# Frozen Binomials on the Web:
# Word Ordering and Language Conventions in Online Text

Katherine Van Koevering
kav64@cornell.edu
Cornell University

Austin R. Benson
arb@cs.cornell.edu
Cornell University

Jon Kleinberg
kleinberg@cornell.edu
Cornell University

## ABSTRACT

There is inherent information captured in the order in which we write words in a list. The orderings of binomials — lists of two words separated by 'and' or 'or' — has been studied for more than a century. These binomials are common across many areas of speech, in both formal and informal text. In the last century, numerous explanations have been given to describe what order people use for these binomials, from differences in semantics to differences in phonology. These rules describe primarily 'frozen' binomials that exist in exactly one ordering and have lacked large-scale trials to determine efficacy.

Text in online social media such as Reddit provides a unique opportunity to study these lists in the context of informal text at a very large scale. In this work, we expand the view of binomials to include a large-scale analysis of both frozen and non-frozen binomials in a quantitative way. Using this data, we then demonstrate that most previously proposed rules are ineffective at predicting binomial ordering. By tracking the order of these binomials across time and communities we are able to establish additional, unexplored dimensions central to these predictions and demonstrate the global structure of the binomials across communities.

Expanding beyond the question of individual binomials, we also explore the global structure of binomials in various communities, establishing a new model for these lists and analyzing this structure for non-frozen and frozen binomials. Additionally, novel analysis of trinomials — lists of length three — suggests that none of the binomials analysis applies in these cases. Finally, we demonstrate how large data sets gleaned from the web can be used in conjunction with older theories to expand and improve on old questions.

## 1 INTRODUCTION

Lists are extremely common in text and speech, and the ordering of items in a list can often reveal information. For instance, orderings can denote relative importance, such as on a to-do list, or signal status, as is the case for author lists of scholarly publications. In other cases, orderings might come from cultural or historical conventions. For example, 'red, white, and blue' is a specific ordering of colors that is recognizable to those familiar with American culture.

The orderings of lists in text and speech is a subject that has been repeatedly touched upon for more than a century. By far the most frequently studied aspect of list ordering is the *binomial*, a list of two words usually separated by a conjunction such as 'and' or 'or', which is the focus of our paper. The academic treatment of binomial orderings dates back more than a century to Jespersen [12], who proposed in 1905 that the ordering of many common English binomials could be predicted by the rhythm of the words. In the case of a binomial consisting of a monosyllable and a disyllable, the prediction was that the monosyllable would appear first followed by the conjunction 'and'. The idea was that this would give a much more standard and familiar syllable stress to the overall phrase, e.g., the binomial 'bread and butter' would have the preferable rhythm compared to 'butter and bread.'

This type of analysis is meaningful when the two words in the binomial nearly always appear in the same ordering. Binomials like this that appear in strictly one order (perhaps within the confines of some text corpus), are commonly termed *frozen binomials* [9, 14]. Examples of frozen binomials include 'salt and pepper' and 'pros and cons', and explanations for their ordering in English and other languages have become increasingly complex. Early work focused almost exclusively on common frozen binomials, often drawn from everyday speech. More recent work has expanded this view to include nearly frozen binomials, binomials from large data sets such as books, and binomials of particular types such as food, names, and descriptors [5, 7, 10, 18, 19, 27]. Additionally, explanations have increasingly focused on meaning rather than just sound, implying value systems inherent to the speaker or the culture of the language's speakers (one such example is that men are usually listed before women in English [26]). The fact that purely phonetic explanations have been insufficient suggests that list orderings rely at least partially on semantics, and it has previously been suggested that these semantics could be revealing about the culture in which the speech takes place [7]. Thus, it is possible that understanding these orderings could reveal biases or values held by the speaker.

Overall, this prior research has largely been confined to pristine examples, often relying on small samples of lists to form conclusions. Many early studies simply drew a small sample of what the author(s) considered some of the more representative or prominent binomials in whatever language they were studying [2, 3, 6–8, 12, 14, 23, 25]. Other researchers have used books or news articles [5, 9], or small samples from the Web (web search results and Google books) [18]. Many of these have lacked a large-scale text corpus and have relied on a focused set of statistics about word orderings.

Thus, despite the long history of this line of inquiry, there is an opportunity to extend it significantly by examining a broad range of questions about binomials coming from a large corpus of online text data produced organically by many people. Such an analysis could produce at least two types of benefits. First, such a study could help us learn about cultural phenomena embedded in word orderings and how they vary across communities and over time. Second, such an analysis could become a case study for the extension of theories developed at small scales in this domain to a much larger context.

**The present work: Binomials in large-scale online text.** In this work, we use data from large-scale Internet text corpora to study binomials at a massive scale, drawing on text created by millions of users. Our approach is more wholesale than prior work — we focus on all binomials of sufficient frequency, without first restricting to small samples of binomials that might be frozen. We draw our data from Reddit, which in addition to large volume, also let us characterize binomials in new ways, and analyze differences in binomial orderings across communities and over time. Furthermore, the subject matter on Reddit leads to many lists about people and organizations that lets us study orderings of proper names — a key setting for word ordering which has been difficult to study by other means.

We begin our analysis by introducing several new key measures for the study of binomials, including a quantity we call *asymmetry* that measures how frequently a given binomial appears in some ordering. By looking at the distribution of asymmetries across a wide range of binomials, we find that most binomials are not frozen, barring a few strong exceptions. At the same time, there may still be an ordering preference. For example, '10 and 20' is not a frozen binomial; instead, the binomial ordering '10 and 20' appears 60% of the time and '20 and 10' appears 40% of time.

We also address temporal and community structure in collections of binomials. While it has been recognized that the orderings of binomials may change over time or between communities [2, 3, 6, 8, 14, 18], there has been little analysis of this change. We develop new metrics for the *agreement* of binomial orderings across communities and the *movement* of binomial orderings over time. Using subreddits as communities, these metrics reveal variations in orderings, some of which suggest cultural change influencing language. For example, in one community, we find that over a period of 10 years, the binomial 'son and daughter' went from nearly frozen to appearing in that order only 64% of the time.

While these changes do happen, they are generally quite rare. Most binomials — frozen or not — are ordered in one way about the same percentage of the time, regardless of community or the year. We develop a null model to determine how much variation in binomial orderings we might expect across communities and across time, if binomial orderings were randomly ordered according to global asymmetry values. We find that there is less variation across time and communities in the data compared to this model, implying that binomial orderings are indeed remarkably stable.

Given this stability, one might expect that the dominant ordinality of a given binomial is still predictable, even if the binomial is not frozen. For example, one might expect that the global frequency of a single word or the number of syllables in a word would

predict ordering in many cases. However, we find that these simple predictors are quite poor at determining binomial ordering.

On the other hand, we find that a notion of 'proximity' is robust at predicting ordering in some cases. Here, the idea is that the person producing the text will list the word that is conceptually "closer" to them first — a phenomenon related to a "Me First" principle of binomial orderings suggested by Cooper and Ross [7]. One way in which we study this notion of proximity is through sports team subreddits. For example, we find that when two NBA team names form a binomial on a specific team's subreddit, the team that is the subject of the subreddit tends to appear first.

The other source of improved predictions comes from using word embeddings [17]: we find that a model based on the positions of words in a standard pre-trained word embedding can be a remarkably reliable predictor of binomial orderings. While not applicable to all words, such as names, this type of model is strongly predictive in most cases.

Since binomial orderings are in general difficult to predict individually, we explore a new way of representing the global binomial ordering structure — for a given community (subreddit), we form a directed graph where an edge from $i$ to $j$ means that $i$ tends to come before $j$ in binomials. These graphs show tendencies across the English language and also reveal peculiarities in the language of particular communities. For instance, in a graph formed from the binomials in a sports community, the names of sports teams and cities are closely clustered, showing that they are often used together in binomials. Similarly, we identify clusters of names, numbers, and years. The presence of cycles in these graphs are also informative. For example, cycles are rare in graphs formed from proper names in politics, suggesting a possible hierarchy of names, and at the same time very common for other binomials. This suggests that no such hierarchy exists for most of the English language, further complicating attempts to predict binomial order.

Finally, we expand our work to include multinomials, which are lists of more than two words. There already appears to be more structure in trinomials (lists of three) compared to binomials. Trinomials are likely to appear in exactly one order, and when they appear in more than one order the last word is almost always the same across all instances. For instance, in one section of our Reddit data, 'Fraud, Waste, and Abuse' appears 34 times, and 'Waste, Fraud, and Abuse' appears 20 times. This could point to, for example, recency principles being more important in lists of three than in lists of two. While multinomials were in principle part of the scope of past research in this area, they were difficult to study in smaller corpora, suggesting another benefit of working at our current scale.

## 1.1 Related Work

Interest in list orderings spans the last century [2, 14], with a focus almost exclusively on binomials. This research has primarily investigated frozen binomials, also called *irreversible binomials*, *fixed coordinates*, and *fixed conjuncts* [25], although some work has also looked at non-coordinate freezes where the individual words are nonsensical by themselves (e.g., 'dribs and drabs') [25]. One study has directly addressed *mostly* frozen binomials [18], and we expand the scope of this paper by exploring the general question of how frequently binomials appear in a particular order. Early research

investigated languages other than English [2, 14], but most recent research has worked almost exclusively with English. Overall, this prior research can be separated into three basic categories — phonological rules, semantic rules, and metadata rules.

**Phonology.** The earliest research on binomial orderings proposed mostly phonological explanations, particularly rhythm [12, 23]. Another highly supported proposal is Panini's Law, which claims that words with fewer syllables come first [22]; we find only very mild preference for this type of ordering on Reddit. Cooper and Ross's work expands these to a large list of rules, many overlapping, and suggests that they can compound [7]; a number of subsequent papers have expanded on their work [6, 22, 25, 26].

**Semantics.** There have also been a number of semantic explanations, mostly in the form of categorical tendencies (such as 'desirable before undesirable') that may have cultural differences [2, 14]. The most influential of these may be the 'Me First' principle codified by Cooper and Ross. This suggests that the first word of a binomial tends to follow a hierarchy that favors 'here', 'now', present generation, adult, male, and positive. Additional hierarchies also include a hierarchy of food, plants vs. animals, etc. [7].

**Frequency.** More recently, it has been proposed that the more cognitively accessible word might come first, which often means the word the author sees or uses most frequently [15]. There has also been debate on whether frequency may encompass most phonological and semantic rules that have been previously proposed [5, 8]. We find that frequency is in general a poor predictor of word ordering.

**Combinations.** Given the number of theories, there have also been attempts to give a hierarchy of rules and study their interactions [5, 18]. This research has complemented the proposals of Cooper and Ross [7]. These types of hierarchies are also presented as explanations for the likelihood of a binomial becoming frozen [18].

**Names.** Work on the orderings of names has been dominated by a single phenomenon: men's names usually come before women's names. Explanations range from a power differential, to men being more 'agentic' within 'Me First', to men's names being more common or even exhibiting more of the phonological features of words that usually come first [7, 8, 10, 11, 15, 18, 26, 27]. However, it has also been demonstrated that this preference may be affected by the author's own gender and relationship with the people named [11, 27], as well as context more generally [13].

**Orderings on the Web.** List orderings have also been explored in other Web data, specifically on the ordering of tags applied to images [21]. There is evidence that these tags are ordered intentionally by users, and that a bias to order tag A before tag B may be influenced by historical precedent in that environment but also by the relative importance of A and B [21]. Further work also demonstrates that exploiting the order of tags on images can improve models that rank those images [20].

## 2 DATA

We take our data from Reddit, a large social media website divided into subcommunities called 'subreddits' or 'subs'. Each subreddit has a theme (usually clearly expressed in its name), and we have focused our study on subreddits primarily in sports and politics, in part because of the richness of proper names in these domains: r/nba, r/nfl, r/politics, r/Conservative, r/Libertarian, r/The_Donald,
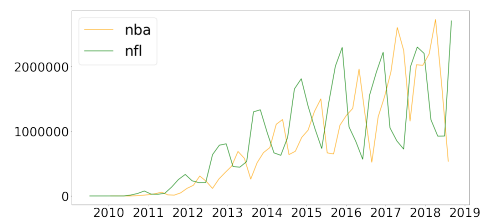


**Figure 1: Histogram of comment timestamps for r/nba and r/nfl. Both subreddits exhibit a seasonal structure. The number of comments is increasing for all subreddits.**

r/food, along with a variety of NBA team subreddits (e.g., r/rockets for the Houston Rockets). Apart from the team-specific and food subreddits, these are among the largest and most heavily used subreddits [24]. We gather text data from comments made by users in discussion threads. In all cases, we have data from when the subreddit started until mid-2018. (Data was contributed by Cristian Danescu-Niculescu-Mizil.) Reddit in general, and the subreddits we examined in particular, are rapidly growing, both in terms of number of users and number of comments.

Some of the subreddits we looked at (particularly sports subreddits) exhibited very distinctive 'seasons', where commenting spikes (Fig. 1). These align with, e.g., the season of the given sport. When studying data across time, our convention is to bin the data by year, but we adjust the starting point of a year based on these seasons. Specifically, a year starts in May for r/nfl, August for r/nba, and February for all politics subreddits.

We use two methods to identify lists from user comments: 'All Words' and 'Names Only', with the latter focusing on proper names. In both cases, we collect a number of lists and discard lists for any pair of words that appear fewer than 30 times within the time frame that we examined (see Table 1 for summary statistics).

The All Words method simply searches for two words $A$ and $B$ separated by 'and' or 'or', where a word is merely a series of characters separated by a space or punctuation. This process only captures lists of length two, or binomials. We then filter out lists containing words from a collection of stop-words that, by their grammatical role or formatting structure, are almost exclusively involved in false positive lists. No metadata is captured for these lists beyond the month and year of posting.

The Names Only method uses a curated list of full names relevant to the subreddit, focusing on sports and politics. For sports, we collected names of all NBA and NFL player active during 1980–2019 from *basketball-reference.com* and *pro-football-reference.com*. For politics, we collected the names of congresspeople from *the @unitedstates project* [1]. To form lists, we search for any combination of any part of these names such that at least two partial names are separated by 'and', 'or', 'v.s.', 'vs', or '/' and the rest are separated by ','. While we included a variety of separators, about 83% of lists include only 'and', about 17% include 'or' and the rest of the separators are negligible. Most lists that we retrieve in this way are of length 2, but we also found lists up to length 40 (Fig. 2). Finally, we also captured full metadata for these lists, including a timestamp, the user, any flairs attributed to the user (short custom text that appears next to the username), and other information.

**Table 1: Summary statistics of subreddit list data that we investigate in this paper.**

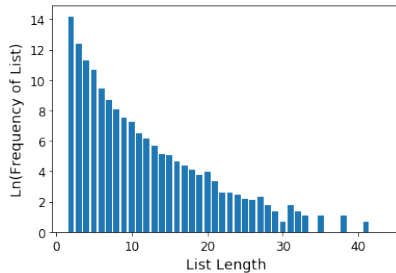|  | Total Comments | Seasons in Operation | Total Lists | Name Lists | Unique Names |
|---|---|---|---|---|---|
| r/food | 4580723 | 10 | 924457 | - | - |
| r/nba | 44661753 | 10 | 4807623 | 1826090 | 29336 |
| r/nfl | 46515339 | 9 | 5855011 | 1287597 | 35014 |
| r/Conservative | 2772010 | 11 | 859252 | 77593 | 6742 |
| r/Libertarian | 5485366 | 11 | 2318685 | 146729 | 8964 |
| r/politics | 86208050 | 12 | 26142750 | 1854730 | 43316 |
| r/The_Donald | 2576974 | 3 | 3964936 | 388339 | 18891 |



**Figure 2: A histogram of the log frequency of lists of various lengths, where we use name lists for r/nba. In this case, there is no filtering applied, but we cap list length at 50.**

## 3 DIMENSIONS OF BINOMIALS

In this paper we introduce a new framework to interpret binomials, based on three properties: *asymmetry* (how frozen a binomial is), *movement* (how binomial orderings change over time), and *agreement* (how consistent binomial orderings are between communities), which we will visualize as a cube with three dimensions. Again, prior work has focused essentially entirely on asymmetry, and we argue that this can only really be understood in the context of the other two dimensions.

For this paper we will use the convention {A,B} to refer to an unordered pair of words, and [A,B] to refer to an ordered pair where A comes before B. We say that [A,B] and [B,A] are the two possible *orientations* of {A,B}.

### 3.1 Definitions

Previous work has one main measure of binomials — their 'frozenness'. A binomial is 'frozen' if it always appears with a particular order. For example, if the pair {'arrow', 'bow'} always occurs as ['bow', 'arrow'] and never as ['arrow', 'bow'], then it is frozen. This leaves open the question of how describe the large number of binomials that are not frozen. To address this point, we instead consider the *ordinality* of a list, or how often the list is 'in order' according to some arbitrary underlying reference order. Unless otherwise specified, the underlying order is assumed to be alphabetical. If the list ['cat', 'dog'] appears 40 times and the list ['dog', 'cat'] 10 times, then the list {'cat', 'dog'} would have an ordinality of 0.8.

Let $n_{x,y}$ be the number of times the ordered list $[x, y]$ appears, and let $f_{x,y} = n_{x,y}/(n_{x,y} + n_{y,x})$ be the fraction of times that the unordered version of the list appears in that order. We formalize ordinality as follows.

*Definition 3.1 (Ordinality).* Given an ordering $<$ on words (by default, we assume alphabetical ordering), the **ordinality** $o_{x,y}$ of the pair $\{x, y\}$ is equal to $f_{x,y}$ if $x < y$ and $f_{y,x}$ otherwise.

Similarly, we introduce the concept of asymmetry in the context of binomials, which is how often the word appears in its dominant order. In our framework, a 'frozen' list is one with ordinality 0 or 1 and would be considered a high asymmetry list, with asymmetry of 1. A list that appears as ['A', 'B'] half of the time and ['B', 'A'] half of the time (or with ordinality 0.5) would be considered a low asymmetry list, with asymmetry of 0.

*Definition 3.2 (Asymmetry).* The **asymmetry** of an unordered list $\{x, y\}$ is $A_{x,y} = 2 \cdot |o_{x,y} - 0.5|$.

The Reddit data described above gives us access to new dimensions of binomials not previously addressed. We define movement as how the ordinality of a list changes over time

*Definition 3.3 (Movement).* Let $o_{x,y,t}$ be the ordinality of an unordered list $\{x, y\}$ for data in year $t \in T$. The **movement** of $\{x, y\}$ is $M_{x,y} = \max_{t \in T} o_{x,y,t} - \min_{t \in T} o_{x,y,t}$.

And agreement describes how the ordinality of a list differs between different communities.

*Definition 3.4 (Agreement).* Let $o_{x,y,c}$ be the ordinality of an unordered list $x, y$ for data in community (subreddit) $c \in C$. The **agreement** of $\{x, y\}$ is $A_{x,y} = 1 - (\max_{c \in C} o_{x,y,c} - \min_{c \in C} o_{x,y,c})$.

### 3.2 Dimensions

Let the point $(A, M, G)_{x,y}$ be a vector of the asymmetry, movement, and agreement for some unordered list $\{x, y\}$. These vectors then define a 3-dimensional space in which each list occupies a point. Since our measures for asymmetry, agreement, and movement are all defined from 0 to 1, their domains form a unit cube (Fig. 3). The *corners* of this cube correspond to points with coordinates are entirely made up of 0s or 1s. By examining points near the corners of this cube, we can get a better understanding of the range of binomials. Some corners are natural — it is easy to imagine a high asymmetry, low movement, high agreement binomial — such as {'arrow', 'bow'} from earlier. On the other hand, we have found no good examples of a high asymmetry, low movement, low agreement binomial. There are a few unusual examples, such as {10, 20}, which has 0.4 asymmetry, 0.2 movement, and 0.1 agreement and is clearly visible as an isolated point in Fig. 3.
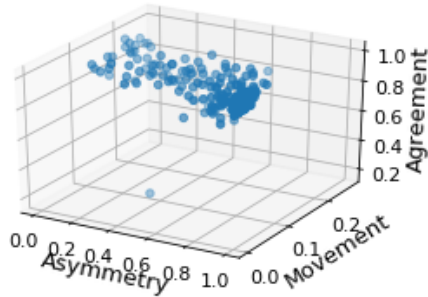
**Figure 3: 309 binomials that occur at least 30 times per year in r/politics, r/nba, and r/nfl mapped on to the 3-dimensional cube. The point on the bottom left is {'10', '20'}.**

**Asymmetry.** While a majority of binomials have low asymmetry, almost all previous work has focused exclusively on high-asymmetry binomials. In fact, asymmetry is roughly normally distributed across binomials with an additional increase of highly asymmetric binomials (Fig. 4). This implies that previous work has overlooked the vast majority of binomials, and an investigation into whether rules proposed for highly asymmetric binomials also functions for other binomials is a core piece of our analysis.
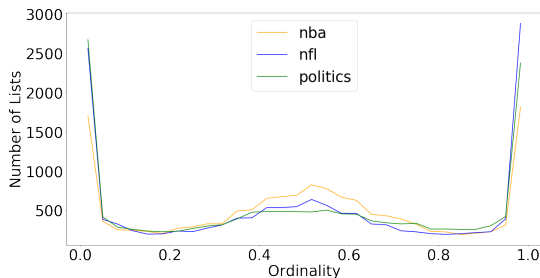


**Figure 4: Histograms of the alphabetical orientation of the 14920 most common binomials within r/nba, r/nfl and r/politics. Note that while there are many frozen binomials (with orientation of 0 or 1), the rest of the binomials appear to be roughly normally distributed around 0.5.**

**Movement.** The vast majority of binomials have low movement. However, the exceptions to this can be very informative. Within r/nba a few of these pairs show clear change in linguistics and/or culture. The binomial ['rpm', 'vorp'] (a pair of basketball statistics) started at 0.74 ordinality and within three years dropped to 0.32 ordinality, showing a potential change in users' representation of how these statistics relate to each other. In r/politics, ['daughter', 'son'] moved from 0.07 ordinality to 0.36 ordinality over ten years. This may represent a cultural shift in how users refer to children, or a shift in topics discussed relating to children. And in r/politics, ['dems', 'obama'] went from 0.75 ordinality to 0.43 ordinality from 2009–2018, potentially reflecting changes in Obama's role as a defining feature of the Democratic Party. Meanwhile the ratio of unigram frequency of 'dems' to 'obama' actually increased from

**Table 2: Covariance table for movement in r/nba, r/nfl and r/politics.**

|  | r/nba | r/nfl | r/politics |
|---|---|---|---|
| r/nba | 0.00426 | 0.00240 | 0.00260 |
| r/nfl | 0.00240 | 0.00468 | 0.00320 |
| r/politics | 0.00560 | 0.00320 | 0.00579 |

10% to 20% from 2010 to 2017. Similarly, ['fdr', 'lincoln'] moved from 0.49 ordinality to 0.17 ordinality from 2015–2018. This is particularly interesting, since in 2016 'fdr' had a unigram frequency 20% higher than 'lincoln', but in 2017 they are almost the same. This suggests that movement could be unrelated to unigram frequency changes. Note also that the covariance for movement across sub-reddits is quite low (Table 2), and movement in one subreddit is not necessarily reflected by movement in another.

**Agreement.** Most binomials have high agreement (Table 3) but again the counterexamples are informative. For instance, ['score', 'kick'] has ordinality of 0.921 in r/nba and 0.204 in r/nfl. This likely points to the fact that American football includes field goals. A less obvious example is the list ['ceiling', 'floor']. In r/nba and r/nfl, it has ordinality 0.44, and in r/politics, it has ordinality 0.27.

There are also differences among proper nouns. One example is ['france', 'israel'], which has ordinality 0.6 in r/politics, 0.16 in r/Libertarian, and 0.51 in r/The_Donald (and the list does not appear in r/Conservative). And the list ['romney', 'trump'] has ordinality 0.48 in r/poltics, 0.55 in r/The_Donald, and 0.73 in r/Conservative.

## 4 MODELS AND PREDICTIONS

In this section, we establish a null model under which different communities or time slices have the same probability of ordering a binomial in a particular way. With this, we would expect to see variation in binomial asymmetry. We find that our data shows smaller variation than this null model predicts, suggesting that binomial orderings are extremely stable across communities and time. From this, we might also expect that orderings are predictable; but we find that standard predictors in fact have limited success.

### 4.1 Stability of Asymmetry

Recall that the asymmetry of binomials with respect to alphabetic order (excluding frozen binomials) is roughly normal centered around 0.5 (Fig. 4). One way of seeing this type of distribution would be if binomials are ordered randomly, with $p = 0.5$ for each order. In this case, if each instance $l$ of a binomial $\{x, y\}$ takes value 0 (non-alphabetical ordering) or 1 (alphabetical ordering), then $l \sim \text{Bernoulli}(0.5)$. If $\{x, y\}$ appears $n$ times, then the number of instances of value 1 is distributed by $W \sim \text{Bin}(n, 0.5)$, and $W/n$ is approximately normally distributed with mean 0.5.

One way to test this behavior is to first estimate $p$ for each list within each community. If the differences in these estimates are not normal, then the above model is incorrect. We first omit frozen binomials before any analysis. Let $L$ be a set of unordered lists and $C$ be a set of communities. We estimate $p$ for list $l \in L$ in community $c \in C$ by $\hat{p}_{l,c} = o_{l,c}$, the ordinality of $l$ in $C$. Next, for all $l \in L$ let $p_l^* = \max_{c \in C}(\hat{p}_{l,c}) - \min_{c \in C}(\hat{p}_{l,c})$. The distribution of

**Table 3: The average difference in asymmetry between the same binomial in various subreddits. The difference between r/nba and r/nfl is 0.062.**

|  | r/Conservative | r/Libertarian | r/politics | r/the_Donald |
|---|---|---|---|---|
| r/Conservative |  | 0.057 | 0.049 | 0.055 |
| r/Libertarian | 0.057 |  | 0.053 | 0.059 |
| r/politics | 0.049 | 0.053 |  | 0.052 |
| r/The_Donald | 0.055 | 0.059 | 0.052 |  |

$p_l^*$ over $l \in L$ has median 0, mean 0.0145, and standard deviation 0.0344. We can perform a similar analysis over time. Define $Y$ as our set of years, and $\hat{p}_{l,y} = o_{l,y}$ for $y \in Y$ our estimates. The distribution of $p_l' = \max_{y \in Y}(\hat{p}_{l,y}) - \min_{y \in Y}(\hat{p}_{l,y})$ over $l \in L$ has median 0.0216, mean 0.0685, and standard deviation 0.0856. The fact that $p$ varies very little across both time and communities suggests that there is some $p_l$ for each $l \in L$ that is consistent across time and communities, which is not the case in the null model, where these values would be normally distributed.

We also used a bootstrapping technique to understand the mean variance in ordinality for lists over communities and years. Specifically, let $o_{l,c,y}$ be the ordinality of list $l$ in community $c$ and year $y$, $O_l$ be the set of $o_{l,c,y}$ for a given list $l$, and $s_l$ be the standard deviation of $O_l$. Finally, let $\bar{s}$ be the average of the $s_l$. We re-sample data by randomizing the order of each binomial instance, sampling its orderings by a binomial random variable with success probability equal to its ordinality across all seasons and communities ($p_l$). We repeated this process to get samples estimates $\{\bar{s}_1, \ldots, \bar{s}_k\}$, where $k$ is the size of the set of seasons and communities. These averages range from 0.0277 to 0.0278 and are approximately normally distributed (each is a mean over an approximately normal scaled Binomial random variable). However, $\bar{s} = 0.0253$ for our non-randomized data. This is significantly smaller than the randomized data and implies that the true variation in $p_l$ across time and communities is even smaller than a binomial distribution would predict. One possible explanation for this is that each instance of $l$ is not actually independent, but is in fact anti-correlated, violating one of the conditions of the binomial distribution. An explanation for that could be that users attempt to draw attention by intentionally going against the typical ordering [14], but it is an open question what the true model is and why the variation is so low. Regardless, it is clear that the orientation of binomials varies very little across years and communities (Fig. 5).

## 4.2 Prediction Results

Given the stability of binomials within our data, we now try to predict their ordering. We consider deterministic or rule-based methods that predict the order for a given binomial. We use two classes of evaluation measures for success on this task: (i) by token — judging each instance of a binomial separately; and (ii) by type — judging all instances of a particular binomial together. We further characterize these into weighted and unweighted.

To formalize these notions, first consider any unordered list $\{x, y\}$ that appears $n_{x,y}$ times in the orientation $[x, y]$ and $n_{y,x}$ times in the orientation $[y, x]$. Since we can only guess one order, we will have either $n_{x,y}$ or $n_{y,x}$ successful guesses for $\{x, y\}$ when
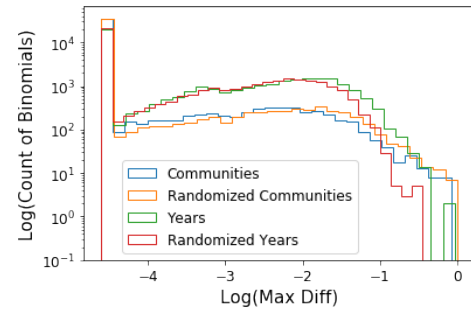


**Figure 5: Histogram of the maximum difference in $p_l$ for all lists $l$ across communities and years, on a log-log scale. We add 0.01 to all differences to show cases with a difference of 0, which is represented as the bar on the left of the graph (mostly due to frozen binomials). We sampled 40000 instances for this graph, since there was variation in the number of binomials across years and communities.**

**Table 4: Accuracy of binomial orientation predictions using a number of basic rules. The scoring was done based on "unweighted type" scoring, and statistics are given based on the scores across the subreddits.**

|  | Min | Mean | Max |
|---|---|---|---|
| Alphabetical | 0.365 | 0.462 | 0.512 |
| Length | 0.483 | 0.563 | 0.639 |
| Phonemes | 0.477 | 0.561 | 0.642 |
| Syllables | 0.588 | 0.654 | 0.714 |
| Unigram Frequency | 0.553 | 0.599 | 0.629 |

guessing by token. The unweighted token score (UO) and weighted token score (WO) are the macro and micro averages of this accuracy.

If predicting by type, let $S$ be the lists such that the by-token prediction is successful at least half of the time. Then the unweighted type score (UT) and weighted type score (WT) are the macro and micro averages of $S$.

**Basic Features.** We first use predictors based on rules that have previously been proposed in the literature: word length, number of phonemes, number of syllables, alphabetical order, and frequency. We collect all binomials but make predictions only on binomials appearing at least 30 times total, stratified by subreddit. However, none of these features appear to be particularly predictive across the board (Table 4). A simple linear regression model predicts close to random, which bolsters the evidence that these classical rules for frozen binomials are not predictive for general binomials.

**Table 5: Count for number of syllables in first and second word of all binomials in r/politics. First word is rows, second word is columns. Overall, shorter words are significantly more likely to come before longer words (see also Table 6).**

|   | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 24388 | 17964 | 8707 | 3269 | 719 |
| 2 | 17354 | 18296 | 10507 | 4364 | 1139 |
| 3 | 7758 | 8462 | 7271 | 3929 | 1134 |
| 4 | 3148 | 3313 | 2889 | 1975 | 721 |
| 5 | 500 | 621 | 642 | 472 | 197 |

**Table 6: Similar to Table 5, this table displays the fraction of of lists with this ordering of that size. So, [2,1] is 0.491 and [1,2] is 0.509, meaning monosyllables slightly tend to come before disyllables. This is cut off at 5 syllables, as more syllables has too few instances.**

|   | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 1.0 | 0.509 | 0.529 | 0.509 | 0.590 |
| 2 | 0.491 | 1.0 | 0.554 | 0.568 | 0.647 |
| 3 | 0.471 | 0.446 | 1.0 | 0.576 | 0.639 |
| 4 | 0.491 | 0.432 | 0.424 | 1.0 | 0.604 |
| 5 | 0.410 | 0.353 | 0.361 | 0.396 | 1.0 |

Perhaps the oldest suggestion to explain binomial orderings is that if there are two words A and B, and A is monosyllabic and B is disyllabic, then A comes before B [12]. Within r/politics, we gathered an estimate of number of syllables for each word as given by a variation on the CMU Pronouncing Dictionary [4] (Tables 5 and 6). In a weak sense, Jespersen was correct that monosyllabic words come before disyllabic words more often than not; and more generally, shorter words come before longer words more often than not. However, as predictors, these principles are close to random guessing.

**Paired Predictions.** Another measure of predictive power is predicting which of two binomials has higher asymmetry. In this case, we take two binomials with very different asymmetry and try to predict which has higher asymmetry by our measures (we use the top-1000 and bottom-1000 binomials in terms of asymmetry for these tasks). For instance, we may predict that ['red', 'turquoise'] is more asymmetric than ['red', 'blue'] because the differences in lengths is more extreme. Overall, the basic predictors from the literature are not very successful (Table 7).

**Word Embeddings.** If we turn to more modern approaches to text analysis, one of the most common is *word embeddings* [17]. Word embeddings assign a vector $x_i$ to each word $i$ in the corpus, such that the relative position of these vectors in space encode information linguistically relevant relationships among the words. Using the Google News word embeddings, via a simple logistic model, we produce a vector $v^*$ and predict the ordering of a binomial on words $i$ and $j$ from $v^* \cdot (x_i - x_j)$. In this sense, $v^*$ can be thought of as a "sweep-line" direction through the space containing the word vectors, such that the ordering along this sweep-line is the predicted ordering of all binomials in the corpus. This yields surprisingly accurate results, with accuracy ranging from 70% to 85% across various subreddits (Table 9), and 80-100% accuracy on frozen
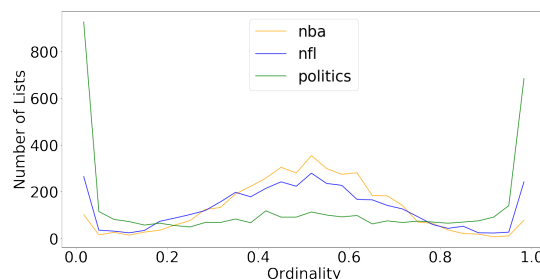


**Figure 6: Histogram of asymmetry for lists of names in r/nfl, r/nba and r/politics.**

binomials. This is by far the best prediction method we tested. It is important to note that not all words in our binomials could be associated with an embedding, so it was necessary to remove binomials containing words such as names or slang. However, retesting our basic features on this data set did not show any improvement, implying that the drastic change in predictive power is not due to the changed data set.

## 5 PROPER NOUNS AND THE PROXIMITY PRINCIPLE

Proper nouns, and names in particular, have been a focus within the literature on frozen binomials [7, 8, 10, 11, 13, 15, 16, 18, 26, 27], but these studies have largely concentrated on the effect of gender in ordering [7, 8, 10, 11, 13, 15, 18, 26, 27]. With Reddit data, however, we have many conversations about large numbers of celebrities, with significant background information on each. As such, we can investigate proper nouns in three subreddits: r/nba, r/nfl, and r/politics. The names we used are from NBA and NFL players (1970–2019) and congresspeople (pre-1800 and 2000–2019) respectively. We also investigated names of entities for which users might feel a strong sense of identification, such as a team or political group they support, or a subreddit to which they subscribe. We hypothesized that the group with which the user identifies the most would come first in binomial orderings. Inspired by the 'Me First Principle', we call this the *Proximity Principle*.

### 5.1 NBA Names

First, we examined names in r/nba. One advantage of using NBA players is that we have detailed statistics for ever player in every year. We tested a number of these statistics, and while all of them predicted statistically significant numbers ($p < 1e\text{-}6$) of binomials, they were still not very predictive in a practical sense (Table 10). The best predictor was actually how often the player's team was mentioned. Interestingly, the unigram frequency (number of times the player's name was mentioned overall) was not a good predictor. It is relevant to these observations that some team subreddits (and thus, presumably, fanbases) are significantly larger than others.

### 5.2 Subreddit and team names

Additionally, we also investigated lists of names of sports teams and subreddits as proper nouns. In this case we exploit an interesting structure of the r/nba subreddit which is not evident at scale in other subreddits we examined. In addition to r/nba, there exists

**Table 7: Paired prediction results.**

| predictor | r/Conservative | r/Libertarian | r/politics | r/The_Donald | r/nba | r/nfl | r/food |
|---|---|---|---|---|---|---|---|
| Alphabetical | 0.501 | 0.417 | 0.539 | 0.631 | 0.360 | 0.351 | 0.234 |
| Length | 0.445 | 0.481 | 0.330 | 0.505 | 0.482 | 0.324 | 0.447 |
| Phonemes | 0.430 | 0.460 | 0.324 | 0.513 | 0.428 | 0.447 | 0.417 |
| Syllables | 0.357 | 0.439 | 0.505 | 0.328 | 0.477 | 0.257 | 0.453 |

**Table 8: The accuracy using "unweighted type" for only frozen binomials, here defined as binomials with asymmetry above 0.97. The results suggest that these rules are equally ineffective for frozen and non-frozen binomials.**

|  | Min | Mean | Max |
|---|---|---|---|
| Alphabetical | 0.46 | 0.51 | 0.55 |
| Length | 0.37 | 0.45 | 0.54 |
| Phonemes | 0.33 | 0.44 | 0.51 |
| Sylables | 0.39 | 0.45 | 0.55 |
| Unigrams | 0.34 | 0.47 | 0.58 |

**Table 9: Results of logistic regression based on word embeddings. This is by far our most successful model. Note that not all words in our binomials were found in the word embeddings, leaving about 70–97% of the binomials usable.**

|  | Accuracy |
|---|---|
| r/nba | 0.70 |
| r/nfl | 0.75 |
| r/Conservative | 0.83 |
| r/Libertarian | 0.87 |
| r/politics | 0.74 |
| r/The_Donald | 0.70 |
| r/food | 0.85 |

**Table 10: Prediction results for a number of statistics about players in the NBA and NFL. A '-' denotes that the statistic is unavailable for the sport.**

|  | r/nba | r/nfl |
|---|---|---|
| Alphabetical | 0.49 | 0.48 |
| Num Chars | 0.40 | 0.40 |
| Unigram Freq | 0.50 | 0.49 |
| Years Playing | 0.56 | 0.65 |
| Team Mentions | 0.70 | 0.83 |
| Games Played | 0.54 | 0.62 |
| PER | 0.51 | - |
| VORP | 0.51 | - |
| Games Started | - | 0.60 |
| Linear Model | 0.56 | 0.51 |

a number of subreddits that are affiliated with a particular NBA team, with the purpose of allowing discussion between fans of that team. This implies that most users in a team subreddit are fans of that team. We are then able to look for lists of NBA teams by name, city, and abbreviation. We found 2520 instances of the subreddit

**Table 11: If two sports subreddits are listed in a sports subreddit, the subreddit of origin (r/nba in top row, r/nfl in bottom row) usually comes first, in terms of the weighted token evaluation (number of occurrences in parentheses). A '-' means that there are fewer than 30 such lists.**

|  | r/nba | r/nfl | r/hockey | r/soccer | r/baseball |
|---|---|---|---|---|---|
| r/nba | - | 0.73 (204) | 0.73 (34) | 0.44 (43) | - |
| r/nfl | 0.61 (204) | - | 0.66 (35) | 0.60 (75) | 0.76 (30) |

**Table 12: Political name ordering by party across political subreddits. Note that r/politics is left-leaning.**

|  | (D,D) | (D,R) | (R,D) | (R,R) |
|---|---|---|---|---|
| r/politics | 3167 | 2263 | 1980 | 4442 |
| r/Conservative | 313 | 260 | 225 | 610 |
| r/Libertarian | 342 | 311 | 235 | 612 |

team coming first, and 1894 instances of the subreddit team coming second. While this is not a particularly strong predictor, correctly predicting 57% of lists, it is one of the strongest we found, and a clear illustration of the Proximity Principle.

We can do a similar calculation with subreddit names, by looking between subreddits. While the team subreddits are not large enough for this calculation, many of the other subreddits are. We find that lists of subreddits in r/nba that include 'r/nba' often start with 'r/nba', and a similar result holds for r/nfl (Table 11).

While NBA team subreddits show a fairly strong preference to name themselves first, this preference is slightly less strong among sport subreddits, and even less strong among politics subreddits. One potential factor here is that r/politics is a more general subreddit, while the rest are more specific — perhaps akin to r/nba and the team subreddits.

### 5.3 Political Names

In our case, political names are drawn from every congressperson (and their nicknames) in both houses of the US Congress through the 2018 election. It is worth noting that one of these people is Philadelph Van Trump. It is presumed that most references to 'trump' refer to Donald Trump. There may be additional instances of mistaken identities. We restrict the names to only congresspeople that served before 1801 or after 1999, also including 'trump'.

One might guess that political subreddits refer to politicians of their preferred party first. However, this was not the case, as Republicans are mentioned first only about 43%–46% of the time in all subreddits (Table 12). On the other hand, the Proximity Principle does seem to come into play when discussing ideology. For
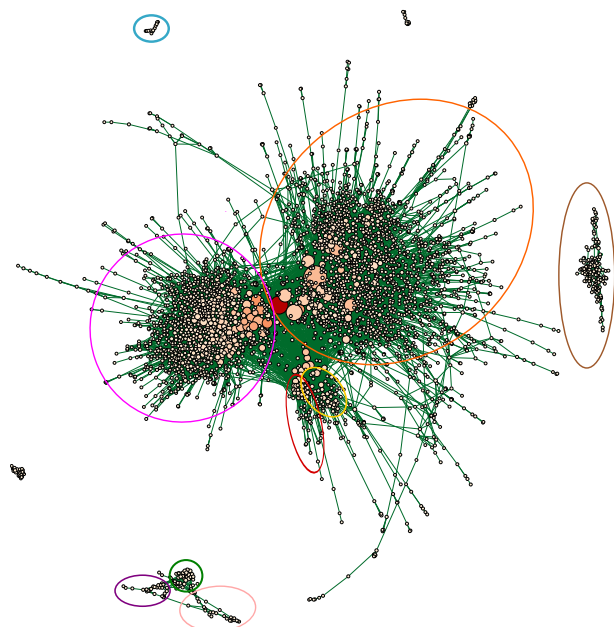
**Figure 7: The r/nba binomial graph, where nodes are words and directed edges indicate binomial orientation.**

instance, r/politics — a left-leaning subreddit — is more likely to say 'democrats and republicans' while the other political subreddits in our study — which are right-leaning — are more likely to say 'republicans and democrats'.

Another relevant measure for lists of proper nouns is the ratio of the number of list instances containing a name to the unigram frequency of that name. We restrict our investigation to names that are not also English words, and only names that have a unigram frequency of at least 30. The average ratio is 0.0535, but there is significant variation across names. It is conceivable that this list ratio is revealing about how often people are talked about alone instead of in company.

## 6 GLOBAL STRUCTURE

We can discover new structure in binomial orderings by taking a more global view. We do this by building directed graphs based on ordinality. In these graphs, nodes are words and an arrow from A to B indicates that there are at least 30 lists containing A and B and that those lists have order [A,B] at least 50% of the time. For our visualizations, the size of the node indicates how many distinct lists the word appears in, and color indicates how many list instances contain the word in total.

If we examine the global structure for r/nba, we can pinpoint a number of patterns (Fig. 7). First, most nodes within the purple circle correspond to names, while most nodes outside of it are not names. The cluster of circles in the lower left are a combination of numbers and years, where dark green corresponds to numbers, purple corresponds to years, and pink corresponds years represented as two-digit numbers (e.g., '96'). On the right, the brown circle contains adjectives, while above the blue circle contains heights (e.g., 6'5"), and in the two circles in the lower middle, the left contains

cities while the right contains team names. The darkest red node in the center of the graph corresponds to 'lebron'.

We can also consider the graph as we change the threshold of asymmetry for which an edge is included. If the asymmetry is large enough, the graph is acyclic, and we can consider how small the ordinality threshold must be in order to introduce a cycle. These cycles reveal the non-global ordering of binomials. The graph for r/nba begins to show cycles with a threshold asymmetry of 0.97. Three cycles exist at this threshold: ['ball', 'catch', 'shooter'], ['court', 'pass', 'set', 'athleticism'], and ['court', 'plays', 'set', 'athleticism'].

Restricting the nodes to be names is also revealing. Acyclic graphs in this context suggest a global partial hierarchy of individuals. For r/nba, the graph is no longer acyclic at an asymmetry threshold of 0.76, with the cycle ['blake', 'jordan', 'bryant', 'kobe']. Similarly, the graph for r/nfl (only including names) is acyclic until the threshold reaches 0.73 with cycles ['tannehill', 'miller', 'jj watt', 'aaron rodgers', 'brady'], and ['hoyer', 'savage', 'watson', 'hopkins', 'miller', 'jj watt', 'aaron rodgers', 'brady'].

Figure 8 shows these graphs for the three political subreddits, where the nodes are the 30 most common politician names. The graph visualizations immediately show that these communities view politicians differently. We can also consider cycles in these graphs and find that the graph is completely acyclic when the asymmetry threshold is at least 0.9. Again, this suggests that, at least among frozen binomials, there is in fact a global partial order of names that might signal hierarchy. (Including non-names, though, causes the r/politics graph to never be acyclic for any asymmetry threshold, since the cycle ['furious', 'benghazi', 'fast'] consists of completely frozen binomials.) We find similar results for r/Conservative and r/Libertarian, which are acyclic with thresholds of 0.58 and 0.66, respectively. Some of these cycles at high asymmetry might be due to English words that are also names (e.g. 'law'), but one particularly notable cycle from r/Conservative is ['rubio', 'bush', 'obama', 'trump', 'cruz'].

## 7 MULTINOMIALS

Binomials are the most studied type of list, but trinomials — lists of three — are also common enough in our dataset to analyze. Studying trinomials adds new aspects to the set of questions: for example, while binomials have only two possible orderings, trinomials have six possible orderings. However, very few trinomials show up in all six orderings. In fact, many trinomials show up in exactly one ordering: about 36% of trinomials being completely frozen amongst trinomials appearing at least 30 times in the data. To get a baseline comparison, we found an equal number of the most common binomials, and then subsampled instances of those binomials to equate the number of instances with the trinomials. In this case, only 21% of binomials are frozen. For trinomials that show up in at least two orderings, it is most common for the last word to keep the same position (e.g., [a, b, c] and [b, a, c]). For example, in our data, ['fraud', 'waste', 'abuse'] appears 34 times, and ['waste', 'fraud', 'abuse'] appears 20 times. This may partially be explained by many lists that contain words such as 'other', 'whatever', or 'more'; for instance, ['smarter', 'better', 'more'] and ['better', 'smarter', 'more'] are the only two orderings we observe for this set of three words.

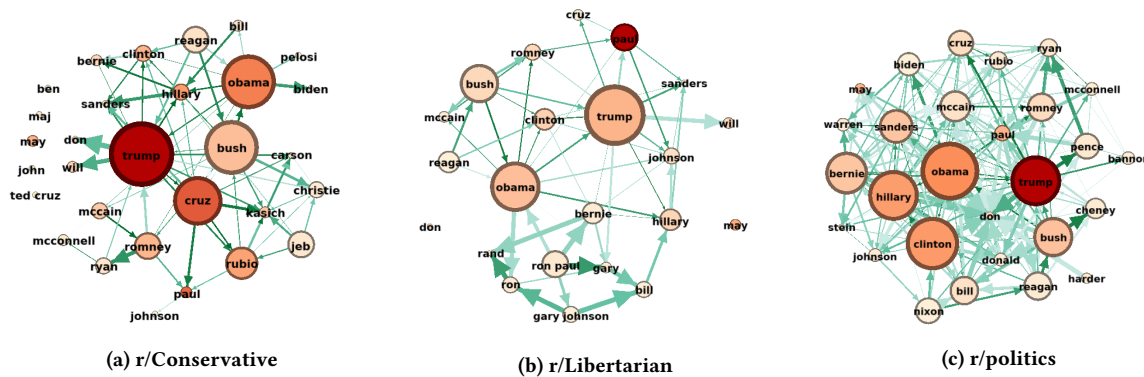(a) r/Conservative     (b) r/Libertarian     (c) r/politics

**Figure 8: Graphs of some of the 30 most common names in r/Conservative, r/Libertarian, and r/politics. Nodes are names, and an edge from A to B represents a list where the dominant order is [A,B]. Node size is the number of lists the word comes first in, node color is the total number of lists the node shows up in, edge color is the asymmetry of the list.**

Additionally, each trinomial [a, b, c] contains three binomials within it: [a, b], [b, c], and [a, c]. It is natural to compare orderings of {a, b} in general with orderings of occurrences of {a, b} that lie inside trinomials. We use this comparison to define the *compatibility* of {a, b}, as follows.

*Definition 7.1.* **Compatibility** Let {a, b} be a binomial with dominant ordering [a, b]; that is, [a, b] is at least as frequent as [b, a]. We define the *compatibility* of {a, b} to be the fraction of instances of {a, b} occurring inside trinomials that have the order [a,b].

There are only a few cases where binomials have compatibility less than 0.5, and for most binomials, the asymmetry is remarkably consistent between binomials and trinomials (Fig. 9). In general, asymmetry is larger than compatibility — this occurs for 4569 binomials, compared to 3575 where compatibility was greater and 690 where the two values are the same. An extreme example is the binomial {'fairness', 'accuracy'}, which has asymmetry 0.77 and compatibility 0.22. It would be natural to consider these questions for tetranomials and longer lists, but these are rarer in our data and correspondingly harder to draw conclusions from.
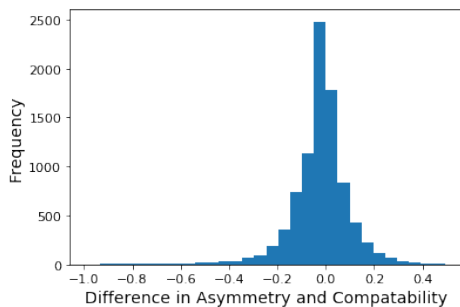


**Figure 9: Histogram of difference in asymmetry and compatibility for binomials within trinomials on r/politics.**

## 8 DISCUSSION

Analyzing binomial orderings on a large scale has led to surprising results. Although most binomials are not frozen in the traditional

sense, there is little movement in their ordinality across time or communities. A list that appears in the order [A, B] 60% of the time in one subreddit in one year is likely to show up as [A, B] very close to 60% of the time in all subreddits in all years. This suggests that binomial order should be predictable, but there is evidence that this is difficult: the most common theories on frozen binomial ordering were largely ineffective at predicting binomial ordering in general.

Given the challenge in predicting orderings, we searched for methods or principles that could yield better performance, and identified two promising approaches. First, models built on standard word embeddings produce predictions of binomial orders that are much more effective than simpler existing theories. Second, we established the Proximity Principle: the proper noun with which a speaker identifies more will tend to come first. This is evidenced when commenters refer to their sports team first, or politicians refer to their party first. Further analysis of the global structure of binomials reveals interesting patterns and a surprising acyclic nature in names. Analysis of longer lists in the form of multinomials suggests that the rules governing their orders may be different.

We have also found promising results in some special cases. We expect that more domain-specific studies will offer rich structure.

It is a challenge to adapt the long history of work on the question of frozen binomials to the large, messy environment of online text and social media. However, such data sources offer a unique opportunity to re-explore and redefine these questions. It seems that binomial orderings offer new insights into language, culture, and human cognition. Understanding what changes in these highly stable conventions mean — and whether or not they can be predicted — is an interesting avenue for future research.

# REFERENCES

[1] 2019. *the @unitedstates project.* https://theunitedstates.io/
[2] Richard D. Abraham. 1950. Fixed Order of Coordinates: A Study in Comparative Lexicography. *The Modern Language Journal* 34, 4 (1950), 276–287. http://www.jstor.org/stable/318913
[3] Keith Allan. 1987. Hierarchies and the choice of left conjuncts (with particular attention to English). *Journal of linguistics* 23, 1 (1987), 51–77.
[4] Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2009. On the Syllabification of Phonemes. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Boulder, Colorado, 308–316.
[5] Sarah Bunin Benor and Roger Levy. 2006. The chicken or the egg? A probabilistic analysis of English binomials. *Language* (2006), 233–278.
[6] Dwight L Bolinger. 1962. Binomials and pitch accent. *Lingua* 11 (1962), 34–44.
[7] William E. Cooper and John Robert Ross. 1975. Word Order. In *Papers from the Parasession on Functionalism*, Robert E. Grossman, Jame L. San, and Timothy J. Vance (Eds.). Chicago Linguistic Society, Goodspeed Hall, 1050 East 59th Street, Chicago, Illinois 60637, Chapter 6, 63–111.
[8] Gertraud Fenk-Oczlon. 1989. Word frequency and word order in freezes.
[9] Marita Gustafsson. 1976. The frequency and "frozenness" of some English binomials. *Neuphilologische Mitteilungen* (1976), 623–637.
[10] P Hagerty. 2015. Ladies and gentlemen: Word order and gender in English. *The expression of gender* (2015), 69–86.
[11] Peter Hegarty, Nila Watson, Laura Fletcher, and Grant McQueen. 2011. When gentlemen are first and ladies are last: Effects of gender stereotypes on the order of romantic partners' names. *British Journal of Social Psychology* 50, 1 (2011), 21–35.
[12] Otto Jespersen. 1905. *Growth and Structure of the English Language*. B.G. Teubner. 233–234 pages.
[13] Selin Kesebir. 2017. Word order denotes relevance differences: The case of conjoined phrases with lexical gender. *Journal of personality and social psychology* 113, 2 (2017), 262.
[14] Yakov Malkiel. 1959. Studies in irreversible binomials. *Lingua* 8 (1959), 113–160.

[15] Janet L McDonald, Kathryn Bock, and Michael H Kelly. 1993. Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology* 25, 2 (1993), 188–230.
[16] William J. McGuire and Claire V. McGuire. 1992. Psychological Significance of Seemingly Arbitrary Word-Order Regularities: The Case of Kin Pairs. In *Language, Interaction and Social Cognition*, Gün R. Semin and Klaus Fiedler (Eds.). SAGE Publications Ltd, London, Chapter 11, 214–236.
[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[18] Sandra Mollin. 2012. Revisiting binomial order in English: Ordering constraints and reversibility. *English Language & Linguistics* 16, 1 (2012), 81–103.
[19] Heiko Motschenbacher. 2013. Gentlemen before ladies? A corpus-based study of conjunct order in personal binomials. *Journal of English Linguistics* 41, 3 (2013), 212–242.
[20] Amandianeze Nwana and Tsuhan Chen. 2017. QUOTE:" Querying" Users as Oracles in Tag Engines-A Semi-Supervised Learning Approach to Personalized Image Tagging. *IEEE MultiMedia* (2017).
[21] Amandianeze O Nwana and Tsuhan Chen. 2016. Who ordered this?: Exploiting implicit user tag order preferences for personalized image tagging. In *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–6.
[22] Steven Pinker and David Birdsong. 1979. Speakers' sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior* 18, 4 (1979), 497–508.
[23] Fred Newton Scott. 1913. The Order of Words in Certain Rhythm-Groups. *Modern Language Notes* 28, 8 (1913), 237–239. http://www.jstor.org/stable/2916276
[24] SheMedia. 2019. *RedditMetrics*. https://redditmetrics.com/
[25] Włodzimierz Sobkowiak. 1993. Unmarked-before-marked as a freezing principle. *Language and Speech* 36, 4 (1993), 393–414.
[26] Saundra Wright and Jennifer Hay. 1975. Fred and Wilma: A Phonological Conspiracy. In *Gendered Practices in Language*, Sarah Benor, Mary Rose, Devyani Sharma, Julie Sweetland, and Qing Zhang (Eds.). CSLI Publications, Stanford, California, Chapter 9, 175–192.
[27] Saundra K Wright, Jennifer Hay, and Tessa Bent. 2005. Ladies first? Phonology, frequency, and the naming conspiracy.