Unzipping the Causality of Zipf's Law and Other Lexical Trade-offs

Amanda Doucette[1] **Timothy J. O'Donnell**[1,2] **Morgan Sonderegger**[1]
[1]Dept. of Linguistics, McGill University [2]Canada CIFAR AI Chair, Mila
amanda.doucette@mail.mcgill.ca timothy.odonnell@mcgill.ca
morgan.sonderegger@mcgill.ca

Abstract

There are strong constraints on the structure of a possible lexicon. For example, the negative correlation between word frequency and length known as Zipf's law of abbreviation, and a negative correlation between word length and phonotactic complexity appear to hold across languages. While lexical trade-offs like these have been examined individually, it is unclear how they interact as a system. In this paper, we propose causal discovery as a method for identifying lexical biases and their interactions in a set of variables. We represent the lexicon as a causal model, and apply the Fast Causal Discovery algorithm (Spirtes et al., 1995) to identify both causal relationships between measured variables and the existence of possible unmeasured confounding variables. We apply this method to lexical data including measures of word length, frequency, phonotactic complexity, and morphological irregularity for 25 languages and find evidence of universal associations involving word length with a high likelihood of involving an unmeasured confounder, suggesting that additional variables need to be measured to determine how they are related. We also find evidence of variation across languages in relationships between the remaining variables, and suggest that given a larger dataset, causal discovery algorithms can be a useful tool in assessing the universality of lexical biases.[1]

1 Introduction

Although lexicons vary significantly across languages, they exhibit striking regularity. For example, it is well documented that the most frequent words in a language tend to be the shortest (Zipf's law of abbreviation;[2] Zipf, 1935; Piantadosi, 2014), and that frequent words tend to be phonotactically simple (Mahowald et al., 2018). Some of these trade-offs appear to be linguistic universals while others, such as the relationship between frequency and morphological irregularity, are more debated and appear to display more variation across languages (Fratini et al., 2014; Yang, 2016; Wu et al., 2019; Doucette et al., 2024). Each of these trade-offs imposes limitations on the set of possible natural language lexicons. Although it is possible to construct a lexicon where, for example, the most frequent words are the longest, no human language follows this pattern.

Many of these trade-offs have been attributed to universal cognitive pressures. For example, a pressure for efficient communication may explain why frequent words tend to be short and phonotactically simple (Zipf, 1935; Mahowald et al., 2018; Piantadosi et al., 2011; Graff, 2012; Gibson et al., 2019; Levshina, 2022). Hay and Baayen (2003) attribute a tendency for frequent words to be morphologically irregular to a constraint on processing – it is more efficient to access these frequent irregulars as whole words, rather than parse them into component morphemes. However, a pressure for efficient communication could also imply the opposite pattern: It is more memory-efficient to store component morphemes in the lexicon, so irregulars should be *infrequent*. Another trade-off, a negative correlation between word length and phonotactic complexity demonstrated by Pimentel et al. (2020), has been attributed to a pressure towards uniform information density: A consistent rate of information requires shorter words to be more complex (Pellegrino et al., 2011; Coupé et al., 2019; Meister et al., 2021). However, Doucette et al. (2024) showed that this correlation becomes positive when only morphologically complex words are examined. Such complex and potentially contradictory results suggest a network of interacting pressures influencing the shape of a lexicon. A cognitive pressure that neatly explains one trade-off may be

[1]Code is available at https://osf.io/g8b35.

[2]Not the Zipf's law from Zipf (1949), which states that a word's frequency is inversely proportional to its frequency rank.

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 66–76

contradicted by another. Because we are examining interactions among sets of variables, we will refer to the limitations on possible lexicons imposed by these trade-offs as *lexical biases*, independent of the *cognitive constraints* that may cause them. In order to study the cognitive constraints shaping the lexicon, we cannot only consider data representing a single trade-off in the lexicon. Instead, we need an understanding of what lexical biases exist, how they interact with each other, and whether and how they vary across the world's languages.

In this paper, we propose a method for identifying lexical biases and their interactions: causal discovery. Much of the previously described work on lexical biases implicitly suggests a causal relationship – that there is some process where words that become more frequent are shortened over time, for example. It is also possible that this type of direct causal process does not exist, and instead word length and frequency share some common cause – a *confounder*. The structural causal modeling framework introduced by Pearl (1995) is useful in assessing these types of causal structures. A causal model includes a set of random variables and the causal relationships between them, represented by a graph. We can represent the situation where a change in word length (WL) causes a change in frequency (FR) as $WL \longrightarrow FR$, and the situation where both share an unknown common cause (U) as $WL \longleftarrow U \longrightarrow FR$. These graphs represent data generating processes. In the first graph, a word length is sampled, then its frequency is determined based on that value. In the second graph, we sample a value of U, which determines the values of WL and FR. Causal discovery allows us to identify causal graphs consistent with a sample of observational data. Identifying a causal model of the lexicon through causal discovery allows us to examine the networks of lexical biases across languages and ultimately identify the cognitive constraints that underlie them.

Although many questions about language involve causality, causal analyses have only been applied to linguistic data in a few cases. For example, in identifying the causes of lenition (Priva, 2017; Priva and Gleason, 2020), examining causality in child language acquisition (Irvin et al., 2016; Spokoyny et al., 2016), in language change (del Prado Martín, 2014; Moscoso del Prado Martín and Brendel, 2016; Dellert, 2019, 2024), and in examining cross-linguistic trade-offs between case marking and word order (Levshina, 2021). Causal models have not yet been used to investigate lexical biases, which we do using causal discovery.

We apply this method to data described in Doucette et al. (2024): measures of word length, frequency, phonotactic complexity, and morphological irregularity in 25 languages. Through causal discovery, we are able to identify the well-known associations between word length and frequency, and word length and phonotactic complexity, as well as the association between word length and morphological irregularity identified by Doucette et al. (2024). However, we are also able to identify possible unmeasured confounding variables in each of these relationships, suggesting that the direct causal relationship implied by previous studies may not exist. Furthermore, we find evidence of variation in relationships between the remaining pairs of variables: an association only exists in approximately half of the languages in the sample, and where it does exist there is the possibility of confounding. These results demonstrate that in order to determine the causal structure of lexical biases, a larger set of variables need to be considered. Causal discovery allows us to both identify relationships between aspects of the lexicon and determine where more data is needed to make conclusions about causal structure.

2 Data

In this paper, we examine data from Doucette et al. (2024), which was used to study compensation relationships between word length, frequency, morphological irregularity, and phonotactic complexity. It contains 25 languages selected from UniMorph, a database of morphologically annotated corpora (Batsuren et al., 2022), with between 334 and 96,196 word forms per language (median 8,061), converted to IPA transcriptions using Epitran (Mortensen et al., 2018). We note that this data does not represent a random sample from each lexicon: UniMorph largely consists of words with multiple morphemes, with few monomorphemic words. We return to this point in the discussion.

In this data, word length is measured in number of phones, and frequency is calculated from Wikipedia as log count per million. The phonotactic complexity measure, defined by Pimentel et al. (2020), comes from a neural network model trained to estimate the probability of a word w given the rest of the language $\mathcal{L}$. Phonotactic complexity is a measure of bits per phoneme:

Edge type	Interpretation
X $\longrightarrow$ Y	X causes Y, Y does not cause X
X $\circ\!\!\rightarrow$ Y	either X causes Y, or an unobserved confounder causes both X and Y, but not both
X $\longleftrightarrow$ Y	an unobserved confounder causes both X and Y
X $\circ\!\!-\!\!\circ$ Y	One of the following holds: 1. X causes Y; 2. Y causes X; 3. an unobserved confounder causes X and Y; 4. both 1 and 3 hold; 5. both 2 and 3 hold
X - - - Y	no association between X and Y

Table 1: Partial Ancestral Graph (PAG) edge types and their interpretations.

$\log p(w \mid \mathcal{L})/|w|$. The morphological irregularity measure, from Wu et al. (2019), is a neural estimate of the predictability of the surface form of an inflected word from its lemma. A neural network is trained to predict an inflected form from a lemma ℓ, a set of morphological features σ, and the rest of the language with the target lemma removed $\mathcal{L}_{-\ell}$, and the morphological irregularity measure is $\log\left(p(w \mid \ell, \sigma, \mathcal{L}_{-\ell}) / \left[1 - p(w \mid \ell, \sigma, \mathcal{L}_{-\ell})\right]\right)$.

3 Causal Graphs and Causal Discovery

In Pearl's (1995) structural causal modeling framework, a causal model is represented by a directed acyclic graph (DAG), $G = (\mathbf{V}, \mathbf{E})$, a tuple with a finite set of vertices $\mathbf{V}$ representing random variables, and a finite set of edges $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ representing causal relationships. An edge $X \rightarrow Y$ implies that X directly causes Y. The value of any variable in a causal graph is completely determined by a function of its parents. In a DAG $X \rightarrow Y \leftarrow Z$, Y is caused by X and Z, and $Y = f(X, Z)$. There is a large literature on causal modeling, which we only partially and briefly summarize here. For more background, see Pearl et al. (2016) or Hernán and Robins (2024).

Causal discovery algorithms aim to recover the true graph G from a sample dataset. To do this, several assumptions about the relationship between a graph G and the joint probability distribution over its random variables $P(\mathbf{V})$ are needed. The graph G must satisfy the *Causal Markov Condition*: every variable is conditionally independent of its nondescendents given its parents. The probability distribution associated with G must decompose recursively as $P(\mathbf{V}) = \prod_{X \in \mathbf{V}} P(X \mid Pa(X))$, where $Pa(X)$ is the parents of X in G. We also need to assume *faithfulness*: that all independence relationships among the variables $\mathbf{V}$ are represented in G. In other words, the conditional independence relationships in the data are closely tied to the structure of the DAG G. Causal discovery algorithms identify conditional independencies in data, then construct a graph consistent with them.

There are many causal discovery algorithms, each with different assumptions (see Zanga et al. (2022) or Malinsky and Danks (2018) for a summary). Many assume *causal sufficiency*: that any cause of a variable in $\mathbf{V}$ is also contained in $\mathbf{V}$. If there are any unmeasured common causes in the data, an algorithm assuming causal sufficiency will not output a correct causal graph. For example, if we are trying to discover a true DAG $X \longrightarrow U \longleftarrow Y$, but do not have measurements of U, an algorithm assuming causal sufficiency will fail to recover the correct DAG. The unmeasured variable U is a confounder that can induce a spurious correlation between X and Y even though there is no causal relationship between X and Y.

It is likely that most lexical data is not causally sufficient: there are likely to be additional causes outside of the set of variables included in the data. Therefore, we use an algorithm that does not assume sufficiency: the Fast Causal Inference (FCI) algorithm (Spirtes et al., 1993, 1995), which takes a set of observations of random variables as input and ouputs a Partial Ancestral Graph (PAG), a causal graph with additional edge types in order to represent unmeasured confounders and uncertainty. In a PAG, directed edges $\longrightarrow$ and $\longleftarrow$ have the same meaning as in a DAG: they represent a direct causal relationship. A PAG represents the presence of an unmeasured confounder with a bidirected edge, $\longleftrightarrow$. For example, an edge $X \longleftrightarrow Y$ means there is some unmeasured variable that causes both X and Y, and that there is no direct causal relationship between X and Y. PAGs also add circle endmarks to edges, representing uncertainty. For example, $X \circ\!\!-\!\!\circ Y$ corresponds to one of several possibilities: X causes Y, Y causes X, there is an unmeasured common cause of both X and Y, or there is both an unmeasured common cause and a direct causal relationship (i.e. $X \longrightarrow Y$ *and* $X \longleftrightarrow Y$). The interpretation of all possible edges in a PAG are listed in Table 1.

The FCI algorithm starts with a complete undirected graph, where all random variables are con-

nected by undirected edges. Next, a series of conditional independence tests are conducted. An undirected edge A — B is removed if A and B are conditionally independent given some set of variables $\mathbf{C}$. The resulting graph after no more edges can be removed is called a *skeleton*. The unoriented edges in a graph skeleton do not have a causal interpretation, but can be useful for examining statistical *associations* between the random variables. In the next step of the FCI algorithm, all edges in the skeleton begin as unoriented ○–○ edges. Edges are then oriented following a series of rules based on graph structure.

We use an implementation of the FCI algorithm and an implementation of the Fisher's Z conditional independence test from the R package `pcalg` (Kalisch et al., 2012). This conditional independence test assumes a Gaussian distribution, and requires a significance level. This assumption may not be reasonable, and we will return to it in the discussion. For each language in the dataset, we used a bootstrapping procedure to resample the data 1000 times. The FCI algorithm was run on each sample with a significance level of 0.01, and the proportion of edge types discovered for each pair of variables was recorded.

4 Results

For each language, the most frequently occurring edge type for each pair of variables in the bootstrap samples was selected to create a "most-likely graph" for that language. These are shown in Figure 1, where we can see that there is significant variation in the most-likely graphs discovered for each language – there are 18 unique graphs identified across 25 languages. At most three languages share the same graph. One of these groups, containing Chewa and Zulu, can be explained by typological relatedness, but the others have no clear explanation. The variation in these most-likely graphs may suggest that there is no universal set of lexical biases shared across languages. We return to this point in the discussion.

We also see that the graphs for many languages contain edges with circle marks (←○ , ○→ , ○–○), which suggests that there is not enough information in the dataset to fully determine causal relationships. It is likely that there are unmeasured confounding variables. If we instead examine the graph skeletons discovered by FCI, we can examine associations between variables. In these undirected graphs, the presence of an edge implies an association between variables – a correlation that is not necessarily causal. In Figure 1, groups of languages sharing the same skeleton are outlined. When considering the graph skeletons, we see larger groups of languages emerge, suggesting less variation in lexical biases across languages. It is unclear why certain languages share the same skeleton structure. Many of the groups in Figure 1 are not typologically related, such as Polish, Dutch, Czech, French, and Ukrainian. A larger set of languages is needed to determine if there is any typological explanation behind these groupings.

To further examine these individual language graphs, we created a cross-linguistic most-likely graph by selecting the most frequently occurring edge type for each pair of variables from the graphs in Figure 1. This is shown in Figure 2. A most-likely skeleton, shown in Figure 3 was created by following the same procedure with the graph skeletons. To examine the distribution of edge types in the bootstrap sampling procedure, we plotted histograms for each pair of variables showing the proportion of bootstrap samples where types of edges were found across all languages. A right-skewed histogram implies that an edge was discovered in most languages, while a left-skewed histogram implies that no edge was discovered in most languages. Figure 4A shows the proportion of samples where an edge of any type was discovered (⟶ , ⟵ , ⟷ , ○→ , ←○ , ○–○), Figure 4B shows the proportion where a directed edge was discovered (⟶ or ⟵), and Figure 4C shows the proportion of edges discovered with confounding variables (⟷) or with potential confounding variables (○–○ , ←○ , or ○→).

4.1 Word Length and Frequency

Due to Zipf's law of abbreviation (Zipf, 1935), where word length and frequency are negatively correlated, we expect to find an association between word length and frequency. This is what we find: in Figure 3, we see that the most likely skeletons for all languages have an edge between these variables. In Figure 4, we also see that an edge was discovered in nearly all bootstrap samples for nearly all languages. However, directed edges (⟶ or ⟵) were not found in nearly all samples, as shown in Figure 4. Instead, as can be seen in Figure 4, nearly all bootstrap samples indicate the possible presence of an unmeasured confounding variable. In Figure 2, a ○–○ edge is most likely between word

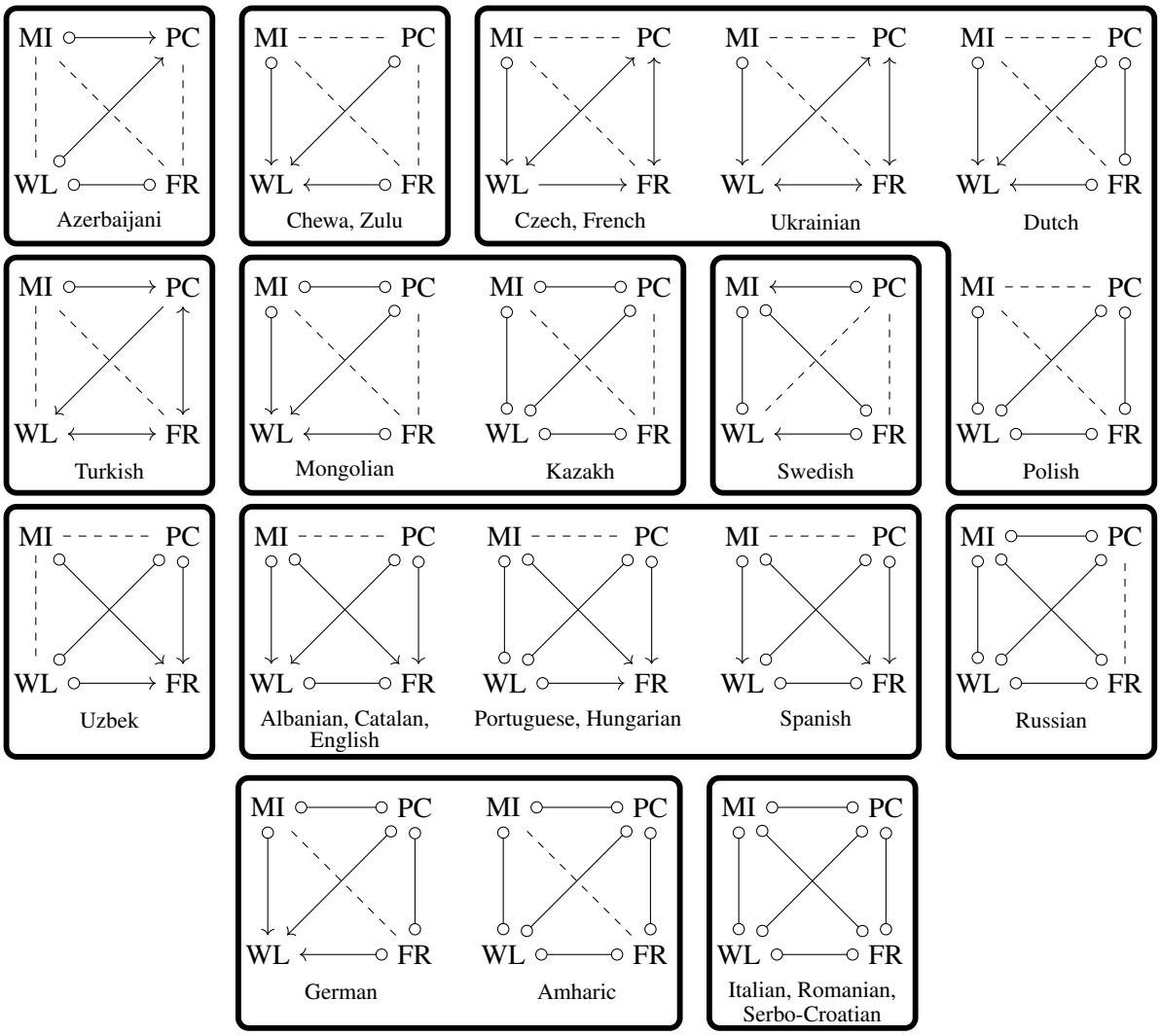

Figure 1: Most likely PAGs for individual languages. Languages with the same undirected graph outlined in black. MI: morphological irregularity; PC: phonotactic complexity; WL: word length; FR: frequency.

length and frequency, implying an association with unknown causal direction and possible confounding. This suggests that we are not able to infer the causality of this relationship from the variables in this dataset.

This is consistent with previous work where it has been argued that average surprisal, or predictability in context, correlates more strongly with word length than frequency does (Piantadosi et al., 2011), although other analyses have failed to reproduce this finding (Meylan and Griffiths, 2021; Pimentel et al., 2023). This dataset does not include a measure of average surprisal, but it is possible that surprisal is the unmeasured confounding variable, and causal discovery could help determine the relationship between frequency, word length, and surprisal given a dataset with measurements of average surprisal. Because word length correlates with surprisal, surprisal may also be a confounder in its relationships with morphological irregularity and phonotactic complexity. In order to make conclusions about causal relationships involving word length and frequency, more variables need to be measured than those included here.

4.2 Word Length and Phonotactic Complexity

In Figure 3, we also see that an edge was discovered between word length and phonotactic complexity in all languages, as predicted by Pimentel et al.'s (2020) finding that these variables are negatively correlated. The most frequently occurring edge type between word length and phonotactic complexity is $\circ\!\!-\!\!\circ$, implying that either word length is constrained by phonotactic complexity (PC $\longrightarrow$ WL), phonotactic complexity is constrained by word length (PC $\longleftarrow$ WL) or there is an unmeasured confounding variable (PC $\longleftrightarrow$ WL). However, despite being the most frequent edge type between

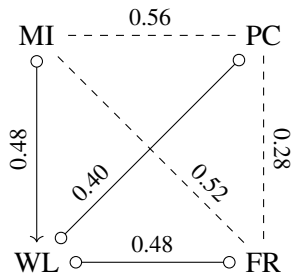

Figure 2: Most likely PAG across all languages, with proportion of languages where edge type occurred. MI: morphological irregularity; PC: phonotactic complexity; WL: word length; FR: frequency.

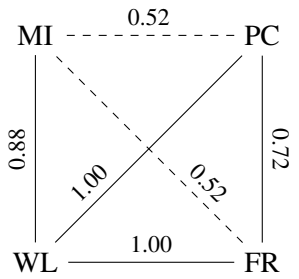

Figure 3: Most likely undirected graph skeleton across all languages, with proportion of languages where edge type occurred. MI: morphological irregularity; PC: phonotactic complexity; WL: word length; FR: frequency.

these variables, it occurs in less than half of languages, suggesting variation in the causal structure of this relationship. This can be seen in Figure 1, where there is little consistency in the type of edge between word length and phonotactic complexity. Figure 4 shows that a directed edge was found in very few samples, and that a possible confounder was identified in nearly all samples. Again, this shows that while an association between word length and phonotactic complexity may be universal, we do not have enough data to determine their causal relationship.

4.3 Word Length and Morphological Irregularity

Figure 3 and Figure 4 show that an edge was also discovered between word length and morphological irregularity in most languages, consistent with Doucette et al.'s (2024) finding of a negative correlation. Again, directed edges were discovered in very few samples and a possible confounder was discovered in most samples, as shown in Figure 4. In Figure 2, the most common edge type between word length and morphological irregularity is $\leftarrow\!\circ$, discovered in 12 of 25 languages. Like the relationship between word length and phonotactic complexity, an association between these variables is near-universal, but it is likely confounded by a variable outside of this dataset.

4.4 Frequency and Phonotactic Complexity

The remaining three pairs of variables display less consistency in whether or not an edge is present. For frequency and phonotactic complexity, the most likely scenario is that an edge does exist, shown in Figure 3. However, the most likely PAG shows no edge between these variables. Although the presence of an edge between these variables is slightly more likely than not, the type of edge varies across languages, as shown in Figure 1. Mahowald et al. (2018) found that a negative correlation between phonotactic complexity and frequency was robust after controlling for the effect of word length as a confounding factor. The FCI algorithm is able to identify confounding relationships, but this relationship does not appear as robustly as previously found, even in Doucette et al.'s (2024) analysis of the same data. This suggests that it may not be correct to conclude that there is an association between frequency and phonotactic complexity while only considering word length as a confounder. Once another factor like morphological irregularity is included, the relationship becomes less clear.

4.5 Frequency and Morphological Irregularity

While Wu et al. (2019) found a positive correlation between morphological irregularity and frequency, we found no association in about half of the 25 languages, as shown in both the most-likely skeleton in Figure 3. and the most-likely PAG in Figure 2. In Figure 4, we can see that for some languages, this edge occurred in almost no bootstrap samples, while for other languages, it occurred in nearly all samples. Very few languages fall in the center of the histogram, with the edge being discovered in some samples, but not others. This U-shaped histogram suggests that the existence of an association between morphological irregularity and frequency may be a point of variation across languages – some definitively have an association, while others do not.

The histogram for possible confounders in Figure 4 is similarly U-shaped. This suggests if an edge between frequency and morphological irreg-

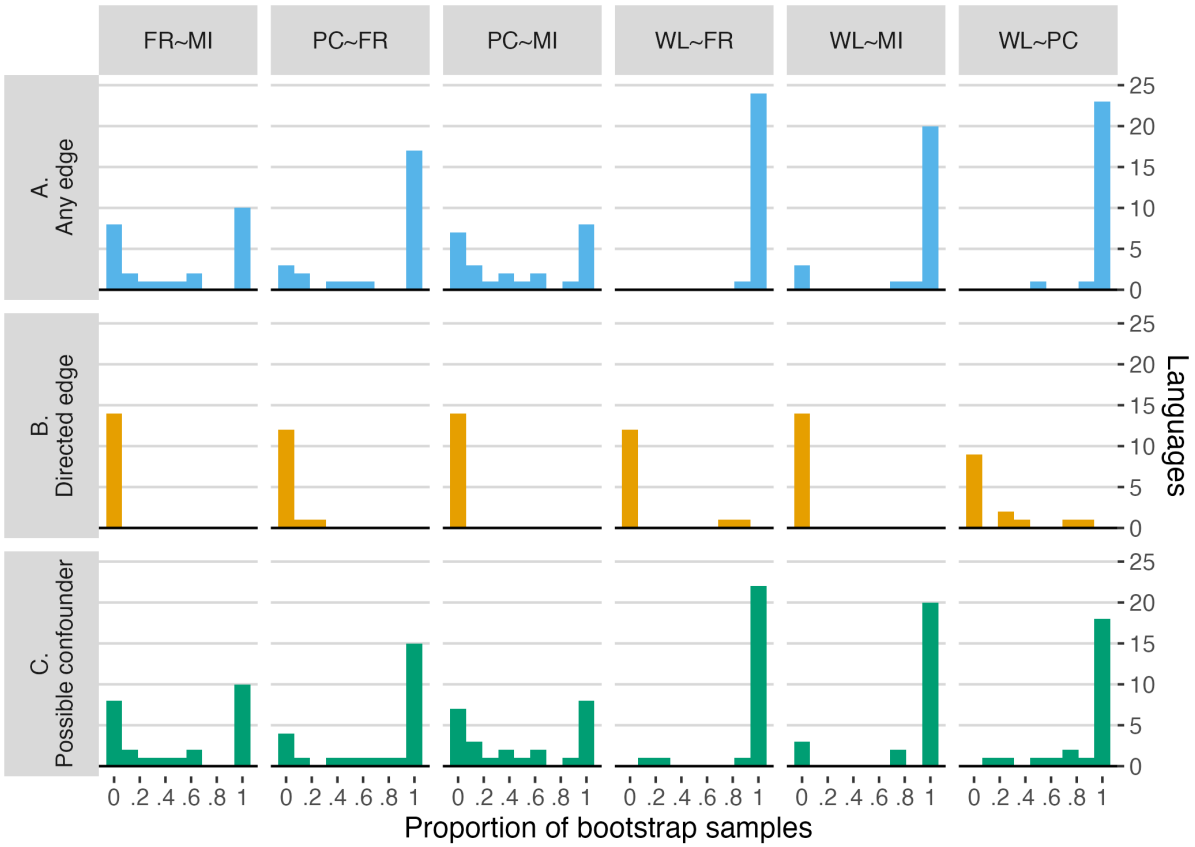

Figure 4: Histograms of proportion of bootstrap samples including edge types for each pair of variables. A: $\longrightarrow$ $\longleftarrow$, $\longleftrightarrow$, $\circ\!\!-\!\!\circ$, $\circ\!\!\rightarrow$, or $\leftarrow\!\!\circ$; B: $\longrightarrow$ or $\longleftarrow$; C: $\longleftrightarrow$, $\circ\!\!-\!\!\circ$, $\circ\!\!\rightarrow$, or $\leftarrow\!\!\circ$. MI: morphological irregularity; PC: phonotactic complexity; WL: word length; FR: frequency.

ularity exists for a language, it is likely to have a possible confounding variable. Like previously discussed pairs of variables, directed edges are rarely identified between frequency and morphological irregularity.

4.6 Morphological Irregularity and Phonotactic Complexity

We see a similar pattern in the histograms in Figure 4 for phonotactic complexity and morphological irregularity: Directed edges almost never occur, and the edge existence and possible confounder histograms are U-shaped, suggesting that languages vary in whether or not there is an associatino between these variables. This is consistent with previous findings that a relationship between morphological irregularity and phonotactic complexity may exist in some languages (Hay and Baayen, 2003; Burzio, 2002; Hay, 2003), but not others (Doucette et al., 2024).

5 Discussion

Although previous work on biases in the lexicon have implied that certain lexical trade-offs are cross-linguistic universals, our findings suggest that evidence of these universals may not be as strong as previously thought. For example, Zipf's law of abbreviation has been studied extensively, showing that a relationship between word length and frequency holds cross-linguistically. Previous work has shown a strong negative correlation across languages (Piantadosi, 2014), which indeed does hold in all 25 languages in our data. However, it is unclear if there is a direct *causal* relationship between frequency and word length. The FCI algorithm allows us to identify possible unmeasured confounding variables in a causal model. In our analysis, possible confounding in the relationship between word length and frequency is identified in nearly all languages. If only word length and frequency are considered, the causal model underlying Zipf's law cannot be identified. It is possible that surprisal is the confounding variable in this relationship, as

suggested by Piantadosi et al. (2011), or it could be something else. Structural causal modeling and causal discovery provide a framework for testing this, which we leave to future work.

We also identified possible confounding in the relationships between word length and phonotactic complexity and word length and morphological irregularity. This potential confounding occurred consistently across languages, as it did in the relationship between word length and frequency. This suggests that there may be universal lexical biases involving either word length or some other cause of word length. Again, this could be surprisal. The confounding variable could also differ across languages, but the existence of an association between word length and these variables appears to be universal.

While relationships involving word length occur consistently across languages, we find strong evidence of variation in the other relationships examined. In approximately half of the languages in our sample, there is no association between frequency and morphological irregularity, frequency and phonotactic complexity, and morphological irregularity and phonotactic complexity. In the languages where associations do exist, there is a probable unmeasured confounder. This suggests that lexicons may vary in whether or not these relationships are constrained, and that a set of universal lexical biases may not exist. Previous work has claimed that these relationships are universal, but considering a larger set of variables with causal discovery shows that this may not be true. If the properties of a lexicon are the result of some universal cognitive pressure (towards efficient communication, for example), only minor variation would be expected, rather than qualitative variation in whether or not a trade-off exists. This suggests that strong claims about universality and causal structure in the lexicon may need to be reconsidered.

We also note that the notion of causality in the lexicon implies diachronic language change, while our data represents observations of lexicons at a single point in time. Although in the ideal case we would examine changes in lexicons across time to determine causality, a causal model of a synchronic lexicon still has a useful interpretation. Pearl (2019) argues that a causal model can be interpreted as constraints on a mathematical system. The lexical biases we examine are exactly that: although they are likely caused by some underlying cognitive constraint, they impose constraints on possible lexicons.

However, our results are not without limitations. We examine a larger number of variables than many previous studies of lexical trade-offs, but the four variables we investigate are still not causally sufficient. Several possible unmeasured confounders are identified, leading to a causal graph that is not fully specified. Although we are able to identify patterns in causal structure across languages, we are not able to make any strong claims about direct causal relationships – there is simply not enough data. Our analysis is also limited by the dataset, which mainly includes morphologically complex words. It is possible that morphologically simple words may follow a different pattern. Additional data is needed to assess this possibility, which we plan on investigating in future work. We also note that the independence test used by the FCI algorithm in our analyses assumes that data is normally distributed. While this may be a reasonable assumption, further work is needed to assess how non-normality impacts the output of FCI. The data for word length, frequency, and phonotactic complexity have roughly normal distributions, but in many languages the morphological irregularity measure has a bimodal distribution. Nonparametric conditional independence testing is an active area of research (Li and Fan, 2020; Kim et al., 2022; Bianchi et al., 2023), and there may be tests with assumptions that better fit lexical data. There are also many different causal discovery algorithms, each with different assumptions about the data, and it should be explored how different algorithms can affect results. We leave investigating these possibilities to future work.

6 Conclusion

In this paper, we have shown that causal discovery methods can help identify relationships between statistical properties of the lexicon, providing more information about these relationships than correlations or regression models of pairs or small sets of variables. Given a dataset with more variables measured, and more languages, causal discovery may make it possible to determine exactly how the lexicon of a language is constrained, and how these biases vary across languages. In future work, we plan on applying causal discovery to a larger dataset with more languages and more variables with the goal of identifying a more specified causal model.

Acknowledgements

We thank the Montreal Computational & Quantitative Linguistics Lab for helpful feedback. The first author was supported by funding from the Fonds de recherche du Québec - Société et culture (FRQSC). The third author acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (RGPIN-2023-04873) and the Canada Research Chairs program. The second author also gratefully acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-06385) and the Canada CIFAR AI Chairs Program. This research was enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada.

References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855.

Pascal Bianchi, Kevin Elgui, and François Portier. 2023. Conditional independence testing via weighted partial copulas. *Journal of Multivariate Analysis*, 193:105120.

Luigi Burzio. 2002. Missing players: Phonology and the past-tense debate. *Lingua*, 112(3):157–199.

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594.

Fermín Moscoso del Prado Martín. 2014. Grammatical change begins within the word: Causal modeling of the co-evolution of Icelandic morphology and syntax. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Johannes Dellert. 2019. *Information-theoretic causal inference of lexical flow*. Language Science Press.

Johannes Dellert. 2024. Causal inference of diachronic semantic maps from cross-linguistic synchronic polysemy data. *Frontiers in Communication*, 8.

Amanda Doucette, Ryan Cotterell, Morgan Sonderegger, and Timothy J. O'Donnell. 2024. Correlation does not imply compensation: Complexity and irregularity in the lexicon. In *Proceedings of the Society for Computation in Linguistics 2024*, pages 117–128.

Viviana Fratini, Joana Acha, and Itziar Laka. 2014. Frequency and morphological irregularity are independent variables. Evidence from a corpus study of Spanish verbs. *Corpus Linguistics and Linguistic Theory*, 10(2):289–314.

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

Peter Graff. 2012. *Communicative efficiency in the lexicon*. Ph.D. thesis, Massachusetts Institute of Technology.

Jennifer Hay. 2003. *Causes and Consequences of Word Structure*. Routledge.

Jennifer Hay and Harald Baayen. 2003. Phonotactics, parsing and productivity. *Italian Journal of Linguistics*, 15:99–130.

M.A. Hernán and J.M. Robins. 2024. *Causal Inference: What If*. Chapman & Hall/CRC.

Jeremy Irvin, Daniel Spokoyny, and Fermín Moscoso Martın, del Prado Martín. 2016. Dynamical systems modeling of the child–mother dyad: Causality between child-directed language complexity and language development. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 38.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. 2012. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. 2022. Local permutation tests for conditional independence. *The Annals of Statistics*, 50(6):3388 – 3414.

Natalia Levshina. 2021. Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations. *Frontiers in Psychology*, 12.

Natalia Levshina. 2022. *Communicative Efficiency*. Cambridge University Press.

Chun Li and Xiaodan Fan. 2020. On nonparametric conditional independence tests for continuous variables. *WIREs Computational Statistics*, 12(3):e1489.

Kyle Mahowald, Isabelle Dautriche, Edward Gibson, and Steven T. Piantadosi. 2018. Word forms are structured for efficient use. *Cognitive Science*, 42(8):3116–3134.

Daniel Malinsky and David Danks. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980.

Stephan C. Meylan and Thomas L. Griffiths. 2021. The challenges of large-scale, web-based language datasets: Word length and predictability revisited. *Cognitive Science*, 45(6):e12983.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Fermín Moscoso del Prado Martín and Christian Brendel. 2016. Case and cause in Icelandic: Reconstructing causal networks of cascaded language changes. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2421–2430.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Judea Pearl. 2019. On the interpretation of do(x). *Journal of Causal Inference*, 7(1):20192002.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

François Pellegrino, Christophe Coupé, and Egidio Marsico. 2011. A cross-language perspective on speech information rate. *Language*, 87(3):539–558.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Tiago Pimentel, Clara Meister, Ethan Wilcox, Kyle Mahowald, and Ryan Cotterell. 2023. Revisiting the optimality of word lengths. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2240–2255.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18.

Uriel Cohen Priva. 2017. Informativity and the actuation of lenition. *Language*, 93(3):569–597.

Uriel Cohen Priva and Emily Gleason. 2020. The causal structure of lenition: A case for the causal precedence of durational shortening. *Language*, 96(2):413–448.

Peter Spirtes, Clark Glymour, and Richard Scheines. 1993. *Causation, Prediction, and Search*. Springer.

Peter Spirtes, Christopher Meek, and Thomas Richardson. 1995. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, page 499–506.

Daniel Spokoyny, Jeremy Irvin, and Fermín Moscoso del Prado Martín. 2016. Explicit causal connections between the acquisition of linguistic tiers: Evidence from dynamical systems modeling. In *Proceedings of the 7th workshop on Cognitive Aspects of Computational Language Learning*, pages 73–81.

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

Charles Yang. 2016. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. MIT press.

Alessio Zanga, Elif Ozkirimli, and Fabio Stella. 2022. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 151:101–129.

George Kingsley Zipf. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human eoclogy*. Addison-Wesley Press.