously feel the need to use that construction.

Another possibility is that repetitiveness is particularly beneficial in situations with increased cognitive load. Indeed, it has been proposed that children resort to repetitiveness in language learning more so than adults specifically because they have reduced cognitive capacity [35, 60]. Because code comprehension is challenging, repetitiveness may be extra beneficial for code compared to natural language. Finally, repetitiveness may arise from pedagogical choices, or from coding standards.

Threats We knowledge potential threats. *Internal validity* threats might arise from a few sources. First, our transforms must be sound.

We carefully reviewed the code of and hand-checked a large sample of the results to ensure correctness and diminish the possibility of error. Second, while we primarily used lexical ngram based models, we validated the robustness of our corpus study with LSTM models on several examples. Moreover, for the small-scope localized transforms we use, we believe they are adequate, as such models have been used in prior work of this nature [14].

Regarding *external validity*, one issue is our choice of projects. We have chosen a reasonable sample of projects in two widely-used languages, and our results largely hold up. We believe it’s likely that our results will generalize to languages similar to Java and Python. It’s possible that other languages (e.g. Haskell) with tightly-knit, highly-skilled user-bases may behave differently. We have also only focused on small-scale transformations to expressions. It’s possible that that larger transforms may have different effects, and may require different modeling techniques as mentioned above. Finally, coding style guides can influence how code is written. We searched our projects for references to style guides and found several variants. Some projects had explicit style checks, and others were much looser²⁹. Virtually all the guidelines were largely unrelated to our transformations, and more focused on naming & whitespace. We did find a few, limited references in Java to using parentheses as needed for clarity; and just one project specified that *null* values should come second, but otherwise nothing that would affect our transforms. *Construct validity*: we measure prevalence using surprisal from language models. Language models are highly-refined methods for estimating occurrence frequency, and have proven value in natural language processing, so this is well-justified.

Actionability and Future Work: We acknowledge that our work thus far is more science than engineering, but it does have practical implications. While sound, meaning-preserving transforms aren’t realistic for natural language, they are for code! We confirm that surprisal of code is strongly associated with human preference, thus providing theoretical support for a tool that aims to rewrite code into a meaning-preserving form preferred by programmers. We plan to see if better neural models, such as the Transformer [66], align with human preference more strongly. Psycholinguistic experiments have demonstrated that natural language comprehension and production are robustly sensitive to surprisal across a wide range of measures (e.g. forced-choice preferences, reading times, production times, comprehension accuracy, neural measures, etc.) [40, 46, 50, 58, 68]. In future work we plan to extend the human subjects work here to test whether ease of code comprehension and production similarly relates to surprisal.

6 RELATED WORKS

First, we briefly note the recent work by Rahman *et al.* claiming the greater repetitiveness of source code over English is diminished once syntactic tokens are removed [55]. They compared a fixed baseline of English without syntactic markers against code with and without these tokens. However, Casalnuovo *et al.* performed *two* pairwise English-Java corpus comparisons, with and without syntactic (closed-category) tokens. This more balanced comparison reveals that *without the markers of syntax*, the gap of predictability between code and English actually *increases* [14].

²⁹For instance, Apache Tomcat.

6.1 Program Understanding

While we draw theoretical inspiration from the “bimodality” of software [3, 7], we also note that this theory is only a recent reformulation of a much older idea. The idea of programs serving a dual purpose, for machine and human, is decades old [9]. In the study of program understanding, this connects to the ideas of *top-down* and *bottom-up* comprehension. *Top-down comprehension* arguably relates the human-human channel, where past experience guides a reader to seek out expected cues called *beacons*, that help her decipher the program’s meaning [9, 10]. In contrast, *bottom-up comprehension* involves processing individual pieces of code, storing them in memory as *semantic chunks* and constructing the meaning out of these pieces [53, 61]. This in some ways resembles the way a machine would process code, where understanding arises out of precise operational semantics.

Program understanding using fMRI³⁰ and eye-tracking as humans read programs has seen recent focus. A study by Seigmund *et al.* used fMRI to study both top-down and bottom-up comprehension in a programming environment [62]. Using brain activation to measure “neural efficiency” (which associates lower brain activation with greater cognitive ease [47, 62]), they find that top-down comprehension is more efficient than bottom-up comprehension. This supports the theory that the availability of highly probable “beacons” expected by humans facilitates code reading on approaches that rely on bottom-up construction of semantics.

Meanwhile, eye-tracking studies (see survey by Obaidallah *et al.* [49]) also help explain how humans understand code and how they do so *differently* from natural language. For instance, studies show that while natural language follows a linear reading pattern (left to right in English), code readers jump around quite a bit, *e.g.* from a variable or function use to its declaration [11, 37]. Indeed, expert readers tend to show more non-linear eye traces than novices [37]. These techniques have also been used together: Fritz *et al.* used eye-tracking in combination with EEG-based measures of electrical brain activity, to predict the difficulty of programming tasks [26], and recent calls for similar studies combining eye and brain methods highlight their potential for understanding program comprehension [52]. Fakhoury *et al.* used fNIRS (similar to fMRI) and eye tracking [23] to relate cognitive load to lexical but not structural anti-patterns in code, though both led to worse task performance. Finally, in natural language, surprisal is known to relate to eye movement and comprehension [21, 25, 29, 41, 42]; our work suggests that lower surprisal in code will also ease reading & comprehension, which we hope to pursue in future work.

Finally, we note that Gopstein *et al.* [27, 28] found that style guides advocating for using only necessary curly braces aren’t empirically well founded; sometimes superfluous braces aid program understanding. This is consistent with our finding that parentheses are preferentially included to indicate evaluation order (even when not needed) to serve a similar role in segmenting expressions as curly braces do in control flow. Although the predictability of source code and human understanding are different metrics, our results on parentheses removal suggest a similar phenomenon - that developers use them sometimes to benefit from easier readability.

6.2 Generated vs. Stored Language

As we focused on repeated patterns used to express the same meaning, we highlight work on natural language *exemplar* theories. These *exemplars* are examples (at the word or phrase level) that are learned from usage and then generalized [12, 30]. Importantly, some phrases are neither stored entirely in memory nor strictly generated from grammar, forming a category in between the two. Recent work by Morgan *et al.* examined this on order preferences in binomial expressions, such as *bread and butter vs butter and bread*, to see how frequency of usage and abstract linguistic preferences³¹ determined what humans prefer. They found that both effects played roles in preference, but frequency of expression overwhelms the effects of underlying preferences and codifies a norm [46]. We use similar human subject study, but leverage code’s structure to combine this with a natural experiment on code corpora.

We also briefly note a specific type of stored language, *idioms*, as they have been explored in software before. Idioms can quickly and efficiently convey a meaning for to those who know them [59]. Idiomatic language has been mined from source code by Allamanis *et al.*, finding syntactic fragments across programs that possess the same meaning [4], though they focus on extracting them over studying controlled preferences as we do.

Finally, we briefly note recent work in approaches to generating program variants in mutation testing and program obfuscation. Mutation testing seeks to create semantically *different* programs to expose deficits in test suites [51]. One relevant work has tried using language models to find *natural* mutants, finding that the mutants tended to be less natural (more improbable) but did not have success using the metric to guide mutation selection [38]. Additionally, obfuscation transformations generally retains meaning, though they can produce different error behavior and run much slower [17, 18]. Recent work [43] has used naturalness to combine obfuscation operators in a way to minimize the effectiveness of deobfuscation techniques learning patterns out of software [57, 65]. These approaches, however, are more focused on applications than understanding the decisions that inform source code choices.

7 CONCLUSION

Why is code so repetitive? Previous work strongly suggests that it is not merely the restricted syntax of programming languages; and it’s most definitely not because programming languages restrict the possible ways to express computations. In this study, we hypothesize that programmers prefer certain ways to write code. We model familiarity using a language model estimated over a large corpus, and measure the “familiarity” of different ways writing code, while controlling for the meaning. We find that “*familiar*” forms are indeed more preferred by code writers, using surprisal, as scored by a language model, and also align with preferences of human readers via controlled experiment on Mechanical Turk. Finally, we draw connections between our work and the well-established theories from Psycholinguistics.

³⁰Functional Magnetic Resonance Imaging.

³¹As an analogy, an abstract linguistic preference in code might be that variables go before constants, *e.g.*, $i+1$ rather than $1+i$.

REFERENCES

- [1] Miltiadis Allamanis. 2018. The Adverse Effects of Code Duplication in Machine Learning Models of Code. *CoRR* abs/1812.06469 (2018). arXiv:1812.06469 <http://arxiv.org/abs/1812.06469>
- [2] Miltiadis Allamanis, Earl T. Barr, Christian Bird, and Charles Sutton. 2014. Learning Natural Coding Conventions. In *Proceedings of the 22Nd ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2014)*. ACM, New York, NY, USA, 281–293. <https://doi.org/10.1145/2635868.2635883>
- [3] Miltiadis Allamanis, Earl T Barr, Premkumar Devanbu, and Charles Sutton. 2018. A survey of machine learning for big code and naturalness. *Comput. Surveys* (2018).
- [4] Miltiadis Allamanis and Charles Sutton. 2014. Mining Idioms from Source Code. In *Proceedings of the 22Nd ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2014)*. ACM, New York, NY, USA, 472–483. <https://doi.org/10.1145/2635868.2635901>
- [5] Anwar Alqaimi, Patanamon Thongtanunam, and Christoph Treude. 2019. Automatically Generating Documentation for Lambda Expressions in Java. In *Proceedings of the 16th International Conference on Mining Software Repositories (MSR '19)*. IEEE Press, Piscataway, NJ, USA, 310–320. <https://doi.org/10.1109/MSR.2019.00057>
- [6] Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68, 3 (2013), 255 – 278. <https://doi.org/10.1016/j.jml.2012.11.001>
- [7] Earl Barr. 2018. Bimodal Software Engineering. <https://www.youtube.com/watch?v=UJDFmuUMkVE> Machine Learning for Programming Workshop, Federated Logic Conference (FLOC) 2018.
- [8] Kathryn Bock. 1987. An effect of the accessibility of word forms on sentence structures. *Journal of memory and language* 26, 2 (1987), 119–137.
- [9] Ruven Brooks. 1978. Using a behavioral theory of program comprehension in software engineering. In *Proceedings of the 3rd international conference on Software engineering*. IEEE Press, 196–201.
- [10] Ruven Brooks. 1983. Towards a theory of the comprehension of computer programs. *International Journal of Man-Machine Studies* 18, 6 (1983), 543 – 554. [https://doi.org/10.1016/S0020-7373\(83\)80031-5](https://doi.org/10.1016/S0020-7373(83)80031-5)
- [11] Teresa Busjahn, Roman Bednarik, Andrew Begel, Martha Crosby, James H Paterson, Carsten Schulte, Bonita Sharif, and Sascha Tamm. 2015. Eye movements in code reading: Relaxing the linear order. In *Program Comprehension (ICPC), 2015 IEEE 23rd International Conference on*. IEEE, 255–265.
- [12] Joan Bybee. 2003. *Phonology and language use*. Vol. 94. Cambridge University Press.
- [13] Paul-Christian Bürkner. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80, 1 (2017), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- [14] Casey Casalnuovo, Kenji Sagae, and Prem Devanbu. 2018. Studying the difference between natural and programming language corpora. *Empirical Software Engineering* (2018), 1–46.
- [15] Di Chen, Kathryn T. Stolee, and Tim Menzies. 2019. Replication Can Improve Prior Results: A GitHub Study of Pull Request Acceptance. In *Proceedings of the 27th International Conference on Program Comprehension (ICPC '19)*. IEEE Press, Piscataway, NJ, USA, 179–190. <https://doi.org/10.1109/ICPC.2019.00037>
- [16] Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. 2003. Applied multiple correlation/regression analysis for the behavioral sciences. *UK: Taylor & Francis* (2003).
- [17] Christian Collberg, Clark Thomborson, and Douglas Low. 1997. *A taxonomy of obfuscating transformations*. Technical Report. Department of Computer Science, The University of Auckland, New Zealand.
- [18] Christian Collberg, Clark Thomborson, and Douglas Low. 1998. Manufacturing cheap, resilient, and stealthy opaque constructs. In *Proceedings of the 25th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. ACM, 184–196.
- [19] R Dennis Cook and Sanford Weisberg. 1982. *Residuals and influence in regression*. New York: Chapman and Hall.
- [20] Mturk Data. 2018. The BOT problem on Mturk. <http://turkrequesters.blogspot.com/2018/08/the-bot-problem-on-mturk.html>. Accessed: May 2019.
- [21] Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109, 2 (2008), 193 – 210. <https://doi.org/10.1016/j.cognition.2008.07.008>
- [22] Sarah Depaoli and Rens Van de Schoot. 2017. Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological methods* 22, 2 (2017), 240.
- [23] Sarah Fakhoury, Devjeet Roy, Yuzhan Ma, Venera Arnaoudova, and Olusola Adesope. 2019. Measuring the impact of lexical and structural inconsistencies on developers’ cognitive load during bug localization. *Empirical Software Engineering* (08 Aug 2019). <https://doi.org/10.1007/s10664-019-09751-4>
- [24] Victor S Ferreira and Gary S Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive psychology* 40, 4 (2000), 296–340.
- [25] Stefan Frank. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science* 5, 3 (2013), 475–494.
- [26] Thomas Fritz, Andrew Begel, Sebastian C Müller, Serap Yigit-Elliott, and Manuela Züger. 2014. Using psycho-physiological measures to assess task difficulty in software development. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 402–413.
- [27] Dan Gopstein, Jake Iannacone, Yu Yan, Lois DeLong, Yanyan Zhuang, Martin K-C Yeh, and Justin Cappos. 2017. Understanding misunderstandings in source code. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ACM, 129–139.
- [28] Dan Gopstein, Hongwei Henry Zhou, Phyllis Frankl, and Justin Cappos. 2018. Prevalence of Confusing Code in Software Projects: Atoms of Confusion in the Wild. In *MSR '18: 15th International Conference on Mining Software Repositories, May 28–29, 2018, Gothenburg, Sweden*. ACM, 11 pages. <https://doi.org/10.1145/3196398.3196432>
- [29] John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, 1–8.
- [30] Jennifer Hay and Joan Bresnan. 2006. Spoken syntax: The phonetics of giving a hand in New Zealand English. *The Linguistic Review* 23, 3 (2006), 321–349.
- [31] Vincent J Hellendoorn and Premkumar Devanbu. 2017. Are deep neural networks the best choice for modeling source code?. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, ser. ESEC/FSE*. 763–773.
- [32] Abram Hindle, Earl T. Barr, Zhendong Su, Mark Gabel, and Premkumar Devanbu. 2012. On the Naturalness of Software. In *Proceedings of the 34th International Conference on Software Engineering (ICSE '12)*. IEEE Press, Piscataway, NJ, USA, 837–847. <http://dl.acm.org/citation.cfm?id=2337223.2337322>
- [33] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [34] Myles Hollander and Douglas A Wolfe. 1999. *Nonparametric statistical methods*. (1999).
- [35] Carla L Hudson Kam and Elissa L Newport. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language learning and development* 1, 2 (2005), 151–195.
- [36] T Florian Jaeger and Roger P Levy. 2007. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*. 849–856.
- [37] Ahmad Jbara and Dror G Feitelson. 2017. How programmers read regular code: a controlled experiment using eye tracking. *Empirical Software Engineering* 22, 3 (2017), 1440–1477.
- [38] Matthieu Jimenez, Thierry Titchou Checkam, Maxime Cordy, Mike Papadakis, Marinos Kintis, Yves Le Traon, and Mark Harman. 2018. Are mutants really natural?: a study on how naturalness helps mutant selection. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM, 3.
- [39] René Just. 2014. The Major Mutation Framework: Efficient and Scalable Mutation Analysis for Java. In *Proceedings of the 2014 International Symposium on Software Testing and Analysis (ISSTA 2014)*. ACM, New York, NY, USA, 433–436. <https://doi.org/10.1145/2610384.2628053>
- [40] Marta Kutas and Steven A Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 5947 (1984), 161.
- [41] Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition* 106, 3 (2008), 1126 – 1177. <https://doi.org/10.1016/j.cognition.2007.05.006>
- [42] Roger Levy. 2013. Memory and surprisal in human sentence comprehension. In *Sentence processing*. Psychology Press, 90–126.
- [43] Han Liu, Chengnian Sun, Zhendong Su, Yu Jiang, Ming Gu, and Jianguang Sun. 2017. Stochastic optimization of program obfuscation. In *Software Engineering (ICSE), 2017 IEEE/ACM 39th International Conference on*. IEEE, 221–231.
- [44] Lech Madeyski and Norbert Radyk. 2010. Judy—a mutation testing tool for Java. *IET software* 4, 1 (2010), 32–42.
- [45] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Comput. Linguist.* 19, 2 (June 1993), 313–330. <http://dl.acm.org/citation.cfm?id=972470.972475>
- [46] Emily Morgan and Roger Levy. 2016. Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition* 157 (2016), 384–402.
- [47] Aljoscha C Neubauer and Andreas Fink. 2009. Intelligence and neural efficiency. *Neuroscience & Biobehavioral Reviews* 33, 7 (2009), 1004–1023.
- [48] Anh Tuan Nguyen and Tien N. Nguyen. 2015. Graph-based Statistical Language Model for Code. In *Proceedings of the 37th International Conference on Software Engineering - Volume 1 (ICSE '15)*. IEEE Press, Piscataway, NJ, USA, 858–868. <http://dl.acm.org/citation.cfm?id=2818754.2818858>
- [49] Unaizah Obaidallah, Mohammed Al Haek, and Peter C-H Cheng. 2018. A survey on the usage of eye-tracking in computer programming. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 5.
- [50] Richard C Oldfield and Arthur Wingfield. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17, 4 (1965), 273–281.

- [51] Mike Papadakis, Marinos Kintis, Jie Zhang, Yue Jia, Yves Le Traon, and Mark Harman. 2019. Mutation testing advances: an analysis and survey. In *Advances in Computers*. Vol. 112. Elsevier, 275–378.
- [52] Norman Peitek, Janet Siegmund, Chris Parnin, Sven Apel, and André Brechmann. 2018. Toward conjoint analysis of simultaneous eye-tracking and fMRI data for program-comprehension studies. In *Proceedings of the Workshop on Eye Movements in Programming*. ACM, 1.
- [53] Nancy Pennington. 1987. Stimulus structures and mental representations in expert comprehension of computer programs. *Cognitive Psychology* 19, 3 (1987), 295–341. [https://doi.org/10.1016/0010-0285\(87\)90007-7](https://doi.org/10.1016/0010-0285(87)90007-7)
- [54] Gede Artha Azriadi Prana, Christoph Treude, Ferdian Thung, Thushari Atapattu, and David Lo. 2019. Categorizing the Content of GitHub README Files. *Empirical Software Engineering* 24, 3 (01 Jun 2019), 1296–1327. <https://doi.org/10.1007/s10664-018-9660-3>
- [55] Musifiquir Rahman, Dharani Palani, and Peter Rigby. 2019. Natural Software Revisited. In *Proceedings, ICSE*.
- [56] Baishakhi Ray, Vincent Hellendoorn, Saheel Godhane, Zhaopeng Tu, Alberto Bacchelli, and Premkumar Devanbu. 2016. On the "Naturalness" of Buggy Code. In *Proceedings of the 38th International Conference on Software Engineering (ICSE '16)*. ACM, New York, NY, USA, 428–439. <https://doi.org/10.1145/2884781.2884848>
- [57] Veselin Raychev, Martin Vechev, and Andreas Krause. 2015. Predicting Program Properties from "Big Code". In *Proceedings of the 42Nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '15)*. ACM, New York, NY, USA, 111–124. <https://doi.org/10.1145/2676726.2677009>
- [58] Keith Rayner and Arnold D Well. 1996. Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review* 3, 4 (1996), 504–509.
- [59] Norbert Schmitt and Ronald Carter. 2004. Formulaic sequences in action. *Formulaic sequences: Acquisition, processing and use* (2004), 1–22.
- [60] Jessica F Schwab, LEW-WILLIAMS Casey, and Adele E Goldberg. 2018. When regularization gets it wrong: Children over-simplify language input only in production. *Journal of child language* 45, 5 (2018), 1054–1072.
- [61] Ben Shneiderman and Richard Mayer. 1979. Syntactic/semantic interactions in programmer behavior: A model and experimental results. *International Journal of Computer & Information Sciences* 8, 3 (1979), 219–238.
- [62] Janet Siegmund, Norman Peitek, Chris Parnin, Sven Apel, Johannes Hofmeister, Christian Kästner, Andrew Begel, Anja Bethmann, and André Brechmann. 2017. Measuring neural efficiency of program comprehension. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ACM, 140–150.
- [63] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*.
- [64] Zhaopeng Tu, Zhendong Su, and Premkumar Devanbu. 2014. On the Localness of Software. In *Proceedings of the 22Nd ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2014)*. ACM, New York, NY, USA, 269–280. <https://doi.org/10.1145/2635868.2635875>
- [65] Bogdan Vasilescu, Casey Casalnuovo, and Premkumar Devanbu. 2017. Recovering clear, natural identifiers from obfuscated JS names. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ACM, 683–693.
- [66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [67] Eric W Weisstein. 2004. Bonferroni correction. (2004).
- [68] Justine B Wells, Morten H Christiansen, David S Race, Daniel J Acheson, and Maryellen C MacDonald. 2009. Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive psychology* 58, 2 (2009), 250–271.