# The Chicken or the Egg?
# A Probabilistic Analysis of English Binomials

Sarah Bunin Benor
Hebrew Union College – Jewish Institute of Religion
3077 University Avenue
Los Angeles, CA 90007
sbenor@huc.edu

Roger Levy
University of Edinburgh
School of Informatics
2 Buccleuch Place
Edinburgh EH8 9LW
United Kingdom
rlevy@inf.ed.ac.uk

# The Chicken or the Egg?
# A Probabilistic Analysis of English Binomials

Abstract. Why is it preferable to say salt and pepper over pepper and salt? Based on an analysis of 692 binomial tokens from on-line corpora, we show that a number of semantic, metrical, and frequency constraints contribute significantly to ordering preferences, overshadowing the phonological factors that have traditionally been considered important. The ordering of binomials exhibits a considerable amount of variation. For example, although principal and interest is the more frequent order, interest and principal also occurs. We consider three frameworks for analysis of this variation: traditional Optimality Theory, stochastic Optimality Theory, and logistic regression. Our best models – using logistic regression – predict 79.2% of the binomial tokens and 76.7% of types, and the remainder are predicted as less-frequent – but not ungrammatical – variants.[*]

[*] We would like to thank Paul Kiparsky, Joan Bresnan, Elizabeth Traugott, Chris Manning, Gerhard Jäger, Susanne Riehmann, Emily Bender, Art Owen, and Brady Clark for their advice on this paper. Thanks also to the editors and anonymous reviewers for their meticulous reading and helpful suggestions.

1. Introduction. Which comes first, the chicken or the egg? This timeless question may have originated in biology and philosophy, but it is also relevant to linguistic theory. In this paper we address the issue of binomial formation, the process by which a language user determines the ordering of like-category conjoined items in a three-word phrase of the form A and B (e.g. chicken and egg). Existing research using experimental, intuition-based, and corpus-based methods suggests that many factors can play a role under the right conditions, including the semantic relationship between the items, metrical and other phonological properties of the possible orderings, and relative item frequency. What remains poorly understood, however, is exactly how these factors interact, and how salient these factors are in naturally occurring data. We address these questions in this paper.

1.1 Previous Research. A number of scholars have worked on binomials, producing overarching theories of their ordering and small-scale studies of particular constraints. In 1959 Malkiel wrote an overview of the phenomenon of ‘frozen binomials’ – binomials that occur nearly exclusively in one order – including several semantic and phonological principles of their ordering and reasons for their freezing. Bolinger (1962) focused on binomial constructions that are not necessarily frozen. He posited a metrical/phonological explanation for their ordering and backed it up with experimental evidence.

Cooper and Ross (1975) conducted an extensive analysis of frozen binomials and posited overarching semantic (‘Me First’) and phonological (A is smaller than B) constraints. They suggested that the semantic constraints outrank the phonological ones. Fenk-Oczlon (1989) posited a different overarching constraint on the ordering of binomials: the more frequent item precedes the less frequent item. She explains most of the previously posited constraints (e.g. ‘Me First’, vowel quality, number of initial consonants) in these terms. Based on an analysis of 400 frozen binomials in English and German, she determines that the new rule accounts for the ordering of significantly more frozen binomials than any other rule. Most exceptions to the frequency rule can be accounted for by iconic ordering, she found.

More recently, studies have used experimental techniques to investigate the ranking of the various factors in binomial ordering. McDonald et al. (1993) gave experimental evidence that a semantic constraint (animacy) ranks above a metrical one (syllable count). Müller (1997) conducted an Optimality Theory (OT) analysis of frozen binomials in German and found that semantic constraints outrank metrical constraints, which outrank other phonological constraints. Finally, Wright and Hay (2002) studied male and female names and found that male names were significantly more likely to have what they call ‘first-position phonology’, which includes many of the phonological principles posited by Cooper and Ross. They also found that gender (essentially a semantic factor) was more important in respondents’ order preferences than phonology.

Despite this abundance of previous research on binomial ordering, a number of important questions are left unanswered. Several existing studies (Malkiel, Cooper and Ross, Fenk-Oczlon, and Müller) have dealt only with fixed, or frozen, binomials. Furthermore, most existing corpus-based studies have ignored the relationships between constraints. One notable exception is Levelt and Sedee (2004), who found results similar to Müller’s in a stochastic OT analysis of naturally occurring Dutch binomials.

The current study is corpus-based, considers frozen and non-frozen binomials, and deals with how the various constraints interact. This type of inquiry is important for several reasons. First, it is sometimes difficult to determine whether a particular binomial is frozen. As Malkiel explains, even completely irreversible binomials, which would be unidiomatic in the reverse order (e.g. odds and ends), were likely once reversible. If a binomial seems to be in the process of freezing, would we consider it frozen or not? If it is impossible to come up with a definitive list of currently frozen binomials in a language, how can we analyze them in a statistically sound way?

Second, studying only frozen binomials is theoretically problematic. If the principles posited by Cooper and Ross (1975) to govern frozen binomial order are productive, they should also be visible among unfrozen binomials. Third, by looking at a large number of binomials that occur within a fixed corpus, this study allows for quantitative analysis. Malkiel admits that his 'impressionistic pilot study dispenses with any binding statistical computation of frequency' (1959:118) and suggests further research. Cooper and Ross agree:

> Strong support [for the constraints] can only be provided by sampling a very large number of such pairs and stating the statistical probabilities of . . . certain regularities, and, of at least equal importance, the relative strengths of these regularities (1975:79).

Now that we have easy access to millions of words via corpus searches, the studies suggested by Malkiel and Cooper and Ross are more practical.

Gustafsson's (1976) study is corpus-based, and includes unfrozen binomials. However, although she gives important information on frequency and word class, she does not discuss semantic and phonological constraints. Levelt and Sedee (2004) – the only other corpus study that investigates not-necessarily-frozen binomials from a quantitative perspective – uses Internet searches to find frequencies of Dutch binomials. They find that a stochastic OT ranking (Boersma and Hayes 2001) of semantic and phonological constraints accounts for a majority of these binomials.

A number of experimental studies have also investigated unfrozen binomials, focusing primarily on the role of phonological constraints. Bolinger (1962) looks at word combinations like cold and obvious and strong and bitter, as well as nonsense words like plap and plam and briff and brip. Oakeshott-Taylor (1979) tests the order of three-segment words with ten different vowels, such as pit and peat. Wright and Hay (2002) determine how men and women order names like Ben and Karen and Brooke and Bridget. And McDonald et al.'s (1993) experiment uses stimuli like dog and telephone and attorney and desk. If the findings of these experiments are correct, they should extend to naturally occurring binomials as well.

## 2. Definitions and methodology

2.1 Definitions. Malkiel defines a binomial as 'the sequence of two words pertaining to the same form-class, placed on an identical level of syntactic hierarchy, and ordinarily connected by some kind of lexical link' (1959:113). We follow this definition, but for purposes of this paper, we limit the possible lexical links to and. The positions of words within the binomial will be called Slot A and Slot B, and the words themselves the A

Item and B Item or simply A and B. In the binomial salt and pepper, salt is in Slot A, and pepper is in Slot B. A frozen binomial is defined as a pair that occurs almost exclusively in a specific order. An ordered binomial token is a situated instance of the pieces of a binomial in text or speech; a surface (binomial) type refers to an ordered form A and B (not in context), and an input (binomial) type refers to an unordered pair of words {A,B} that can appear together in an ordered binomial.

We use the term constraint to refer to some semantic, metrical, frequency-based, phonological, orthographic, or other feature of a binomial that may influence the order of a binomial. A constraint is active for a binomial if the constraint favors one order over the other; otherwise, it is inactive. An active constraint is aligned with a binomial token if it favors the token's order; if it favors the opposite order, it is aligned against the token. A token is semantically (metrically/frequentistically/phonologically) aligned if some semantic (metrical/frequency/phonological) constraint is aligned with or against it. In the binomial carefully and prudently, for example, carefully is the more frequent word, so a constraint favoring more frequent words in Slot A would be aligned with the binomial.

2.2 Methodology. This study consisted of four stages: a corpus search for binomial tokens; formulation of phonological, semantic, and frequency-based constraints; coding and quantitative analysis of the binomial tokens found; and formulation of three complete models of the data, cast in traditional OT, stochastic OT, and logistic regression.

The corpus search was conducted on three tagged corpora: the Switchboard (spoken), Brown (varied genres, written), and Wall Street Journal (WSJ; newspaper) sections of the Penn Treebank III, available from the Linguistic Data Consortium (Marcus et al. 1993; http://www.ldc.upenn.edu/). These corpora were searched for constructions of N and N, V and V, Adj and Adj, and Adv and Adv, where both X and X were part of the same XP.

The search yielded 3,680 distinct binomials. Using the beginnings and ends of each corpus's search results, we took a total of 411 input binomial types – distinct sets {A,B} for some binomial sequence A and B – for analysis. This total consisted of 120 nouns, 103 verbs (including gerunds and participals), 118 adjectives, and 70 adverbs. We did not include binomials formed from personal names, because idiosyncratic factors frequently determine the ordering of names in a conjunction (however, we did not exclude the names of political entities such as countries or states). We discarded binomials formed with extender phrases, such as and stuff, as they are not in theory reversible (i.e. politics and everything cannot be everything and politics).

For each of these binomials, we noted whether we considered each to be frozen (for example, by and large and north and south are frozen; honest and stupid and slowly and thoughtfully are not). We then searched for all occurrences of each binomial and its reverse in all three corpora, and included all such occurrences in our final corpus, yielding 692 tokens. Like Gustafsson (1976), we found that very few of the binomials occurred more than once in the three corpora. Most of those that did are frozen binomials, such as back and forth, which occurred 49 times.

Throughout this paper, we assume that every corpus instance of a binomial was generated as follows. First, the speaker/writer determines the individual words constituting the binomial, as well as the context surrounding the binomial. Given the words and context, the speaker/writer then chooses an order in which to produce the words. We make no assumptions about the conscious/unconscious nature of this decision; we are solely interested in the relationships between a variety of semantic, metrical, phonological, and pragmatic factors and the chosen order.

3. Constraints. In this section, we consider twenty semantic, pragmatic, metrical, phonological, and word-frequency factors that may affect the ordering of binomials, including several that have been previously suggested in the literature and some new ones that we predict based on linguistic principles.

3.1 Semantic-Pragmatic Constraints. Previous literature discusses a number of semantic constraints that affect the ordering of binomials, including 'animate > inanimate,' 'male > female,' 'positive > negative,' and 'alcoholic > non-alcoholic' ('>' means 'precedes'). Cooper and Ross (1975) organize their semantic constraints into 19 categories and then reduce almost all of them to one umbrella principle, called 'Me First'. This constraint says that 'first conjuncts refer to those factors which describe the prototypical speaker.' The Me in Me First is personified by Archie Bunker,[1] who, according to Cooper and Ross, is Here, Now, Adult, Male, Positive, Singular, Living, Friendly, Solid, Agentive, Powerful, At Home, Patriotic, General (he is a stereotype), and a count noun (1975:67). Malkiel (1959) identifies two categories: 'precedence of the stronger of two polarized traits' and 'priorities inherent in the structure of a society'. In the binomials in our corpus, we identified four semantic constraints, based on this previous literature and related linguistic research published since then. The first two are similar to Malkiel's 'precedence' and the third and fourth are similar to his 'priorities'. The first constraint involves formal linguistic properties, and the next three involve real-world knowledge. Formal Markedness. Cooper and Ross mention markedness and even briefly consider using it as the umbrella concept for the semantic constraints (66-7). The concept of markedness stems back to the Prague School of Linguistics. Jakobson discussed it with regard to oppositions in Russian morphology, 'where one of the terms of the opposition signifies the presence of a certain quality and the other (the unmarked or undifferentiated term . . . ) indicates neither its presence nor its absence' (1984 [1939]:153). Jakobson showed how this concept applies to semantics, as in the pair dévuška 'girl/virgin' (marked) and devíca 'girl' (unmarked). This is clearly relevant to the current paper, as pairs of words are often opposed in a relationship of markedness. An example is pull and tug. 'Pull' is more general in manner than 'tug,' as 'tug' indicates the presence of a quality not necessarily present in 'pull': sudden and quick. It is clear that 'pull' is the unmarked of the pair.

Markedness is relevant to binomial ordering because of Markedness Assimilation (Andersen 1972), the tendency for marked elements to occur in marked contexts and unmarked elements to occur in unmarked contexts. Along the same lines as given information preceding new information, it is logical that the less marked item of a pair would appear in the first slot, Slot A.

The concept of markedness has been defined in several different ways, but in this paper, we restrict our use to a narrow definition. Out of the criteria for markedness discussed by Battistella (1990), we have chosen to use four:

(1) Criteria for lower formal markedness: Less marked items tend to

(i) have a broader, more general meaning
(ii) have greater freedom of distribution
(iii) have a larger number of subcategorical distinctions
(iv) be structurally more simple

The first three qualities apply, for example, in the ordering of flowers and roses, as a rose is a specific type of flower. They also apply in changing and improving, as one can change without improving but not vice versa, and in first and only, as something can be first without being only but not vice versa. Markedness is violated in alterations and sewing, as sewing can include quiltmaking, needlepoint, and alterations, while alterations are a more restricted type of sewing (cf. the occurring binomial sewing and quilting).[2] The fourth quality in (1) applies when one member of a binomial contains a greater amount of semantically potent derivational morphology than the other. It has two incarnations. The first is the absolute case, such as in complete and unabridged, where the items have no shared derivation but one item, unabridged, has a negation morpheme, whereas the other item, complete, does not. In the second, relative case, one item is actually derived from the other, as in poetry and non-poetry and linguistic and paralinguistic. We group the relative case with the general formal markedness constraint, as instances of formal markedness as evaluated by other criteria involve semantic properties of binomials relative to one another; whereas we consider the absolute case an independent semantic constraint. This leaves open the possibility that absolute and relative semantic markedness may on occasion be in opposition to one another. Criterion (iv) also applies in binomials where one item is defined by or discussed in relation to the other, as in there and elsewhere (which appears twice), and in cases where one item is a precondition for the other, as in accept and hire and sewing and alterations.

Perception-Based Markedness. The elements in a binomial sometimes exist in a simple formal relationship determinable by linguistic properties. But more commonly they are in a complex relationship that can be perceived only through extra-linguistic, real-world knowledge. Cooper and Ross's Me First principle describes this important phenomenon: that qualities of prototypical people tend to occur in Slot A. In order to ground our judgments for this constraint independently of actually observed binomial order, we turn to Mayerthaler's (1988 [1981]) research on markedness. Based on experiential evidence, he considers certain properties to be semantically less marked, including the following:

(2)

| Less marked | More marked |
|---|---|
| animate | inanimate |
| singular | plural |
| right | left |
| positive | negative |
| concrete | abstract |
| front | back |
| above | below |
| vertical | horizontal |

He considers the less marked elements to be more closely connected to or more easily perceptible by the speaker. He gives explanations for each category. Since a speaker is animate, singular, and most likely right-handed, elements with these qualities are less marked. Since, according to Mayerthaler, a speaker 'has a positive image of himself' (10), positive items are less marked. Since concrete items are perceptually more accessible, they are less marked. Since a speaker has eyes in her head not in her feet, looks forward rather than backward, and stands upright, elements that are front, above, and vertical are less marked. Using this same reasoning, Mayerthaler also argues that proximal ('here') is more marked than distal ('there'), since one sees others more than oneself (9). While we accept most of Mayerthaler's arguments, we reject his view on the proximal/distal dichotomy and consider proximal as less marked than distal. This decision follows Cooper and Ross (1975), for whom proximal before distal is an important component of analysis. They also discuss proximity in relation to football games (e.g. Harvard students will be more likely to say 'the Harvard-Yale game,' and Yale students will be more likely to say 'the Yale-Harvard game'). We found several supporting examples in our corpus, including Public and International (where 'public' refers to domestic affairs) and here and abroad.

While several of the oppositions in this constraint can be determined according to biological orientation, it should be pointed out that some are also culturally constructed. What some people consider positive or concrete others might consider negative or abstract. And, of course, what is proximal in some cases is distal in others, depending on the vantage point of the speaker. Although we consider the positive > negative markedness distinction to be a linguistic universal, what constitutes positive or negative for any particular speech community is a matter of cultural construction.

Along the same lines as formal markedness, the element that is perceptually less marked for the speaker is more of a given and is more likely to occur in Slot A. Examples include deer and trees and people and soils (animate and inanimate), individually and cumulatively (singular and plural), physical and mental (concrete and abstract), up and down and head and tail (above and below), and high and inside (vertical and horizontal). Since one likely notices age, which is to some extent visually discernible, sooner than mental qualities such as wisdom, we also considered this constraint to apply in older and wiser. There are several common binomial types where the constraint is violated, including back and forth, backward and forward, and left and right.

We found several binomials that we judged to be in a relationship on the dimension of positive and negative, according our understanding of the value judgments of the majority of Americans. Examples include good and bad, honest and stupid, and science and angst.

In (4) below we present several other examples that we judged to involve perceptual markedness, together with brief explanations:

(3) north and south (north is the orienting direction on a compass)
mother and dad (mother is usually more central to the child's upbringing)
day and night (humans usually spend more waking hours during the day)
see and hear, seen and felt (seeing is a more salient form of perception)
oranges and grapefruit, salt and pepper (the former is generally more common)
ugly and bad (ugliness is visible; badness is discernible by moral judgment)

family and friends (family is more central)

There may seem to be a potential for overlap between Formal Markedness and Perception-Based Markedness. In fact, van Langendonck (1986) and Mayerthaler (1988 [1981]) would probably combine the two categories, as they discuss them together in their work. These two concepts are certainly related and can in some cases be used interchangeably, but as Battistella explains, they differ enough to remain separate. In this paper, we separate Formal Markedness from Perception-Based Markedness, because the latter involves perception and real-world knowledge, while the former involves formal linguistic properties.

We also found one trend closely related to perceptual prominence that merits mention as a possible subconstraint: adjectives of temperature precede adjectives of humidity, a pattern apparently unnoticed in previous literature. This subconstraint was unviolated in our corpus and was satisfied by hot and dry (twice) and cold and wet. Preliminary quantitative investigation suggests it is robust: in X and Y searches of the 100-million-word British National Corpus, the temperature-before-humidity order was preferred by ratios of 38:8 for hot/dry, 3:2 for hot/wet, 7:6 for cold/dry, and 50:26 for cold/wet. The numbers for our own binomials corpus are small, however, and we have no proposal for an independent psychological explanation for a temperature > humidity markedness pattern, so we leave further examination of this pattern to future work.

Power. Another constraint that involves real-world relations is Power. This constraint, which stipulates that the more powerful element appears in Slot A of a binomial, encompasses Malkiel's category of 'priorities inherent in the structure of a society'. Malkiel includes gender pairs such as guys and dolls, husband and wife, and Mr. and Mrs; asymmetrical age pairs, such as mother and child; pairs of ruling class and ruled, such as prince and pauper and rich and poor; and animacy pairs, such as man and beast, cat and mouse, and horse and buggy. In all of these binomials, the more powerful element precedes the less powerful. Of course, what is more powerful is determined by subjective values and may differ in various communities.

The Power Constraint predicts that in a mixed-gender pair the man will come first, as in son and daughter and men and women. It also applies to other pairs of items where one is considered more important or central in our society, such as salt and pepper, oranges and grapefruit, and gold and silver. Another incarnation of the Power Constraint is the condiment rule: in complementary pairs, the element perceived as central precedes the element perceived as a side dish, sidekick, or condiment. This applies to food, people, and other things, as in eating and drinking, clergymen and parishioners, and principal and interest. Although we did not include names in our corpus, the condiment rule would also apply in Don Quixote and Sancho Panza, Groucho and Harpo, and Clinton and Gore. Finally, the Power Constraint is also involved in contrasts on a scale of intensity, as in cruel and unusual, where cruel is more powerful, or intense, than unusual.

Iconic/scalar sequencing. When two elements are perceived as existing in a sequence, chronological or otherwise, they should appear in that same sequence within a binomial. Malkiel (1959:146) discusses the temporal aspect of this constraint, including frozen binomials such as wait and see and kiss and tell. Cooper and Ross (1975) also mention it, saying, 'in a freeze of two verbs which are intended to be in temporal sequence, the place 1 verb denotes the earlier action’ (102). Fenk-Oczlon (1989) posits a similar constraint that accounts for both temporal and spatial relationships. She says that this constraint accounts for almost all frozen binomial ordering not accounted for by frequency (see section 3.3). In our corpus, the iconic constraint was particularly common in verbal binomials like slowed and stopped and manufacture and install. It also applies in sequences that are context-specific, such as cooked and shelled, referring to seafood preparation. This constraint also applies to adjectives and adverbs reflecting a chronological or cause-and-effect sequence, such as there and back (one cannot come back before going there); out and about (one must first go out in order to go about); and unconstitutional and severable (the rider restricting the president’s Article II powers was only severable because it was unconstitutional). Finally, it includes numeric and level values, such as eighth and ninth and elementary and high (school), and other items that are considered to exist on a scale, such as months and years and nights and weekends. The iconic sequencing constraint sometimes contrasts with the power constraint, as items that are more intense on a scale may also be considered more powerful.

Comparing the semantic constraints. Cooper and Ross’ ‘Me First’ was an attempt to provide an overarching principle for almost all of the constraints. This is a useful umbrella concept for several factors, but it cannot include all semantic constraints. Although it could include Perception-Based Markedness and some aspects of Power, it cannot include Iconic Sequencing or Formal Markedness. It seems to stretch the categories too far to say that Archie Bunker is more connected to ‘two-sevenths’ than to ‘three-sevenths’ or to ‘pull’ than to ‘tug.’ The same is true for some aspects of the Power Constraint. Contrary to Cooper and Ross’ predictions, Archie Bunker is more similar to ‘patients’ and ‘parishioners’ than to the more powerful ‘psychiatrists’ and ‘clergymen.’

In addition, Me First may be somewhat more relevant for fixed binomials than for non-fixed ones. A binomial may become fixed if it does not violate the Me First Principle for the prototypical speaker, but a naturally occurring non-fixed binomial may be uttered by a speaker who does not fit the parameters of Archie Bunker (e.g. a woman). For this reason we consider it important to divide up binomials that would be listed under Me First. Those that are likely common to most humans are listed under Perception-Based Markedness, and those that are determined by power relations in our society are listed under Power.

Another reason to keep these four semantic constraints separate is the conflicts that arise among them. The fact that some binomials violate one constraint but satisfy another is evidence that the constraints are distinct. For example, the binomial mother and dad violates the Power Constraint, which prefers the male in Slot A, but it satisfies Perception-Based Markedness, as the mother is typically more central to the child than the dad. We also see a satisfaction of Iconicity and violation of Power in harass and punish, as punish is more powerful than harass, but on a scale of increasing intensity harassment precedes punishment. Therefore Power must be listed separately from Iconic Sequencing. Although conflicts among semantic constraints are clearly possible, they turn out to be rare, as we discuss in Section 4.

In addition to these semantic constraints, binomials can be ordered by what we call a ‘set, open construction’, where the A item can exist together with many different B items. These include ‘sit and _____,’ as in sit and wait and sit and think; and ‘good and _____,’ as in good and ready and good and plenty. These binomials may fall under other categories, but we believe the most salient constraint affecting their ordering is the fact that their constructions are conventionalized. The corpus included 17 of these, presented in (4):

(4)
good and thick (x2)
go and vote (x2)
went and voted
went and hid
nice and sunny (x2)
nice and fresh
nice and relaxed
nice and small
nice and toasty
sit and wait
sitting and staring
sitting and watching
sat and cried
try and catch

Finally, a variety of external syntactic and word-order factors can affect binomial ordering; we lump these under the heading of pragmatic constraint. In binomials consisting of modification adjectives, for example, we found that when one item is more closely related to the modified noun, it is preferred in the slot closer to the noun. An example in our corpus is sane and productive (member of society): it is more common to say a 'productive member of society' than a 'sane member of society'. Word order in a neighboring phrase is at work in a token of music and comedy, whose order mimics 'musical comedy' in the following sentence: 'I admit that going back to Ralph Waldo Emerson for humor is like going to a modern musical comedy for music and comedy.' We found 35 binomials satisfying some pragmatic constraint discernible within the sentence, while we found none that clearly violates an intrasentential pragmatic constraint.

Coding Methodology. Evaluating a binomial for the applicability of semantic constraints is a necessarily subjective process. To minimize the possibility of bias marring our judgments of semantic constraints, each co-author independently judged each binomial in the corpus for application and violation of each semantic constraint. We then discussed at length those binomials for which our judgments differed and made final decisions together. Furthermore, it is crucial to evaluate the binomial as it appears in the context of the corpus; for example, examination of the context may reveal whether elements of a binomial appear in chronological and therefore iconic order. We therefore examined the binomials within the sentence in which they occurred. These two processes meant that data coding was quite time-consuming, but our intuition (confirmed by this study) was that semantic constraints were so common and strong that it would be dangerous to attempt analysis of non-semantic factors while ignoring semantic factors. Although the proportion of binomials for which our judgments disagreed on some semantic constraint was not insubstantial (about 10% of the corpus), in only two cases had we judged a single constraint to be active in opposite directions for a single binomial.[3] Discussion revealed that in these as well as most other cases, differences in our judgments resulted from details of the interpretations of semantic constraint descriptions, and we were able to reach final agreement without difficulty. In the few cases where our disagreements persisted, we classified the constraints to be inactive. We also found that semantic constraints were for the most part uncorrelated with phonological constraints (although see Section 4.4 for a possible correlation between perceptual markedness and open main syllables), further suggesting that our semantic judgments were not unduly biased based on phonological factors.

3.2 Metrical Constraints. We coded for a number of metrical constraints, based on previous literature and our own hypotheses.

*A>B (Syl#) – A should not be longer than B.
Many studies about the ordering of binomials claim that the number of syllables is the main metrical constraint at work (e.g. Cooper and Ross 1975, Pinker and Birdsong 1979). A short-before-long preference is also widely known to exist in other aspects of English word order variation (see Wasow 2002 for recent work and references). We therefore hypothesized that longer items would tend to follow shorter items in our corpus.

*Lapse (*ww) – The binomial should not have more than one consecutive weak syllable between strong ones.

According to Selkirk (1984), there is a constraint against more than 2 consecutive weak syllables. As Nespor and Vogel (1989) argue, based on Selkirk (1984:52), 'Any weak position on a metrical level may be preceded by at most one weak position on that level.' Green and Kenstowicz (1995) present this constraint in the framework of Optimality Theory and call it *www.

The present paper is the first corpus study to investigate lapse in the ordering of naturally occurring binomials. In experimental work, McDonald et al. (1993) investigated the interaction between length and lapse in recall and ordering preferences. Although lapse avoidance seemed to help experimental subjects recall binomials more than short-before-long ordering, short-before-long ordering but not lapse had a significant effect on ordering preferences. These inconclusive results call for further investigation. In coding for this constraint, we considered a binomial to violate the lapse constraint if its maximum number of consecutive weak syllables was higher than the maximum number of weak syllables in its reverse (hence the *ww constraint). Note that our criterion is therefore more stringent than Selkirk's, as a binomial with only two consecutive weak syllables can be in violation if its reverse has no consecutive weak syllables. This seemed appropriate, because many binomials differ by one weak syllable, as in fuzzy and warm vs. warm and fuzzy. We did not consider syllables with secondary stress to be weak. For example, complete and unabridged has only one weak syllable between stressed ones when we take into account the secondarily stressed un syllable: wS w swS.

*Ultimate Stress of B – B should not have ultimate stress (abbreviated as *BStr).

Müller (1997:23) accounts for metrical tendencies in German binomials not with a lapse constraint but with 'Foot-Accent' and 'Word-Accent'. He argues that the active constraint is that the main stress of the B item should be on its penult. Looking at English binomials, Bolinger (1962) posits that oxytonic (ultimate) stress will be uncommon in the B element, because the binomial is often followed by a word with a stressed initial syllable. He gives experimental evidence for this, using speakers' judgments of non-fixed binomials that precede a noun. It is possible that stress on the final syllable of a binomial phrase is uncommon, as there is a universal yet violable constraint against word-final stress (Anttila 1997:51). We expected this constraint to be somewhat active in the ordering of binomials.

3.3 Frequency Constraints. Fenk-Oczlon (1989) provides convincing evidence that word frequency plays an important role in the ordering of binomials: the more frequent item precedes the less frequent item in a binomial. Research on lexical access gives this constraint a transparent psychological motivation: latency—the amount of time that a person takes to name an object presented in a picture— is lower for more-frequent words, as shown by Oldfield and Wingfield (1965) and Wingfield (1968) in an object-naming task for English, and replicated by Jescheniak and Levelt (1994) for Dutch. In Fenk-Oczlon's study, 84% of binomials were consistent with this constraint, a higher proportion than for any other constraint. Accordingly, we hypothesized that more frequent words would tend to precede less frequent words in the binomials of our corpus. To measure word frequency in our dataset, we used the number of occurrences in the corpus from which the individual binomial was culled.

One potential problem with simple word counts is if the frequency of a specialized usage or sense of a word in the binomial is poorly reflected by the frequency of the wordform. For example, in English and Americans, the words had frequencies of 61 and 27 respectively. However, these numbers include for English the meaning of the language in addition to the people. Another example is wiry and fit, whose frequencies in the WSJ corpus were 2 and 32 respectively, including, of course, more meanings for fit than simply 'in shape'. This may not be a problem, as people's lexical access might be different for different uses of a word or for homophonous words.

Fenk-Oczlon's frequency information included multiple word forms for one stem. However, since many of the words in our dataset were already derived, or polymorphemic (e.g. hurtling and plunging, sleepily and friendlily), we searched only for the exact forms. This procedure actually allowed for this constraint to predict some forms that would not have otherwise been predicted, as in the following binomials:

| (5) | Binomial | Frequency of (A and B), (B and A) |
|---|---|---|
| | math and sciences | 16, 1 |
| | science and math | 47, 16 |

Another explanation for the tendency of derived forms to appear in the B slot is that derived forms are generally longer than the non-derived form in the A slot. See Fenk-Oczlon (1989) for more discussion of the connection between word frequency and metrical and other constraints.

3.4 Non-Metrical Phonological Constraints. In this section we discuss the non-metrical phonological constraints included in our analysis. These constraints interact with metrics in various ways. As Müller observes, the B item of a binomial tends to be more stressed than the A item. We argue that the tendency toward greater stress cannot be attributed solely to phrase-final lengthening, as some researchers have suggested. We show how a number of other possible phonological constraints could follow from the greater stress of B.

Because phrase-final lengthening leads to longer vowels and more heavily stressed syllables, a number of studies have mentioned phrase-final lengthening in their explanations of certain constraints. The stimuli in Oakeshott-Taylor's (1984) and Gustafsson's (1974) experiments were free-standing binomials, so they were necessarily phrase-final. But we must ask if naturally occurring binomials also occur in phrase-final position. Samples from our corpus indicate that although adjectival, adverbial, and nominal binomials are almost exclusively phrase-final, about two-thirds of verbal binomials precede a constituent they govern in their own phrase. For example, the binomial in (6a) ends its phrase, but the one in (6b) governs a following NP.

(6a) I do [$_{NP}$ a lot of [$_{NP}$ cross-stitching and painting]] (Switchboard)
(b) Those persons who were lucky enough [$_{IP}$ to [$_{VP}$ [$_{V}$ [$_{V}$ see] and [$_{V}$ hear]] [$_{NP}$ the performance of his work]]] (Brown)

However, the NP complement of the compound verb is so long that see and hear likely forms its own phonological phrase (Nespor & Vogel 1986). In this case, phrase-final lengthening may be the reason that hear is more stressed than see.

We also investigated those non-phrase-final verbal binomials whose following constituents consist of a single word, and which therefore are not likely to undergo phrase-final lengthening. First, we extracted all verbal binomials with a right sister from the Treebank.[4] Among these, we found that the immediately following constituent was a single word long 20.0%, 15.2%, and 25.5% of the time in the WSJ, Brown, and Switchboard sections of the Treebank respectively, for a total of 87 tokens. These included verb phrases such as:

(7a) sitting and staring silently (Brown)
(b) check and discipline himself (Brown)
(c) owns and operates hotels (WSJ)
(d) attract and train ringers (WSJ)
(e) see and do things (Switchboard)

In (b) and (e) the final word in the phrase seems to be included in the phonological phrase of the binomial, but the main stress of the phrase is on the B item of the binomial. Even in cases such as (a),(c), and (d), where the final word in the phrase does receive the main stress of the VP, there is a clear tendency for greater stress among binomial items on B than on A. This greater stress would likely be manifested as a lengthened syllable or a contrast in pitch or volume (Ladd 1996). Therefore, we conclude that the greater stress of B must be a quality of binomial phrases, independent of phrase-final lengthening. Although we do not have recordings of these sentences to verify this point acoustically, a quick attempt to stress the A item more than the B item (nów and agàin) shows that a phrase with this stress pattern sounds less like a binomial and more like a content word with a discourse marker.  We see this reverse stress pattern in smog and stuff and politics and everything, which are, as discussed above, not included in our study.

As a final note, although we believe the evidence is strong that greater stress on B is an intrinsic property of binomials rather than an epiphenomenon of phrase-final lengthening, this is not crucial to the larger picture of deriving phonological constraints from the tendency toward greater stress on B.  Even if the tendency toward greater stress on B was not intrinsic, we would expect that phonological factors favoring the stress of one item in a binomial input would be more harmonically realized if that item were placed in the B slot.  We examine a number of such phonological factors in the remainder of this section.

Vowel Length.  Several linguists have argued that vowel length affects the ordering of binomials, saying that B should have a longer vowel.  Gustaffson's experiment (1974) shows that the B item is almost always rendered longer in duration than A, and Oakeshott-Taylor (1984) shows that the vowel of B is almost always lengthened.  In an experiment designed to test factors in isolation, Pinker and Birdsong (1979) found a significant preference for B to have a longer vowel.  They attribute this preference to ease of processing, as the item with longer phonetic material will be harder to process.  We expected that longer vowels would be preferred in B, but we attribute this to B's greater stress, rather than to ease of processing.  The English stress system is partly based on syllable weight, which is determined by vowel length and coda.  Therefore, a longer vowel would likely be attracted to the stressed position.

In coding for vowel length, we used the following 2-way phonemic distinction, as diphthongs tend to pattern with long vowels in English:

(8) short vowels:  æ, ɛ, ɪ, ʌ, ʊ
long vowels, including diphthongs:  ɑ, e, i, o, u, ɔ, æʊ, aɪ, aʊ, ṛ, Vr

Syllabic [ṛ] was considered long, as it can form a word-final syllable of its own.  And Vr combinations were considered diphthongs, following Veatch's (1991) finding that /r/ patterns as a glide and is part of the preceding vowel.[5]  Front vowels before [ŋ] were considered short (the vowel in pinks was considered [ɪ], not [i]).  We considered items in a binomial to differ in phonemic vowel length if one item had a short main vowel and the other had a long main vowel or diphthong.

To further investigate the question of vowel length and binomial order, we tried another measurement criterion, using intrinsic phonetic duration instead of phonemic length.  To determine vowels' phonetic length, we followed Crystal and House's (1988) calculations of mean intrinsic duration of American English vowels.  They analyzed vowels from several speakers' slow and fast readings of a set passage and calculated the length of these vowels in various environments: primary stress, secondary stress, unstressed.  Since all of the vowels we are coding are in primary stressed syllables, we used Crystal and House's values for primary stressed vowels:

(9) Inherent duration (mean for several speakers) of primary stressed vowels (in ms)
Short:

| | |
|---|---|
| ɪ | 75 |

| | |
|---|---|
| ʊ | 85 |
| ʌ | 103 |
| ɛ | 106 |
| Long: | |
| i | 119 |
| u | 126 |
| e | 136 |
| ɑ | 140 |
| ɔ | 148 |
| æ | 159 |
| o | 162 |
| Diphthongs: | |
| aɪ | 172 |
| aʊ | 202 |
| ɔɪ | 298 |
| Rhotic: | |
| r̩ | 123 |

For purposes of coding, we grouped these vowels into six groups. Although Vr combinations are not included in Crystal and House's study, we included them in the diphthong group.

(10) Groups of vowels used in coding, arranged from shortest to longest

1. ɪ, ʊ (range of inherent duration: 71-90 ms)
2. ʌ, ɛ (91-110 ms)
3. i, r̩, u (111-130 ms)
4. e, ɑ, ɔ (131-150 ms)
5. æ, o (151-170 ms)
6. aɪ, aʊ, ɔɪ, Vr (171+ ms)

Note that phonetic vowels length differs from phonemic vowel length mostly in the number of distinctions made. But one vowel, [æ], is in a much different location in the two measurements. Although [æ] is phonetically long, it patterns phonologically with short vowels.

We coded the binomials according to these groups. For example, in greasy and dirty the vowels are of equal length ([i], [r̩]), and in sane and productive A's vowel ([e]) is longer than B's ([ʌ]) . If the vowels of A and B are in the same group but one is followed by a voiced coda consonant and the other is followed by a voiceless coda consonant, we coded the pre-voiced-coda vowel as longer. Examples are: hit and killed, big and thick, and down and out.

Vowel Backness. Several scholars say that Slot B's main vowel should be backer than Slot A's main vowel. Cooper and Ross's data for this constraint come mostly from coordinate words without the conjunctive link, such as flimflam and zigzag, many of which are stressed on the first element. Pinker and Birdsong (1979) and Pordany (1986) disagree with this constraint, arguing that vowel height has more of an effect. But Oakeshott-Taylor (1984) provides experimental evidence that backer vowels are preferred in Slot B. He tested British and South African subjects' ordering preferences for nonsense-word binomials where the only difference between the two words was the vowel quality, and he found that backness had a significant effect.

However, we see no phonological reason for a preference. It is possible that backness plays a role in experiments only because speakers have in mind similar lexical items, many of which place the backer item second. And this may be due to a confounding of frontness and height. Binomials like spic and span or beck and call, and even compounds without the conjunction, such as riff-raff, exhibit a preference for B to have a vowel that is both backer and lower. We expected no independent preference for backer main vowels in the B slot. We used the following scale to determine vowel backness alignment:

(11) u, o, ɔ, ʌ, ʊ, r̩, ɑ > æ, ɛ, e, ɪ, i

Vowel Height. Pordany (1986:124) argues that vowel height is more important than vowel backness in determining the ordering of binomials. He gives little evidence for this claim, basing it on only a few examples from English, Hungarian, and German. Pinker and Birdsong (1979) give experimental evidence for a cross-linguistic preference for B to have a lower vowel. Similarly, Müller (1997), looking at German binomials, says that high vowels precede low vowels and that, among vowels of the same height, backer vowels go first. However, Oakeshott-Taylor's experiment found that vowel height had no effect on the ordering. Our hypothesis was that low vowels would be preferred in the B slot for reasons of greater stress. This is in line with Anttila's (1997) findings that Finnish stems ending in lower vowels prefer the strong variant of the genitive plural, which gives preference to endings that are heavier and stressed. We used the following scale for determining vowel height alignment:[6]

(12) i, u, ɪ, ʊ > e, o, ɛ, ɔ, ʌ, ə > æ, a

Initial Consonants. Much of the literature, starting with Cooper and Ross (1975), assumes that the B item is more likely to have more initial consonants. Cooper and Ross base this constraint on word pairs without 'and,' as well as on a few binomials of the form A and B, such as fair and square and sea and ski. Wright and Hay (2002) disagree with this constraint. They point out that there is more phonetic motivation for an initial cluster to be disfavored in the B slot. A cluster there creates an even longer sequence of consonants because it immediately follows 'and' (which is likely reduced to [n̩]). They cite examples such as flora and fauna, in which the A item has the larger initial consonant cluster, but violates the semantic constraint of more animate before less animate. Is this a case where the (possible) trend for B not to have an initial consonant cluster outranks other phonological and semantic constraints? Wright and Hay's experiment, in which participants were asked to order pairs of names, finds a weak (but insignificant) preference for cluster-initial names to be preferred in the A Slot.

We predicted no trends for alignment with initial consonant cluster differences, as this factor is not related to the greater stress of B. In coding, we used Cooper and Ross's formulation, so that a binomial is in alignment if B has more initial consonants than A. But we ignored differences when both items had two or more initial consonants (so that the constraint was aligned with cauliflower and broccoli, but inactive for stress and pressure).

Final Consonants. Cooper and Ross (1975) say that the B element of a binomial should have fewer final consonants, based on a few examples, such as wax and wane and betwixt and between. However, Pinker and Birdsong (1979) give experimental evidence to the contrary. They found a marginally significant trend for more final consonants to be preferred in the B slot.

Going along with other expected phonological trends, a preference in this category would also be related to weight and stressability. Since there is an overall tendency for greater stress on the B item than the A item, we hypothesized that among binomials where B has ultimate stress, there would be a preference for B to have a coda, which would allow for increased main stress on B.

Bolinger (1962) actually tested the reverse prediction in an experiment using monosyllabic nonsense words. His hypothesis was that the B item in a binomial should be as 'open and sonorous as possible' (35) and therefore that A would more likely have a coda than B. He tested this with stimuli where the two words differed only by their coda (e.g. stee and steet or steet and stee). Although he did not present his results for this issue in particular (as it is combined with issues of sonority), he did include all of the response data in his paper. An analysis of the responses for this factor in Bolinger's study finds that there is a slight preference for B to have zero consonants (i.e. respondents preferred broat and broe over broe and broat). However, when we further divide up these data according to the voicing of the final consonant, an interesting pattern emerges. In those stimuli where one of the items ends in a voiceless consonant, respondents preferred that item in the A slot. In those stimuli where one of the items ended in a voiced consonant, respondents preferred that item in the B slot, but the trend was weak and not significant.

This difference can be explained by the fact that voiced tautosyllabic consonants lengthen a preceding vowel. If vowel length has an effect on the ordering of binomials, then the stimuli that include voiced codas are confounding two factors: number of word-final consonants and vowel length. Examining only the stimuli with voiceless codas, we find that there is a strong preference for B to have no coda consonants. Alternatively, one might explain the length difference between words like hit and hid in a different way: that a vowel is shortened by a tautosyllabic voiceless consonant. If this was the case, then we would say that the stimuli with the voiceless codas are confounding two factors, and we would want to examine only the stimuli with voiced codas. Then we would find a slight but insignificant trend in accordance with our hypothesis, contrary to Bolinger's: in a binomial where the items have no coda and a voiced coda, the voiced coda is preferred in the B slot. We pursue this hypothesis further in Section 4.4.

To determine final consonant alignments, we compared the number of final consonants in the A and B items in our corpus, ignoring differences when both items had two or more final consonants.

Openness of Stressed Syllable. For the same reason, we expected words with closed main syllables to be preferred in the B slot (when the main syllables of A and B are not both open or both closed). We treated openness and closedness as a binary property of the main syllable.

In coding for openness, we considered syllables with short vowels followed by ambisyllabic consonants (including flaps) to be open. For example, in rainy and icky, both words are equal in openness of the stressed syllable. Following Veatch (1991, chapter 3), we considered intervocalic glides (/w/, /j/, /r/) to be tautosyllabic but not making a syllable closed. In coding for openness and syllable weight, we considered /æ/ to be a short vowel. Finally, in syllable weight, more than one coda consonant was considered extra-metrical and therefore as not adding weight.

Another question arose often: should inter-vocalic consonant clusters be considered in the previous syllable, the following syllable, or divided between the two? We answered this question with the concept of maximization of the onset: any cluster that could be word-initial in English is considered to be the onset of the following syllable. For example, the [bl] in reestablish was considered an onset to the [ɪ] syllable, but the [g] in magnified was considered a coda to the [æ] syllable. Sometimes maximization of the onset applied even across morphological boundaries, as in push-ups and sit-ups, where the following vowel is not preceded by a glottal stop.

Syllable Weight. Although previous research has made no claims about syllable weight, our analysis of binomial stress patterns leads to a prediction. Since syllable weight is a major determinant of stress in English, and the B element of a binomial has a stronger stress, we would expect B's main syllable to be heavier than A's. We coded syllable weight differences assuming three levels of heaviness:

(13) Heaviness scale for openness of syllable weight
Not heavy: rhymes with short vowels followed by ambisyllabic consonants (e.g. eliminate, scabrous)
Heavy: rhymes with short vowels followed by tautosyllabic consonants (e.g. bender, merry) and rhymes with long vowels or diphthongs (including Vr) and no coda consonant (e.g. maybe, farmer)
Extra-heavy: rhymes with long vowels or diphthongs and a coda consonant (e.g. remainder, suits)

Using the term not heavy, rather than light, does not conflict with the phonological system of English (Kager 1989), in which stress is quantity sensitive, and it also preserves the important distinction between words like tenor and tender.[7]

The coding assumed ambisyllabicity for consonants that follow short vowels, and it considered [æ] to be short, as it patterns phonologically with short vowels despite its phonetic length. Following Veatch (1991), intervocalic glides (j, w, r) were considered tautosyllabic, and the nucleus of a diphthong was considered short. More than one coda consonant was not considered to add weight. We considered a binomial to be aligned with Weight iff the main syllable of B is heavier than that of A.

Initial Segment Sonority. Cooper and Ross (1975) say the initial segment of A will be more sonorant than the initial segment of B. Most of their examples are binomials without the link (e.g. roly poly, jeepers creepers). Pinker and Birdsong (1975) give experimental evidence that this is a trend for English speakers but not for speakers of other languages. We cannot think of a phonetic reason for this, so we hypothesized that there would be no significant difference. We used the following scale to determine differences in initial segment sonority:[8]

(14) vowels, ʔ > h > j > w > r > l > nasals > fricatives > stops

Final Segment Sonority. Several scholars agree that the final segment's sonority should be greater in the B item. Our hypothesis agreed with this idea, because a more sonorous final segment may lead to a more lengthenable final syllable. We used the same scale as for initial segment sonority.

Issues in coding for phonological constraints. An important question that arose during the coding was whose speech variety to follow. The speakers and writers who contributed to the Brown, Switchboard, and Wall Street Journal corpora come from all around the US and perhaps other countries as well. They may have had different pronunciations of the words that make up these binomials, and these differences may affect the ordering. In the coding, we followed Veatch's (1991) 'Reference American' abstraction, which combines the most common aspects of 'standard' American English dialects. For example, this would consider the first syllable of orange to be [or], while some northeasterners would say [ar]. And it would consider been to have the vowel [ɪ] , while some midwesterners would say [ɛ] .

3.5 Alphabetical Order. Because the majority of our corpus is from written sources, it is possible that binomial ordering is influenced by the alphabetic location of the first letter of each word. We expected that this would have more of an effect in the ordering of names, such as business partners or joint authors. Nonetheless, we investigated alphabetic order in our corpus.

3.6 Summary of constraints tested. In short, our analysis included 20 constraints:
Semantic Constraints:

RelForm: Relative Formal Markedness: B should not be less marked than A.
Icon: Iconic Sequencing: If A and B exist in an iconic sequence, they should appear in that sequence.
Power: B should not be more powerful than A.
Percept: Perception-Based Markedness: B should not be less marked than A.
Pragmatic (determined by context)
Set Open Construction (e.g. sit and wait, go and vote)

Metrical Constraints:

*ww: Lapse (2 consecutive weak syllables) is not allowed in the binomial as a whole (takes secondary stress into account)
*A>B: A should not have more syllables than B.
*BStr: B should not have ultimate (primary) stress.

Word Frequency Constraint:

Freq: B should not be more frequent than A (determined according to individual corpus)

Non-metrical Phonological Constraints:

VPhonemic: A’s main stressed vowel should not be longer than B’s main stressed vowel – phonemic, 2 levels of distinction:
phonemically short vowels:  æ, ɛ, ɪ, ʌ, ʊ
phonemically long vowels and diphthongs:  ɑ, e, i, o, u, ɔ, æʊ, aj, oj, r̩, Vr

VPhonetic: A’s main stressed vowel should not be longer than B’s main stressed vowel – phonetic, 6 levels of distinction:
ɪ, ʊ < ʌ, ɛ < i, u, r̩ < e, ɑ, ɔ < æ, o < aɪ, aʊ, ɔɪ, Vr

Backness: A’s main stressed vowel should not be backer than B’s.
u, o, ɔ, ʌ, ʊ, r̩, ɑ > æ, ɛ, e, ɪ, i

Height: B’s main stressed vowel should not be higher than A’s.
i, u, ɪ, ʊ > e, o, ɛ, ɔ, ʌ, r̩ > æ, ɑ

CInit: A should not have more initial consonants than B (more than 2 are considered 2).

CFin: A should not have more final consonants than B (more than 2 are considered 2).

Openness: The primary stressed syllable of B should be closed.

Weight: A’s main stressed syllable should not be heavier than B’s.

SonorInit: The initial segment of B should not be more sonorous than the initial segment of A.
vowels, glottal stop > h > j > w > r > l > nasals > fricatives > stops

SonorFin: The final segment of A should not be more sonorous than the final segment of B.
vowels, glottal stop > h > j > w > r > l > nasals > fricatives > stops

Alphabetic Constraint:
Alpha: The first letter of B should not precede the first letter of A alphabetically.

## 4. Findings

INSERT Table 1 ABOUT HERE

Table 1 presents the satisfaction rates for each individual constraint, together with the number of binomials for which each constraint was active.  The percentage reported is the proportion of binomials that are aligned with the constraint among those for which that constraint is active.  Note that several constraints were found to be violated more often than satisfied (i.e. percentages below 50%), although most of these cases were not statistically significant.  Furthermore, detailed analysis (see below) controlling for constraint correlation reveals no evidence that any constraint truly tends to align against binomials.

Semantic, metrical, and frequency constraints are all significantly aligned with binomial order. Phonological factors turned out to be less consistent: at the level of types, weight, vowel backness, and (marginally) openness of stressed syllable are significantly aligned against binomial order, contrary to previous studies and linguistic evidence; other phonological constraints are uncorrelated with binomial order. In the next several sections we describe trends among constraints in greater detail. We report alignment trends of a constraint in terms of the proportion $\pi_{active}$ of binomials active for that constraint that are aligned with the constraint; *p*-values for these proportions are derived from the null hypothesis of the binomial distribution with parameter ½. We also report associations between many constraint pairs, using two measures. First, two constraints may tend toward or against being active for the same constraints. We report this association using the odds ratio, $\theta$, for constraint activity, and *p*-values are given using Fisher's exact test (Agresti 2002).[9] Second, when two constraints are both active, they may tend toward or against aligning in the same direction. We report this association as the proportion $\pi_{align}$ of same alignment, and calculate *p*-values using the null hypothesis of the binomial distribution with parameter ½. $\theta$ and $\pi_{align}$ are calculated from counts of surface binomial types. In all cases, we consider the conclusions that can be drawn from surface type counts more reliable than those drawn from token counts; token counts are easily skewed by a small number of common frozen binomials, such as back and forth (N=49).

4.1 Semantic constraints. All of our proposed semantic constraints are significantly aligned with binomial order (Table 1). Of these, Iconic Sequencing was the strongest and most frequently active, applying to 77 binomial types (128 tokens), and violated by only two instances of one type: interest and principal, which we judged to violate the constraint because principal causally (and temporally) precedes interest. Even for this input type, there are five reverse tokens of principal and interest, which is aligned with the Iconic Sequencing constraint. Perception-Based Markedness was the next most prevalent, and (with Relative Formal Markedness) the next strongest constraint; violations include always and everywhere, where the less concrete time word precedes the more concrete space word, and animals and humans, where the item less like the speaker precedes the item more like the speaker. Relative Formal Markedness and Power were similar in frequency of activity, with Power being the weakest semantic constraint, satisfied in only 18 of the 26 types ($p$<0.1) and 44 of the 72 tokens ($p$<0.05) to which it applied. Absolute formal markedness is satisfied by one binomial, complete and unabridged, and violated by one, non-instinctive and conscious.

In all, 288 binomial tokens, comprising 144 surface types, satisfied at least one semantic constraint. 102 tokens, composed of 23 types, violated at least one semantic constraint; this count was dominated by two binomial types – back and forth and black and white, occurring 49 and 19 times respectively and violating Perception and Power constraints respectively. No pair of semantic constraints was significantly correlated. Furthermore, as mentioned in Section 3.1, only four binomial types involved satisfaction of one semantic constraint and violation of another, as show below:

(15) Opposition of semantic constraints (constraint pair SATISFIED/VIOLATED):
harass and punish, satisfying Iconic Sequencing and violating Power
mother and dad, satisfying Perception and violating Power
hope and pray, satisfying Relative Formal Markedness and violating Power
unconstitutional and severable, satisfying Iconic Sequencing and violating Relative Formal Markedness

We conclude that semantic constraints in general are quite common within our corpus, being active in over one-third of surface types. When active, they are usually satisfied; only for the Power constraint was the trend for satisfaction somewhat questionable (*p*<0.05).

4.2 Metrical constraints. Since semantic constraints are so strong and pervasive, it is necessary to account for them when determining metrical constraints. This section includes counts both of all binomials and of only those where no semantic constraint is satisfied.

All of the metrical constraints – *A>B (short before long), *ww (avoid lapse), and *BStr (avoid final stress) – show highly significant (*p*<0.001) trends toward satisfaction. The constraint with the strongest satisfaction profile is *BStr, showing 76% token (70% type) satisfaction including and 83% token (78% type) satisfaction excluding semantically aligned binomials.[10] Bolinger's *BStr constraint thus seems to be the most powerful of the three metrical constraints we investigated. However, it should be noted that *BStr is also the most rarely active metrical constraint; only 170 tokens are affected, while 337 and 307 tokens of *A>B and *ww are affected. Müller's more restrictive constraint, that B should have penultimate stress, is not satisfied quite as frequently. Of the 164 surface binomial types in which ordering affects whether B has penultimate stress, an insignificant majority (90, or 55%) have penultimate stress. Even this small majority may be misleading, as well: when binomials with no monosyllabic item are excluded, only 38 of the 91 remaining types (42%) have penultimate stress, an insignificant departure from randomness but nevertheless raising the possibility that any overall trend toward penultimate stress could be an epiphenomenon of *A>B.

INSERT Table 2 ABOUT HERE

While there are strong trends toward satisfaction for all three metrical constraints, they are not all independently active. As can be seen in Table 2, all three metrical constraints are strongly intercorrelated. For each pair of metrical constraints, both constraints tend to be active for the same binomials (Table 2, left-hand side), and among those binomials for which they are active, both constraints tend to be aligned in the same direction (Table 2, right-hand side). Further investigation shows that, with only two exceptions (foot-loose and fancy-free and follow and understand), *BStr is never opposed to either *ww or *A>B, and 73% of input types have identical alignment profiles for *A>B and *ww. *A>B and *ww conflict for 38 tokens (26 types); conflict between these two constraints is discussed in section 5 below.

Only 96 tokens in our corpus (73 types) have neither active semantic constraints nor active metrical constraints (these include tokens such as caring and loving, substantial and persistent, aggressive and persistent, and straight and hard).

4.3 Frequency. As can be seen in Table 1, frequency differentials were almost always involved between items in our binomials dataset, and there is a highly significant (*p*<0.001) trend for more frequent items to precede less frequent items, whether or not semantically aligned binomials are excluded. The rate of constraint satisfaction is highest when both semantically and metrically aligned binomials are excluded: 93 tokens (73 types) are frequency-differentiated, and 62% (68%) have a more frequent A, significant at *p*<0.025 (*p*<0.01). This result shows that frequency is a useful indicator of binomial ordering, and is most reliable for those binomials that are not influenced by semantic or metrical factors. Nevertheless, frequency proves less reliable in our study than Fenk-Oczlon (1989) found for frozen binomials: she found a constraint satisfaction rate of 84% of frozen binomials, whereas in no case does the proportion of constraint satisfaction for our dataset exceed 68%. We suggest that at the time a given binomial froze, it must have had a strong array of constraints favoring one order over the other. This means that any given active constraint is less likely to be aligned against a frozen binomial than against an unfrozen binomial. Although frequency difference is not an inviolable determinant of binomial ordering, it is applicable to nearly all binomials and therefore turns out to be an important component of the multiple-constraint models in Section 5.

As Fenk-Oczlon points out, word frequency is closely connected with semantic and metrical constraints. Table 3 shows the significant correlations of the frequency constraint with semantic and metrical constraints for our data.

INSERT Table 3 ABOUT HERE

Most notably, frequency alignment is strongly correlated with *A>B, consistent with the general principle that more frequent words tend to be shorter (Zipf 1949). Frequency is also strongly correlated with Bolinger’s constraint against final stress, (*BStr) and marginally correlated with avoidance of lapse (*ww), but further investigation indicates that these are likely to be an artifact of the correlation among metrical constraints. Of the binomial types where frequency and *BStr both have non-neutral alignment, *A>B does not share alignment with *BStr in only six cases (*BStr is aligned with frequency in four of these cases, against it in two). And while there are 18 types in which *A>B is inactive and both frequency and *ww have non-neutral alignment, frequency and *ww are actually negatively (though not significantly) correlated in these cases. This suggests that the genuine connection is between frequency and *A>B, and the correlation of frequency with *ww and *BStr is an artifact of mutual correlation with *A>B. The correlations of frequency with both relative semantic markedness and perceptual markedness seem to be direct and understandable: the most frequent forms tend to be the most semantically general (and thus least marked), and less perceptually marked elements such as here, good, and head also tend to be the ones used more commonly than their binomial sisters such as there, bad, and tail. (Note that relative semantic markedness and perceptual markedness are not correlated in our dataset.)

In summary, we found that frequency, when viewed alone or as secondary to semantic constraints, seems strongly justified as a determinant of binomial ordering, although it is not among the most reliable indicators of binomial order. It is strongly correlated with semantic and perceptual markedness, and it has a tight connection with word length that causes a superficial correlation with other metrical constraints. Only nineteen binomial types in our corpus do not involve a frequency differential, and only two of those (rumbles and smolders and pinks and greens) are also unaligned with any metrical or semantic constraint. However, there are many cases in our corpus covered by at least one metrical or phonological constraint where binomial order is not explained by some combination of these three constraint types, such as economically and physically (violating *A>B) and bottles and cans (violating *ww and *BStr). In the next section we investigate whether non-metrical phonological constraints could explain these remaining data.

4.4 Non-metrical phonological constraints.

Vowel Length. Contrary to our expectations, our corpus data do not provide evidence for a phonemic constraint preferring longer main vowels in the B item. Unsurprisingly, phonetic and phonemic vowel length differentials were highly correlated in our corpus ($\tau$=0.50, $p$<0.001). In the corpus as a whole, there is no significant trend for alignment of main vowel length with binomial order (Table 1). When we exclude semantically aligned binomials, we find a trend against phonemically longer B main vowels, but the trend disappears when we also exclude metrically aligned binomials, although the sample size here is much smaller. We believe the superficial trend against phonemically longer B may be due to a correlation with metrical constraints. Phonemic vowel length alignment is significantly negatively correlated in the complete dataset with *A>B ($\theta$=1.07; $\pi_{align}$=0.41, $p$<0.05), and marginally with *BStr ($\theta$=1.37;$\pi_{align}$=0.38, $p$<0.1); when binomials with active semantic constraints are excluded, the correlation with *A>B disappears but the directional correlation with *BStr remains ($\pi_{align}$=0.36, $p$=0.13), and though it is no longer significant the remaining sample size is small ($n$=36). Since *BStr, when active, is a strong determinant of binomial order, the trend toward longer A main vowels may well be explained by a powerful avoidance of final stress. As a tentative explanation of the negative correlation between longer B vowel and final stress avoidance, we note that in our dataset, among binomial types consisting of one monosyllabic word and one non-final-accent polysyllabic word, there are more types (N=28) where the main vowel of the monosyllabic word is long and that of the polysyllabic word is short than types (N=18) where the main vowel of the monosyllabic word is short and that of the polysyllabic word is long, although a simple binomial test indicates that this difference is only marginally significant ($p$<0.1).

Phonetic vowel length, unlike phonemic vowel length, is not significantly correlated with any metrical constraint, whether or not binomials with an active semantic constraint are excluded. However, when we look at those where no semantic, metrical, or frequency constraint is satisfied, there is one interesting trend: of those that are not equal in phonetic vowel length, 61% of tokens have a longer B vowel ($p$<0.5), although no significant trend was present among types (this binomial subset included two high-frequency frozen binomial types: odds and ends [N=12], which has a phonetically longer A main vowel, and black and white [N=19], which has a phonetically longer B main vowel). This raises the possibility that speakers' choices on binomial order might be sensitive to fine phonetic distinctions in length, although the evidence is inconclusive.

Backness. Contrary to the findings of Cooper and Ross and Oakeshott-Taylor, we found a trend toward backer vowels in A both when all binomials were considered and when semantically aligned binomials were excluded (Table 1). While no clear trend remains when both semantically and metrically aligned binomials are excluded, the remaining sample size is quite small. To further investigate this pattern, we looked at the correlations between backness and other constraints, and found a trend toward negative correlation between backness and *A>B ($\theta$=1.11; $\pi_{align}$=0.39, $p$<0.005 in complete dataset; $\theta$=1.11, $\pi_{align}$=0.41, $p$=0.12 excluding semantically aligned binomials). We therefore looked at alignment with backness excluding only metrically aligned binomials, and found no significant trend: 105 of 195 ($p$=0.25) and 32 of 70 types ($p$=0.40) have backer B. Furthermore, Fisher's exact test did not indicate a significant difference between backness ratios by type between this dataset and the dataset excluding both semantically and metrically aligned constraints. We suspect that in a larger dataset controlling for semantic and metrical alignment, we would not see any trend with respect to vowel backness (we found no significant correlation between frequency and backness once metrically aligned binomials were excluded).

Height. In general, we found no significant alignment of height. Among tokens, the unexpected trend – low vowels preferred in A – was significant in the entire dataset ($p$<0.001) and when both semantic and metrical constraints were excluded ($p$<0.05), and it was close to significance when only semantic constraints were excluded ($p$=0.12). Upon inspection, however, this seems to be due to several common, frozen binomials with lower vowels in A, including back and forth (N=49), men and women (N=15), odds and ends (N=12), and now and then (N=12). Among types, there were no significant trends. We conclude that vowel height has no discernible effect on binomial ordering in our corpus.

Initial consonants. There was no significant trend among binomial types in either direction for initial consonants in the entire dataset, or when semantically aligned binomials were excluded. When only metrically aligned binomials were excluded, there was a significant trend in the remaining dataset to prefer more initial consonants in B: 67 of 109 tokens ($p$<0.01) and 42 of 64 types ($p$<0.025). When both metrically and semantically aligned binomials were excluded, directional trends stayed the same, but significance became marginal for tokens and disappeared for types (see Table 1). Investigation suggests that the significant preference for initial consonant clusters in B is masked in the dataset as a whole by a strong negative correlation with the A>B constraint ($\theta$=0.80, $\pi_{align}$=0.27, $p$<0.001 for entire dataset; $\theta$=0.61,$\pi_{align}$=0.28, $p$<0.001 excluding semantically aligned binomials). This negative correlation probably arises from the fact that monosyllabic English open-class words rarely begin with a vowel.[11] While our data seem to support a preference for the B item to have more initial consonants, we wish to point out that the subset of types (N=33) and tokens (N=40) that are not semantically and metrically aligned is quite small.

Final consonants. We expected that among binomials with final stress, the trend would be toward presence of final consonants, to facilitate greater stress on B. Our general prediction was correct: with the exception of a significant trend toward satisfaction for all tokens, apparently due to the high prevalence (n=49) of back and forth, there were no significant trends for alignment. Excluding only metrically aligned binomials revealed a weak and insignificant trend toward longer coda on B (32 of 55 types), as did excluding both semantically and metrically aligned binomials. Among only binomials with ultimate stress, 15 of 22 tokens (7 of 18 types) have B items with more final consonants. Although these findings are not statistically significant, they suggest that the number of consonants ending the stressed syllable may have an effect on the ordering of binomials.

Openness. Contrary to our expectations, there is a trend for the item with the closed main syllable to appear in the A slot, marginally significant in the entire corpus and when semantically aligned binomials are excluded (Table 1). When we exclude only metrically aligned binomials, a significant token-wise trend appears for B to be closed (48 of 77 tokens, $p<0.05$), though this trend is insignificant by types (24 of 41 types, $p=0.35$). This trend holds when semantically aligned binomials are also excluded, but in both cases the sample size is small and the trend is insignificant. These trends seem to arise from strong correlations between openness and all metrical constraints ($\tau>0.40$, $p<0.001$ in all cases). Final syllables are more likely than non-final syllables to have codas, and most polysyllabic words do not have final stress, so monosyllabic words are more likely than polysyllabic words to have closed main syllables. Some B>A examples are drawers and closets, shock and incredulity, and trade and finance; A>B examples include chicken and egg, science and math, and movie and book. Since words with more syllables tend to occur in the B position, words with closed main syllables tend to occur in the A position. Therefore it is not surprising that metrically aligned binomials – common in the corpus as a whole and when semantically aligned binomials are excluded – are more likely to have a closed A item.

It also turns out that there is a significant correlation between openness and perceptual markedness ($\theta=0.96$; $\pi_{align}=0.82$, $p<0.01$ among all binomials; $\theta=1.97$; $\pi_{align}=0.75$, $p<0.1$ excluding metrically aligned), which seems driven by a few common binomials that involve proximal/distal or directional asymmetries such as now and then, here and abroad, and backwards and forwards. Since the Perception constraint strongly affects binomial order, we also looked at the subset of binomials which are neither perceptually nor metrically aligned. Within this subset, there is no significant trend involving openness. We conclude that openness of stressed syllable has no direct effect on binomial ordering in our corpus.

Syllable weight. When we look at all tokens and just those where no semantic constraint is satisfied, A tends to have a heavier stressed syllable than B, contrary to our prediction that heavier stress would tend to fall on the main syllable of B (Table 1). But when we exclude only metrically aligned binomials, the results reverse: a statistically insignificant majority (45 of 77 types) have a heavier-stressed A. Like openness, syllable weight has a strong negative correlation with all metrical constraints; when only metrically aligned binomials are excluded, there is a significant alignment in the expected direction among tokens (which upon inspection seems skewed due to the most common binomial, back and forth, begin aligned), but no significant trend among types. These results are not surprising, as syllable weight is simply a combination of phonemic vowel length and openness. We conclude that syllable weight does not have an effect on the ordering of binomials here, although a larger, more controlled sample excluding both semantically and metrically aligned binomials would be useful.

Sonority. We found no general trends for alignment with either initial sonority or final sonority.[12] However, among those types where B has ultimate stress and the final-segment sonority of the two items is not equal, 13 of 15 have a more sonorous final segment in A ($p<.001$). This can be explained by the trend for B's stressed syllable to be more closed than A's among those with ultimate stress: openness and final sonority were significantly correlated ($\theta=1.60$, $\pi_{align}=0.60$, $p<0.001$). As described above, openness is in turn correlated with metrical constraints, so the apparent role of sonority among ultimately-stressed B tokens is probably an artifact of metrical effects. Therefore, we have no evidence that sonority acts independently and, do not consider it a factor in the ordering of the binomials in this corpus.

Summary of non-metrical phonological constraints. When the confounding effects of semantic and metrical constraints are controlled for, the only phonological constraint for which we found compelling evidence was a preference for larger initial consonant clusters on B. This is in line with most previous literature, although we find no phonological motivation for this trend.

Our inconclusive results on phonological constraints do not rule them out altogether, and a number of examples in our corpus do suggest that they are still active. For example, it seems that phonetic vowel length may be a factor in the ordering of fully and fairly, correct and acute, economic and demographic, semiconductors and supercomputers, and help and serve. But if phonetic vowel length were a powerful constraint, we might expect not to find made and built or big and thick (where the [ɪ] in A is longer because of the voiced coda consonant). Similarly, it seems that openness may have a role in the ordering of toe and fronts, running and jumping, and quality and quantity, but it is violated in ice and snow. In a controlled sample, where semantic, metrical, and frequency factors are excluded, we might find stronger evidence for phonological constraints.

4.5 Alphabetical order. When considered token-by-token, alphabetical order was significantly aligned with binomial order. Upon investigation, however, we found that this happens to have been due to the fact that several of the most common binomials, including back and forth (N=49), black and white (N=19), and here and there (N=16), were always in alphabetical order; all these are frozen binomials whose ordering is governed by a semantic constraint. When binomial types were weighted equally, there was no significant correlation between alphabetical order and binomial ordering.

4.6 Relationships among constraints. Of course, we cannot directly conclude that because binomials are significantly in alignment with a particular constraint, that constraint is directly implicated in determining binomial ordering, because activity is often highly correlated across constraints. We have already investigated many types of constraint correlations above, but it is also informative to look at a more complete set of linkages between correlated constraints. While space constraints prohibit display of the full correlation matrix for all constraints, we present Figure 1, a graph of plausible direct correlations between constraints in our dataset.

INSERT Figure 1 ABOUT HERE

Each link between constraint pairs indicates that the pair is significantly correlated in our dataset, and there is no other intervening constraint that could plausibly explain the correlation between the pair. For example, frequency and perceptual markedness are directly linked, because expressions that refer to perceptually salient entities are likely to have high corpus frequency. But frequency and placement of consonant-initial items in the B slot (CInit) are not directly linked: they are not significantly correlated in this corpus, and even if they were, the preference for longer items in the B slot would be a plausible mediator between the two constraints, since longer words are more likely than shorter words to be consonant-initial. Our linkage diagram shows that frequency and metrical constraints, together with some phonological and semantic constraints, constitute a connected cluster that is clearly implicated in binomial formation patterns. Any future study of binomials must take considerable care to disentangle these factors, given how closely they are related.

5. Interaction of Constraints. In the previous section we quantitatively investigated nineteen constraints in our corpus of binomials, and we found that about half of them were significantly correlated with binomial order. Furthermore, we found that these potentially explanatory constraints are often significantly correlated with each other. We therefore need to investigate constraint interaction and constraint rankings. This is not the first such investigation: Cooper and Ross suggest a possible ordering of constraints and call for further research to test it, and McDonald et al. (1993) give psycholinguistic evidence for the resolution of conflicts between semantic and metrical constraints. Our study differs from most previous work, however, in considering a wider array of constraint types and in applying quantitative methods to investigate constraint interaction.

We investigate constraint interactions in three frameworks: standard Optimality Theory (OT; Prince and Smolensky 1993), Stochastic Optimality Theory (StOT; Boersma 1998, Boersma and Hayes 2001), and logistic regression.[13] All three of these frameworks have the property of expressing linguistic outputs as the result of interacting, violable constraints. The latter two also have the properties crucial for modeling variation:

(16a) Modeling capability: ability to assign arbitrary probabilities (between zero and 1) to linguistic outputs
(b) Learnability: existence of algorithm for training model on variable linguistic input

Standard OT has been compared with logistic regression by Guy (1997), who notes that OT's lack of quantitative constraint ranking and learnability is both theoretically and empirically problematic. Previous comparisons of StOT and logistic regression include Goldwater and Johnson (2003), who found that the two frameworks modeled Finnish genitive plural data from Boersma and Hayes (2001) comparably; Ernestus and Baayen (2003), who used both frameworks to model Dutch neutralized segments;[14] and Jäger and Rosenbach (2004), who found that logistic regression worked much better than StOT for English genitive construction variation. Jäger and Rosenbach point out that StOT is a rather more restrictive probabilistic framework in terms of the kind of constraint conflict patterns it allows. In particular, 'ganging up' – the defeat of a single highly-ranked constraint by multiple constraints of lower rank – is prohibited in OT, and exists in only a very weak form in StOT. In the logistic regression framework, in contrast, the effects of multiple constraints are additive, which permits quite complex forms of ganging up.[15]

Regardless of the framework used to formalize and rank constraints, we can draw on an important theoretical idea from OT: the division of constraints governing surface linguistic forms into faithfulness constraints, determining the harmony of input forms with output forms, and markedness constraints, determining the intrinsic harmony of the output forms themselves. In the next section, we argue that all the relevant constraints on binomial formation are markedness constraints. Subsequently, we compare the formalization of these violable markedness constraints within OT, StOT, and logistic regression in two respects: in terms of their ability to represent constraint priorities and in terms of their ability to accurately model the actual distribution of binomials in our corpus.

5.1 Optimality Theory. Müller (1997) is the first to use an OT framework in analyzing binomials. However, his work is not quantitative, details of his corpus and methodology are omitted, and no exceptions to his constraints are given. Although he posits that the constraints are productive not only in frozen binomials but also in the general grammar, he does not check whether they are productive in the formation of non-frozen binomials. Our study addresses these issues.

One of the foundations of Optimality Theory is the notion of universal violable constraints that derive from general linguistic or psychological principles and are ranked differently in different languages. Many of the constraints discussed above can be generalized. The scalar constraint can be applied to progressions of intensity, such as in a Horn Scale ('You may; in fact, you must') (Levinson 2000). The power constraint can be applied to other areas of word order, such as the placement of the agent and the patient in passive sentences (Kuno and Kaburaki 1977). The perceptual markedness constraint applies to intrasentential information structure, where – focusing constructions such as the English pseudocleft aside – given information (more perceptually salient) often precedes new information (less salient). A lapse constraint has been used in other OT analyses (Green and Kenstowicz 1995), and a length constraint is active in the placement of lengthy phrases at the end of sentences (McDonald et al. 1993). The frequency constraint applies to several areas of the grammar, as Fenk-Oczlon (1989) explains. It is clear that the factors active in the ordering of binomials also exist in the grammar as a whole and can be considered universal and violable constraints.

We first discuss the competition between markedness and faithfulness. In the case of binomials, the following faithfulness constraints are implicitly at work:

(17) Ident-IO (lexical): Input should use the same single-word form-meaning pairs as output
Ident-IO (stress): Input should have the same stress as output
Dep-IO: Output depends on input (do not add segments)

As stated above, we assume that the input for a given binomial is A and B in an unspecified order, and the output is an ordering of A and B. In the OT framework, the generator (GEN) then generates all possible candidates for the binomial, based on the faithfulness and markedness constraints. As Kager explains, 'an output is "optimal" when it incurs the least serious violations of a set of constraints, taking into account their hierarchical ranking' (1999:13). This model has a clear cognitive correlate: a speaker having two words in mind (e.g. swiftly and easily) and then combining them in either order.

We begin by arguing that, as far as our corpus attests, all faithfulness constraints outrank all markedness constraints. In OT, this is possible only when the input leaves some feature unspecified; in the case of binomials, this feature is binomial output order.

Ident-IO (lexical). This analysis assumes an unordered input pair {A, B}, rather than an input of the form than ‘A and x’, where x can be any word. By definition, then, the output words always match the input. Although it is possible to imagine a speaker actually using a word in a binomial with a meaning different than originally intended, this would seem more likely in poetry than in prose. We saw no evidence for lexical faithfulness violations in our corpus.

Ident-IO (stress). Similarly, no token in our corpus involves a change in the stress of one item to conform to the metrical constraints of the binomial. We did find one token where each word seemed to change stress to fit better in its phonological phrase: outspoken and offbeat. This might be rendered ‘oùtspóken and òffbéat’, especially if it appears at the end of a phrase. However, in this case, it modifies a following noun that has initial stress:

(18) ‘The action centers about a group of óutspòken and óffbèat students . . . ’

In order to avoid clash between beat and students, the primary stress likely shifts to the prefix in offbeat. To maintain parallel structure, the stress of ‘outspoken’ likely shifts to the prefix as well. This is not actually an instance of stress change due to binomial formation, as each word could have either stress pattern on its own.

Müller suggests including in OT candidate lists an output in which stress is placed on the link, and. We have found binomials in our corpus where stress may conceivably be placed on and – in particular for binomials with the stress pattern . . . Sw and wSw . . . Some examples are Thailand and Malaysia, linguistics and psychiatry, and changing and improving. Following this same pattern, we might expect to see shifty and evasive and wisely and decisively. However, these do not occur, and their reverses do. Evasive and shifty is consistent with our constraint against lapse, and decisively and wisely is probably given a secondary stress on the first ly, suggesting that a secondary stress may be preferred on a content word over and. These few examples suggest that faithfulness of stress is a strong constraint for our data, although further research that controls for semantic and other metrical factors is needed for a deeper understanding of secondary stress in binomials.

Dep-IO. No binomial in our corpus shows evidence of an item lengthening or shortening to satisfy a markedness constraint. We are, however, aware of one English binomial (which did not appear in our corpus) that does this: mac and cheese, shortened from macaroni and cheese, apparently to satisfy *A>B or *ww. Shortening is common in German, as in the binomials Katz und Maus and Freud und Leid, whose inputs would be /Katze, Maus/ and /Freude, Leid/.

We now turn our attention to markedness constraints. Since multiple outputs are attested for some inputs, no single total ordering of constraints can account for every attested ordering of English binomials. Our approach for the rest of the section is to examine categorical and variable approaches to constraint ranking and output resolution that produce the best overall fit to our corpus. Based on the findings in Section 4, we compare hand-ranked categorical constraints with automatically learned variable constraints from Stochastic Optimality Theory (Boersma 1998). The latter has the advantage of being able to model continuous-valued corpus frequencies, and includes a learning procedure, the Gradual Learning Algorithm (Boersma and Hayes 2001).

5.2 Hand ranking. The investigation above suggested a natural ordering of constraints by constraint type, with semantic and pragmatic constraints outranking metrical constraints, metrical constraints outranking frequency constraints, and frequency constraints outranking phonological and orthographic constraints. This ranking is consistent with direct comparisons of conflicts between constraint types. Of 77 tokens involving conflict between semantic and metrical constraints, metrical constraints win only 12, consisting of the types in (19) below:

(19) Conflicts between semantic and metrical constraints, won by metrical:
peanuts and emeralds, friends and family, everything and everybody, always and everywhere, mental and physical, interest and principal, harass and punish

Similarly, metrical constraints beat frequency constraints in 59% of 175 tokens (67% of 99 tokens if semantically aligned binomials are excluded). Similar analyses of individual constraints yield the ordering in (20):

(20) Hand-ranking of constraints:
Pragmatic > Iconic > RelForm, Power, BStr, > *A>B > *ww > Freq > CInit > other

This ranking correctly derived 76.3% of the surface binomial types and 71.4% of tokens in our corpus. We found that orderings violating the pattern semantic > metrical > frequency > phonological derived a smaller proportion of the corpus, although the relative ranking of frequency and metrical constraints, and of *A>B and *ww, made only a small difference. For the ordering in (20) we found that no constraint below CInit was decisive; that is, no input binomial type had an identical constraint profile for CInit and all higher-ranked constraints.

This finding is in line with McDonald et al. (1993), Müller (1997), and Levelt and Sedee (2004), who argued that semantic constraints outrank metrical constraints. This finding contrasts with the results of Fenk-Oczlon (1989) for frozen binomials, where frequency alone accounted for 84% of the corpus. In our corpus, frequency by itself is aligned only with 55% of the binomial types; a constraint combination can derive over 20% more.

5.3 Variation and automatically learning OT constraints. A number of other linguists have applied OT to variable data, using various modifications to the theory. This has been done in at least four different ways (see Hinskens et al. 1997). Van Oostendorp (1997) posits competing grammars to account for various styles. He says that each speaker has command of multiple grammars, which have different rankings of the same constraints, and can style-shift among them at will. Zubritskaya (1997) and Boersma (1998, Boersma and Hayes 2001) both suggest that intra-speaker variation can be modeled as one system in which each constraint has a numerical weight attached to it. In categorical phenomena, the weights are far enough apart that overlap is minimal, but when two constraints have similar weights, variation can occur.

The other two theories of variable OT both include partial ordering of constraints.

Reynolds (1994) suggests constraints that can 'float' around within a ranking. And Anttila (1997) posits systems of constraints where some are left unranked. Of all the possible grammars, the percentage of those in which a candidate wins predicts its probability of occurrence. For example, in Anttila's data from the genitive plural in Finnish, one environment has three constraints that are unranked with respect to each other. The candidate that violates only one of these three constraints is predicted to occur 2/3 of the time. This particular prediction is vindicated, as this form occurs 62.8% in the large corpus (1997:60).

The best account of binomial orderings with respect to our corpus must come from a modeling framework that can account for variation. We choose Stochastic Optimality Theory for such a model for two reasons: first, it can automatically learn constraints (which Reynolds' and Anttila's models cannot); and second, it does not require additional constraints to model arbitrary probability values between 0 and 1.[16]

We relied on the Gradual Learning Algorithm (GLA; Boersma and Hayes 2001) to find the optimal ranking for our constraint set. In the GLA, the learning process occurs as follows. First, all constraints are given an initial ranking. Next, the grammar is repeatedly presented with stimuli in the form of input-output pairs, randomly sampled from a corpus. For each pair, an output for the input is randomly chosen according to the current state of EVAL. If the correct output is chosen, nothing happens. If the incorrect output is chosen, the constraint rankings change according to the difference in the constraint violation profiles between guessed and true outputs. Each constraint violated in the guessed pair and unviolated in the true pair is demoted; each constraint violated in the true pair and unviolated in the guessed pair is promoted. The size of demotions and promotions is determined by a plasticity factor dependent only on the number of learning steps that have proceeded. Learning is complete when stimuli cease to be presented.[17]

In experiments with our full constraint array, we found that constraint rankings failed to converge; however, the probability of constraint orderings for given constraints did stabilize.[18] We used two different training regimens: one where each attested ordered binomial type was represented with equal weight in the training sample (type-based training) and one where each type was represented proportionately to its frequency of occurrence (token-based training). The learned constraint rankings differed somewhat for the two regimens, as shown in 21 below ( '>' denotes that the difference between constraint rankings was less than the noise constant, and '>>' denotes that the difference was much larger than the noise constant).

(21a) Type-based learned ranking:
Icon > *A>B > Freq >> Percept >> *ww > *BStr >> RelForm >> Power
>> Alpha >> Pragmatic >> VPhonetic >> SFin > CFin >> SInit > CInit
>> Open >> Height > VPhonemic > Weight > Back
(b) Token-based learned ranking
Icon > *A>B > > Alpha > BStr > CFin > Freq >> VPhonetic > *ww >>
Percept >> RelForm >> Weight > VPhonemic >> SFin >> Power >>
Pragmatic >> Back >> Open >> CInit >> SInit >> Height

Surprisingly, the GLA does not achieve the hand-chosen constraint rankings for either type- or token-based learning. In type-based learning, semantic, metrical, and frequency constraints all outrank phonological constraints, but are themselves mixed together. *A>B and *ww outrank *BStr, and Frequency outranks Relative Markedness and Power. This constraint ranking correctly predicts only 72.6% of the binomial types, considerably less than the hand rankings (76.3%). In token-based learning, the results are even more skewed: several phonological constraints outrank semantic, metrical, and frequency constraints. Token for token, however, the automatically learned ranking in (21b) actually matches our dataset better than the hand ranking (see Table 5**Table**).

5.4 Logistic Regression. Logistic regression (Hosmer and Lemeshow 2000) is a widely used statistical methodology for categorical data analysis. As it applies here, the probability *p* of a particular binomial ordering A and B for an input pair {A,B} is assumed to be of the form

$$\log\left(\frac{p}{1-p}\right)=\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n$$

Equation 1. Logistic Regression.

where $\beta_i$ is a real-valued number corresponding to the weight of the $i^{th}$ constraint, and $x_i$ is 1 if the $i^{th}$ predictor is active and aligned with A and B, -1 if active and aligned against *o*, and 0 if inactive. The probability of the alternative binomial ordering B and A will be 1-*p*, since each $x_i$ for B and A will always be the negative of the corresponding $x_i$ for A and B.

Logistic regression shares with StOT the advantage that arbitrary probabilities between 0 and 1 can be modeled without additional numbers of constraints. Unlike StOT, Logistic regression is expressive enough to model the cumulative effects of weaker constraints against stronger constraints. Given three constraints weighted such that $|\beta_1|>|\beta_2|>|\beta_3|$, constraints 2 and 3 can gang up against constraint 1 as long as $|\beta_2|+|\beta_3|>|\beta_1|$.[19] Logistic regression also enjoys one more learnability property unshared by StOT: under typical learning regimes, the optimum is guaranteed to be unique and found.

Table 4 shows the coefficient, or weight, of each constraint in logistic regression models trained on the full constraint set. Note that in logistic regression, a constraint can have a negative weight, meaning that it prefers to be aligned against the binomial. The larger the magnitude of the constraint, the stronger its effect on binomial ordering. Furthermore, constraint weights in the model are on an interval scale, so their magnitudes can be compared numerically. In the model of Table 4, for example, every semantic constraint is more powerful than all metrical constraints combined.

Insert Table 4 about here

Broadly speaking, relative constraint strengths in the model trained on binomial types are consistent with our hand-ranked OT model. Semantic constraints are ranked above metrical constraints, *BStr is the highest-ranked metrical constraint, frequency is ranked among the metrical constraints, and phonological constraints (with the exception of phonetic vowel length) are ranked the lowest.

When we train the regression to optimize on binomial tokens rather than types, we see two major qualitative changes. First, the rankings of non-semantic constraints increase relative to those of semantic constraints. This seems to have to do with the fact that several of the most common binomials have a strongly preferred ordering violating some semantic constraint, including back and forth (N=49), black and white (N=19), off and on (N=7), and interest and principal (N=7). Training on tokens rather on types forces the model to magnify the effects of non-semantic constraints to explain these common binomials. Second, the Power constraint loses its strong positive ranking and becomes slightly negatively ranked. This probably results from the fact that there is one frequent binomial type (N=19), black and white, which violates the Power constraint and is not aligned with any other strongly-weighted constraint. While there is another frequent binomial type (N=15), men and women, which is aligned with the Power constraint, it is also aligned with all three metrical constraints. The fact that constraint effects in logistic regression are additive means that these metrical constraints can 'explain away' the binomial men and women, leaving the model free to reduce the strength of the Power constraint so as not to make black and white too improbable.

Because our full logistic regression model uses a large number of constraints relative to the size of the dataset, it is not possible to draw detailed conclusions from the specific values of resulting constraint weights. It is, however, possible to use these weights to identify broad trends, and to compare the ability of logistic regression to learn a close fit to our data, compared with OT and StOT models.

5.5 Comparison of constraint coverage. We are now in a position to ask the following question: given the constraint profiles for our binomials corpus, which model better captures the ordering realization patterns of our corpus: OT, StOT, or logistic regression? This question follows in the footsteps of Goldwater and Johnson (2003) and Jäger and Rosenbach (2004), who compared StOT models with maximum-entropy models for identical datasets and constraint profiles.

Since we have been investigating the problem of ordering realization for unordered binomial inputs, we evaluate the fit of a model against our corpus by how well it predicts the output ordering for each input binomial. In particular, we focus on two ways of quantifying this fit: hard evaluation, in which a model is assumed to choose the highest-likelihood output for a particular input and constraint profile, and soft evaluation, in which we treat the model output as a probability distribution over output orderings and measure the difference between the predicted and empirical output distributions for each input in the corpus. In hard evaluation, we assume that the grammar uniformly guesses the highest-probability output for each input, and report the percentage of true output binomials (types or tokens, respectively) correctly guessed. In soft evaluation, we report the relative entropy, or Kullback-Leibler divergence, of the true distribution for each output given the input from the guessed distribution (Cover and Thomas 1991).[20] (Lower numbers for relative entropy indicate a better match to the target distribution, with zero indicating an exact match. If some outcome with a non-zero probability in the target distribution is given a zero probability in the guessed distribution, the relative entropy is always infinite.) Whereas hard evaluation indicates a model's ability to accurately guess the output ordering for a single instance of an input type, soft evaluation indicates its ability to match the frequency of output orderings.[21] In both hard and soft evaluation we report results weighted both by input type and input token.

Insert Table 5 about here

Table 5 shows how well the hand-ranked OT, automatically learned StOT, and logistic regression models capture the ordering patterns in our corpus of binomials. By all evaluation measures, the resulting logistic regression model matches binomial ordering patterns more closely than the OT and StOT models do, both in type-based and token-based training. While the hard-evaluation difference between OT and logistic regression models for types is quite small, logistic regression shines most in soft evaluation, which directly tests the model's ability to closely match output frequencies seen in the corpus. This is consistent with arguments proposed by Guy (1997) for the superiority of logistic regression over standard OT, and with the findings of Jäger and Rosenbach (2004), who showed that word order realization for English genitives had additive effects across animacy, topicalty, and possessive relation that resulted in a closer fit from a logistic model than a stochastic OT model.[22]

Why would the logistic model result in a closer fit than OT and StOT? As noted before, logistic regression allows a kind of *ganging up* that is prohibited in StOT: two weaker constraints can combine to overcome a single stronger constraint. Is this the case in our binomials corpus?

5.6 Ganging up. Since none of our models achieves perfect prediction, it is impossible to say *a priori* whether ganging up is required to accurately model our corpus – it is always possible that a different constraint inventory can achieve perfect prediction without any ganging up. However, given that for the existing constraint inventory, logistic regression achieved a better fit to our data than the OT and stochastic OT models, we can ask a related question: within our full logistic regression models, are there binomials for which the preferred ordering is accurately predicted, but is out of alignment with the strongest active constraint? These binomials would be good candidates for ganging up.[23]

When we applied these criteria to our type-trained logistic regression model, we found twelve matching binomial input types.[24] Nine of these involved Height either as the strongest constraint or ganging up against the strongest constraint; we discarded these, as it is our belief that the large negative weight assigned to Height in the model is overfitting.[25] We also discarded *hope and pray*, involving a conflict of Relative Markedness and Power, which are nearly identically ranked in the model. The two remaining matches are *automobiles and factories*, where the strongest constraint, *A>B, is ganged up on by Frequency, CInit, Phonemic Vowel Length, and Backness; and *clerks and postmasters*, where the Power constraint is ganged up on by a combination of all three metrical constraints.

We also note that the binomial *evasive and shifty*, though it involved the Height constraint ganging up against *A>B, also involves *ww, Frequency, and CInit in opposition, suggesting that it is a possible case of ganging up. Finally we also trained a smaller logistic regression model, involving only the relevant constraints from Section 3.6, and found that *clerks and postmasters* remained an instance of ganging up in this simpler model.

In summary, although in the context of modeling variation it is difficult to determine whether there is evidence in a corpus for ganging up within a given set of constraints, we found two related pieces of evidence suggesting that ganging up happens occasionally among naturally occurring English binomials. First, direct comparison of

OT and StOT models, which do not allow ganging up, with logistic regression models, which do, show that logistic regression is able to achieve a better overall fit to the corpus. Second, within the logistic regression model, we found a small number of binomials for which the preferred, accurately predicted order involved a number of weaker constraints overwhelming the strongest active constraint. Judging from our corpus, however, ganging up does not appear to be a primary feature of binomial ordering.

6. Other Factors. There are several other factors that could contribute to the order of naturally occurring binomials. For example, in a non-frozen binomial, the speaker or writer may have used Item A and then thought of adding Item B with a lexical link. Our model does not account for this process, as it assumes an input of {A,B}. Second, there may be pragmatic factors not discernible from the section of the corpus in which it occurred. Perhaps, in a previous part of the conversation or writing, Item A was discussed at length, and Item B is now new information. This would be covered by the current pragmatic constraint, but it cannot be detected in the corpus.

Frozen binomials, as well, may be determined partly by a number of other factors. For example, sugar and spice and various and sundry are fixed binomials where no semantic constraint is satisfied but metrical constraints are violated nonetheless. What factors could be contributing to their ordering? One possibility is the context in which the binomial became frozen. Many binomials are popularized by a well known poem or song. These may violate the optimal ranking because they fit better in the rhyme scheme. Examples are sugar and spice, which violates the metrical constraints but rhymes with ‘everything nice’, and jam and bread, which violates the condiment rule of the Power Constraint but rhymes with ‘(a needle) pulling thread’ in Oscar Hammerstein’s ‘Do-Re-Mi’ (see also Billy Joel’s ‘Piano Man’, where Cooper and Ross’s alcohol rule is violated by ‘tonic and gin’, which rhymes with ‘regular crowd shuffles in’). Or, songs might use a binomial that violates the optimal ranking because it fits better with the imagery of the song. An example is night and day, which violates Perception-Based Markedness but – in Cole Porter’s song – conjures images of an unrequited lover staying up all night with pangs of sorrow and continuing his weeping into the next day. In these binomials there are a number of contextual constraints contributing to the ordering, including constraints imposed by rhyme and imagery.

These constraints may not be acting on the ordering every time the phrases are uttered out of context (e.g. ‘This girl’s so sweet – she’s like sugar and spice!’) or even in partial context (e.g. a quote of the song, ‘Tea, a drink with jam and bread’). However, we can consider the order of these binomials to be partially lexicalized, with the order of lexicalization determined by the context in which the binomial was popularized.

Another such idiosyncratic constraint that may affect the ordering of frozen binomials is loan translation. Milk and honey and flesh and blood are likely translations of the Biblical Hebrew halav udevash and basar vadam; and bread and circuses and divide and rule are likely from Latin panem et circenses and diuide et impera (Malkiel 1959:153-4).

One more historical issue that might affect frozen binomial ordering is changes in sound or meaning. It is possible that – due to phonological changes – a binomial’s words had a different number of syllables or a different stress pattern when the order was crystallized. It is also possible that the meaning of one of the words shifted or that the original sense of the entire binomial is lost. An example of the latter may be back and forth, which violates the markedness constraint but seems to have originally followed the scalar constraint, as it may have denoted a nautical sequence, related to ‘to back and fill’ (Malkiel 1959:148).

As we can see, there are several other factors that may affect the ordering of binomials, including the thought order of the speaker or writer, unidentified contextual effects, rhyme, imagery, loan translation, and historical change. The latter four seem especially applicable to frozen binomials; the former two to unfrozen binomials. Within the context of our corpus and models, we have simply not been able to identify these factors on a systematic basis, and they presumably constitute the remaining unexplained variation in the models.

6.1 Further issues in modeling linguistic variation: finer constraint gradation. Although we coded constraint activity with three discrete values – a constraint could be inactive, aligned with a binomial, or aligned against it – a number of constraints in our inventory could usefully be coded with finer gradation. For two of the three metrical constraints, *A>B and *ww, constraint activity could be measured as the number of syllables (number of consecutive weak syllables for *ww) by which A and B differ, rather than simply the direction of the difference. For the frequency constraint, constraint activity could be measured as the ratio or difference of item frequency. For CInit and CFin, the difference between the number of consonants in the A and B items could be used. And the difference between actual millisecond values for main vowels taken from Crystal and House (1988) could be used to measure vowel length.[26]

Such gradience could be incorporated into a formal model in a variety of ways. In OT and StOT, some types of gradients could be captured as counting violations, although it seems to us that the lack of tied constraints in our model would mean that counting violations would have little effect on results. Alternatively, special ‘multiple-violation’ constraints could be added to an (St)OT model. The situation is even better in logistic regression, which can handle real-valued constraint magnitudes, and requires that a large magnitude always have a stronger effect than a smaller magnitude for a given constraint, which seems natural in this case. We expect that coding and modeling with finer constraint gradations would only increase the insight we could gain into the relative effect of various factors on binomial ordering.

6.2 Variation among frozen binomials. Although the focus of this paper has not been limited to frozen binomials, our data also have some bearing on the question of whether is it possible for the reverse of a frozen binomial to be grammatical. We have found that the answer is yes. We found a few instances where purportedly frozen binomials appear in reverse: principal and interest also appears as interest and principal, and near and dear is also dear and near. There are also a number of binomials where both orders are frozen, such as left and right and right and left; off and on and on and off; and night and day and day and night.

In addition, advertising campaigns have popularized the reverses of certain binomials, such as macaroni and cheese to cheese and macaroni (Kraft) and family and friends to friends and family (Sprint). Similarly, one could imagine a British potato lobby starting a campaign for chips and fish, and the reverse binomial women and men is the title of a song by They Might Be Giants. For many frozen binomials that usually occur in a specific order, one could conceive of a context where the reverse would be appropriate. For example, one could say 'I just pray and hope a lot that she'll be OK,' (where the praying is more central) or 'Out of all the spices, they sold the most pepper and salt' (where pepper outsold salt). Although these are possible grammatical strings, one would still consider hope and pray and salt and pepper to be frozen binomials and to sound better in their canonical order in most circumstances.

However, there are some frozen binomials that almost always appear in one order and would not be nearly as intelligible in the reverse order, phrases that Malkiel (1959) calls Irreversible Binomials. These include binomials where the sum of the parts has come to mean something different from the two items, such as odds and ends, by and large, and high and dry. They also include binomials where one or both of the words are no longer common in the language, such as kith and kin and kit and caboodle. Would these binomials be ungrammatical in the reverse order? Perhaps they would be difficult to understand and very rare. Of course, it is certainly possible to utter them in reverse, and we might expect to encounter them in a metalinguistic context such as a joke.

In terms of distinguishing frozen from unfrozen binomials within a model of binomial order variation, it is important to keep in mind that a model of binomial types is subtly different from a model of binomial tokens. We can interpret a model of binomial types essentially as a model of lexicalization tendencies within binomial ordering. A model of binomial tokens, on the other hand, predicts the actual ordering that will be used for a given binomial in a given instance. If there is clear evidence that a common frozen binomial F has an established, idiosyncratic ordering preference in conflict with the general principles of binomial ordering, such as is the case for back and forth, which violates perceptual markedness but is uncontestably irreversible, it is justified to introduce a lexeme-specific ordering constraint into a model of binomial tokens that applies only to instances of F, and to assign a very high weight to that constraint. This may seem ad hoc, but if our goal is to accurately and parsimoniously explain the distribution of binomial tokens in a corpus, there is nothing wrong with explaining a large number of tokens of F with a type-specific constraint, especially if there is a clear historical explanation for F's anomalous ordering. To introduce the same constraint in a model of binomial types, on the other hand, would be inappropriate, since it would only explain a single data point. The appropriate alternative in this situation would be to identify all binomial types in the corpus with a common type of explanation (say, rhyme or historical sound change) and lump these types under a single constraint.

7. Conclusions. Binomial formation has been the subject of a variety of studies in the past half century, including exploratory essays, cross-linguistic comparisons, and perceptual experiments. But little work on naturally-occurring data has compared multiple types of constraints. The present study fills this gap and adds to our collective wisdom in a few ways. It finds that semantic, metrical, and frequency constraints that others have posited in studies of frozen binomials do apply in non-frozen binomials as well. Among metrical constraints, this paper found that Bolinger's (1962) constraint against ultimate stress of B was the most reliable indicator of binomial order. In line with this finding that the position of stress was an important determinant of binomial ordering, we also suggested a number of phonological factors that might be expected to have an

effect on ordering, based on the greater stress of B. We found evidence for only one of these constraints: the tendency for larger initial consonant clusters in B. We expect that the proposed phonological constraints would show up more in a corpus that includes larger numbers of binomials that are minimal pairs. The main trend we found in our data was the prominence of semantic over metrical constraints, and metrical over frequency constraints. We expect that a similar relationship might be found among these different levels of grammar in phenomena other than binomial formation where semantic, phonological, and frequency factors are also relevant.

Another major conclusion from this study is that a descriptively adequate model of binomial formation and production cannot be complete without the option for reversal. Binomial ordering is a non-categorical phenomenon involving constraint conflict – a finding which has led us to investigate three violable-constraints frameworks: Optimality Theory, stochastic Optimality Theory and logistic regression. All of these frameworks are able to handle the interaction of conflicting constraints in binomial ordering, with OT being the most restrictive, and StOT more restrictive than logistic regression. In all frameworks we found models that accurately predicted over 70% of our corpus data; for StOT and logistic regression, we were able to automatically learn such models. We found that for our full constraint set, logistic regression was able to achieve a better fit to our corpus than both hand-constructed OT and automatically learned StOT models. This is particularly impressive considering that under type-based training, the StOT model was unable to learn as close a fit to the data as we constructed by hand. We suggested that 'ganging up' of weaker constraints on stronger constraints might be the reason for the better fit of logistic regression, although 'ganging up' did not seem to be especially prominent in our data. There was also a considerable amount of residual, unexplained variation in our models, and we discussed a number of extrinsic factors that might determine otherwise inexplicable binomial orderings.

Now we are truly and really able to answer the age-old question: which comes first, the chicken or the egg? The metrics would predict egg and chicken, but perceptual markedness would predict the animate-initial chicken and egg. Since the semantics outrank the metrics, the answer should be chicken and egg. However, this is only a probabilistic determination, and egg and chicken would not be ungrammatical. We now have the answer to the age-old question – probably.

## REFERENCES

Agresti, Alan. 2002. *Categorical Data Analysis*. New York: John Wiley and Sons.

Andersen, Henning. 1972. Diphthongization. *Language* 48.11-50.

Anttila, Arto. 1997. Deriving variation from grammar. *Variation, change, and phonological theory*, ed. by Frans Hinskens, Roeland van Hout, and W. Leo Wetzels. Amsterdam: John Benjamins.

Battistella, Edwin. 1990. *Markedness: The evaluative superstructure of language*. Albany: SUNY Press.

Boersma, Paul. 1998. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics. University of Amsterdam dissertation. Chapter 15. (http://www.fon.hum.uva.nl/paul/)

Boersma, Paul. 2001. Review of variation in Finnish phonology and morphology, by Arto Anttila. *GLOT International* 5/1.31-40.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32.45-86.

Bolinger, Dwight. 1962. Binomials and pitch accent. *Lingua* 11.34-44.

Cedergren, Henrietta J. 1973. On the nature of variable constraints. *New ways of analyzing variation in English*, ed. by C.J. Bailey and R. Shuy, 13-22. Washington: Georgetown University Press.

Cedergren, Henrietta J. and David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50.333-355.

Conover W.J. 1980. *Practical non-parametric statistics*, 2nd ed. New York: John Wiley and Sons.

Cooper, William and John Ross. 1975. World order. *Papers from the parasession on functionalism*, ed. by R. Grossman, L.J. San, and T. Vance, 63-111. Chicago: Chicago Linguistic Society.

Cover, Thomas M. and Joy A. Thomas. 1991. *Elements of information theory*. New York: Wiley.

Crystal, Thomas and Arthur House. 1988. The duration of American-English vowels: An overview. *Journal of Phonetics* 16.263-284.

Ernestus, Mirjam, and R. Harald Baayen. 2003. Predicting the Unpredictable: Interpreting Neutralized Segments in Dutch. *Language* 79.5-38.

Fenk-Oczlon, Gertraud. 1989. Word frequency and word order in freezes. *Linguistics* 27.517-556.

Goldwater, Sharon and M. Johnson 2003. Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, ed. by Jennifer Spenader, Anders Eriksson, and Östen Dahl. 111-120.

Green, Thomas and Michael Kenstowicz. 1995. The lapse constraint. ROA-101-0000.

Gustafsson, Marita. 1974. The phonetic length of the members in present-day English binomials. *Neuphilologische Mitteilungen* 75.663-677.

Gustafsson, Marita. 1976. The frequency and ‘frozenness’ of some English binomials. *Neuphilologische Mitteilungen* 77.623-637.

Guy, Gregory R. 1997. Violable is variable: Optimality theory and linguistic variation. *Language Variation and Change* 9.333-347.

Hazen, Kirk. 2002. Identity and language variation in a rural community. *Language* 78.240-257.

Hinskens, Frans, Roeland van Hout, and W. Leo Wetzels. 1997. Balancing data and theory in the study of phonological variation and change. *Variation, change, and phonological theory,* ed. by Frans Hinskens, Roeland van Hout, and W. Leo Wetzels. Amsterdam: John Benjamins.

Hosmer, David W. and Stanley Lemeshow. 2000. *Applied logistic regression*. New York: Wiley & Sons.

Jäger, Gerhard, and Anette Rosenbach. 2005. The winner takes it all – almost. Cumulativity in grammatical variation. To appear in *Linguistics*.

Jakobson, Roman. 1984 [1939]. Zero sign. *Russian and Slavic grammar studies 1931-1981*, ed. by Linda R. Waugh and Morris Halle, Chapter 11. Berlin: Mouton.

Jescheniak, Jörg, and and Willem Levelt. 1994. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition* 20.824-843.

Kager, René. 1989. *A metrical theory of stress and destressing in English and Dutch*. Dordrecht: Foris.

Kager, René. 1999. *Optimality theory*. Cambridge: Cambridge University Press.

Kuno, Susumu and Kaburaki, Etsuko. 1977. Empathy and syntax. *Linguistic Inquiry* 8.627-672.

Ladd, Robert. 1996. *Intonational phonology*. Cambridge: Cambridge University Press.

Ladefoged, Peter. 1975. *A course in phonetics*. New York: Harcourt Brace Jovanovich.

van Langendonck, Willy. 1986. Markedness, prototypes and language acquisition. *Cahiers de l'institute de linguistique de Louvain* 12.39-76.

Levelt, Claartje and Willemijn Sedee. 2004. De normen en waarden van 'normen en waarden'. Paper presented at TIN-dag (The Linguistic Society of the Netherlands annual meeting), Utrecht, February 2004.

Levinson, Stephen. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge: MIT Press.

Malkiel, Yakov. 1959. Studies in irreversible binomials. *Lingua* 8.113-160.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19.313-330.

Mayerthaler, Willi. 1988 [1981]. *Morphological naturalness*. Translated by Janice Seidler. Ann Arbor: Karoma.

McCarthy, John J. and Alan Prince. 1994. The emergence of the unmarked: Optimality in prosodic morphology. ROA 13-0594. 1-30. Also in *Proceedings of the North East Linguistic Society* 24. M. Gonzàlez, ed., 333-79.

McDonald, Janet L., Kathryn Bock, and Michael H. Kelly. 1993. Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology* 25.188-230.

Müller, Gereon. 1997. Beschrankungen fur Binomialbildungen im Deutschen. *Zeitschrift fur Sprachwissenschaft* 16, 1/2.5-51.

Nespor, Marina and Irene Vogel. 1986. *Prosodic phonology.* Dordrecht: Foris.

Nespor, Marina and Irene Vogel. 1989. On clashes and lapses. *Phonology* 6.69-116.

Oakeshott-Taylor, John. 1984. Phonetic factors in word order. *Phonetica* 41.226-237.

Oldfield, R.C., and Arthur Winfield. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17.273-281.

van Oostendorp, Marc. 1997. Style levels in conflict resolution. *Variation, change, and phonological theory*, ed. by Frans Hinskens, Roeland van Hout, and W. Leo Wetzels, 207-30. Amsterdam: John Benjamins.

Paolillo, John. 2001. *Understanding linguistic variable rule analysis*. Stanford: CSLI.

Pinker, Steven and David Birdsong. 1979. Speakers' sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior* 18.497-508.

Pordany, Laszlo. 1986. A comparison of some English and Hungarian freezes. *Papers and Studies in Contrastive Linguistics* 21.119-127.

Prince, Alan and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. ms. Rutgers University and University of Colorado.

R Development Core Team. 2004. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Reynolds, William. 1994. *Variation and optimality theory*. Philadelphia: University of Pennsylvania dissertation.

Selkirk, Elizabeth. 1984. *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Veatch, Thomas Clark. 1991. *English vowels: Their surface phonology and phonetic implementation in vernacular dialects*. Philadelphia: University of Pennsylvania dissertation (single-spaced version).

Wasow, Thomas. 2002. *Postverbal behavior*. CSLI Press.

Wingfield, Arthur. 1968. Effects of frequency on identification. *American Journal of Psychology* 81.226-234.

Wright, Saundra and Jennifer Hay. 2002. Fred and Wilma: A phonological conspiracy. *Gender and linguistic practice*, ed. by Sarah Benor, Mary Rose, Devyani Sharma, Julie Sweetland, and Qing Zhang, 175-191. Stanford: CSLI Press.

Zipf, George K. 1949. *Human_behavior and the principle of least effort*. Reading, MA: Addison-Wesley.

Zubritskaya, Katya. 1997. Mechanism of sound change in Optimality Theory. *Language Variation and Change* 9.121-148.

CORPORA

Brown: http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM

Switchboard: http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html#summary

Wall Street Journal: http://www.ldc.upenn.edu/Catalog/LDC95T7.html

Appendix. Alphabetical list of binomial types.

N is the number of tokens of the input binomial type found.
% is the proportion of binomial tokens found in the order presented in the table.

+/-[constraint] means that the alphabetical ordering of the binomial is aligned with/against the constraint.

I = Iconicity
Pw = Power
Pt = Perceptual Markedness
R = Relative Formal Markedness
A = Absolute Formal Markedness

| Binomial | N | % | Sem |
|---|---|---|---|
| Americans and English | 1 | 0 | |
| By and large | 1 | 1 | |
| Connecticut and Massachusetts | 1 | 1 | |
| Czechoslovakia and Hungary | 1 | 1 | |
| England and Ireland | 1 | 1 | +Pw |
| Iowa and Nebraska | 2 | 1 | |
| Lotus and WordPerfect | 1 | 0 | |
| Malaysia and Thailand | 1 | 0 | |
| Slowly and thoughtfully | 1 | 1 | |
| T-Ball and soccer | 2 | 1 | |
| about and out | 3 | 0 | -I |
| abroad and here | 2 | 0 | -Pt |
| abused and neglected | 1 | 1 | |
| accept and hire | 1 | 1 | +R,+I |
| accepted and proposed | 1 | 0 | -I |
| accurately and promptly | 1 | 0 | |
| accurately and quickly | 1 | 0 | |
| acetate and cotton | 1 | 0 | |
| achieved and maintained | 1 | 1 | +I |
| action and conversation | 1 | 0 | |
| actively and continually | 1 | 1 | |
| acute and correct | 1 | 0 | |
| adamant and calm | 1 | 0 | |
| adding and using | 1 | 1 | +I |
| administrating and running | 1 | 1 | |
| administrative and scientific | 1 | 0 | -Pw |
| administrative and technical | 1 | 1 | |
| admired and knew | 1 | 0 | |
| again and now | 3 | 0 | -I |
| aggressive and persistent | 1 | 1 | |
| aggressively and swiftly | 1 | 0 | |
| alterations and sewing | 1 | 1 | -R |
| altogether and finally | 1 | 1 | |
| always and everywhere | 1 | 1 | -Pt |
| amply and cheerfully | 1 | 1 | |
| anger and anxiety | 1 | 1 | |
| anger and spite | 1 | 1 | |
| angst and science | 2 | 0 | -Pt |
| animals and humans | 1 | 1 | -Pt |
| animated and magnified | 1 | 0 | |
| answer and ask | 1 | 0 | -I |
| answers and questions | 1 | 0 | -I |
| anthropology and linguistics | 1 | 1 | |
| anxiously and eagerly | 1 | 1 | -Pt |
| appraisingly and coldly | 1 | 0 | |
| appreciate and understand | 1 | 0 | -R |
| appropriate and reasonable | 1 | 0 | |
| approved and commended | 1 | 1 | +R |
| approved and welcomed | 1 | 0 | +R |
| around and round | 1 | 0 | |
| attract and train | 1 | 1 | +I |
| attracting and keeping | 1 | 1 | +I |
| automobiles and factories | 1 | 1 | |
| back and forth | 49 | 1 | -Pt |
| back and there | 1 | 0 | -I |
| backward and forward | 1 | 0 | -Pt |
| backwards and forwards | 1 | 1 | -Pt |
| bad and good | 2 | 0 | -Pt |
| bad and ugly | 2 | 0 | |
| bananas and strawberries | 1 | 0 | |
| bar and pie | 1 | 0 | |
| been and gone | 1 | 1 | +I |

| | | | |
|---|---|---|---|
| better and interesting | 1 | 0 | |
| big and thick | 1 | 1 | |
| bitter and resentful | 1 | 1 | |
| black and innocent | 1 | 1 | +Pt |
| black and white | 19 | 1 | -Pw |
| bland and neutral | 1 | 1 | |
| boards and two-by-fours | 1 | 1 | +R |
| bobbed and gobbled | 1 | 1 | |
| bold and entertaining | 1 | 1 | |
| book and movie | 1 | 0 | |
| bookkeeping and taxes | 1 | 0 | |
| born and raised | 3 | 1 | +I |
| bottles and cans | 2 | 1 | |
| bought and sold | 10 | 1 | +I |
| brief and shallow | 1 | 0 | |
| broccoli and cauliflower | 1 | 1 | +Pw |
| brothers and sisters | 3 | 1 | +Pw |
| brown and thick | 1 | 0 | |
| build and operate | 4 | 1 | +I |
| built and made | 1 | 0 | |
| busily and profitably | 1 | 1 | +I |
| business and government | 2 | 0.5 | |
| buy and sell | 11 | 1 | +I |
| buying and holding | 1 | 1 | +I |
| calm and relaxed | 1 | 1 | |
| calmly and carefully | 1 | 1 | |
| capturing and taking | 1 | 1 | -R |
| carefully and prudently | 1 | 1 | |
| caring and compassionate | 1 | 1 | |
| caring and loving | 1 | 0 | |
| catch and try | 1 | 0 | |
| certain and quick | 1 | 0 | |
| champagne and dessert | 1 | 1 | |
| changing and improving | 1 | 1 | +R |
| chanted and chortled | 1 | 1 | +Pt |
| charming and pleasant | 1 | 0 | |
| chattered and coughed | 1 | 0 | |
| check and discipline | 1 | 1 | +I |
| chicken and egg | 1 | 1 | +Pt |
| chilling and muddling | 1 | 0 | |
| civil and criminal | 1 | 0 | |
| clean and dry | 1 | 1 | |
| clean and straight | 1 | 1 | |
| cleaner and faster | 1 | 0 | |
| clergymen and parishioners | 1 | 1 | +Pw |
| clerks and postmasters | 1 | 1 | -Pw |
| closets and drawers | 1 | 0 | |
| cold and wet | 1 | 1 | |
| come and go | 4 | 1 | +I |
| come and stay | 1 | 1 | +I |
| comedy and humor | 6 | 0 | -R |
| comedy and music | 1 | 0 | |
| comedy and satire | 1 | 1 | |
| comfortable and cool | 1 | 0 | |
| commercially and scientifically | 1 | 0 | -I |
| commoners and kings | 1 | 1 | -Pw |
| complete and unabridged | 1 | 1 | +R |
| completely and unselfishly | 1 | 1 | |
| confuse and disorient | 1 | 0 | |
| congressional and presidential | 1 | 1 | +I |
| conscious and non-instinctive | 1 | 0 | +A,-R |
| consider and rate | 1 | 1 | +I |
| convicted and tried | 1 | 0 | -I |
| cook and eat | 1 | 1 | +I |
| cooked and shelled | 1 | 1 | +I |
| cordial and loyal | 1 | 1 | |
| correct and erase | 1 | 0 | -I |
| country and western | 1 | 1 | +Pw |
| crack and whine | 1 | 0 | -I |
| cracked and snarled | 1 | 0 | |
| cried and sat | 1 | 0 | |
| crime and sports | 1 | 1 | |
| crochet and knit | 1 | 1 | |
| cross-stitching and painting | 1 | 1 | |
| cruel and unusual | 1 | 1 | |
| culturally and socially | 1 | 1 | |
| cumulatively and individually | 1 | 0 | -Pt |
| cut and dried | 1 | 1 | +I |
| cut and dry | 1 | 1 | +I |
| dad and mother | 1 | 0 | +Pw,-Pt |
| dancing and dinner | 1 | 0 | -I,-Pw |
| daughter and son | 1 | 0 | -Pw |
| day and night | 6 | 0.5 | +Pt |
| dead and hideous | 1 | 1 | +I |
| dear and near | 2 | 0.5 | |
| deceptive and frothy | 1 | 0 | |
| decisively and wisely | 1 | 1 | |
| deer and trees | 1 | 1 | +Pt |

| | | | |
|---|---|---|---|
| deliberately and slowly | 1 | 0 | -R |
| demographic and economic | 1 | 0 | |
| despoiling and sacking | 1 | 0 | |
| develops and markets | 2 | 1 | +I |
| dilates and relaxes | 1 | 0 | -I |
| diminishing and dwindling | 1 | 0 | |
| directly and immediately | 1 | 0 | |
| dirty and dusty | 1 | 0 | |
| dirty and greasy | 1 | 0 | |
| dirty and mean | 1 | 0 | |
| dirty and tough | 1 | 0 | |
| discarded and explored | 1 | 0 | -I |
| distributes and makes | 2 | 0 | -I |
| down and out | 1 | 1 | +Pt |
| down and up | 17 | 0 | -Pt |
| dresses and suits | 2 | 0.5 | |
| drinking and eating | 1 | 0 | -Pw |
| drinks and food | 1 | 0 | -Pw |
| dry and high | 3 | 0 | |
| dry and hot | 2 | 0 | |
| dubious and surprised | 1 | 0 | -I |
| dull and gray-looking | 1 | 1 | |
| easily and swiftly | 1 | 0 | |
| east and west | 3 | 1 | +Pt |
| easy and fast | 1 | 0 | |
| economic and educational | 1 | 0 | |
| economically and physically | 1 | 1 | |
| effectively and purposively | 1 | 0 | -I |
| eighth and ninth | 1 | 1 | +I |
| elementary and high-school | 1 | 1 | +I |
| elsewhere and there | 2 | 0 | -R |
| emeralds and peanuts | 2 | 0 | +Pw |
| emotion and meaning | 1 | 1 | |
| ending and starting | 1 | 0 | -I |
| ends and odds | 12 | 0 | |
| energetic and young | 1 | 0 | |
| engineering and psychology | 1 | 0 | |
| enthusiastically and punctually | 1 | 0 | |
| erroneous and unconstitutional | 1 | 1 | +R |
| evasive and shifty | 1 | 1 | |
| everybody and everything | 1 | 0 | +Pt |
| excessive and unjustified | 1 | 1 | +R |
| exercise and fitness | 2 | 0.5 | -R |
| fairly and fully | 1 | 0 | |
| family and friends | 3 | 0.67 | +Pt |

| | | | |
|---|---|---|---|
| fancy-free and foot-loose | 1 | 0 | |
| far and wide | 1 | 1 | |
| felt and seen | 2 | 0 | -Pt |
| few and unfavorable | 1 | 1 | +Pt |
| figuratively and literally | 1 | 0 | -Pt |
| file and rank | 1 | 0 | |
| finance and trade | 1 | 0 | |
| fired and restructured | 1 | 1 | +I |
| firm and healthy | 1 | 1 | |
| first and foremost | 6 | 1 | |
| first and only | 2 | 1 | +R |
| fiscal and monetary | 2 | 0 | -R |
| fit and straighten | 1 | 0 | -I |
| fit and wiry | 1 | 0 | |
| fits and starts | 1 | 1 | |
| flowers and roses | 1 | 1 | +R |
| follow and understand | 1 | 1 | +I |
| fought and won | 1 | 1 | +I |
| frankly and simply | 1 | 1 | |
| fresh and nice | 1 | 0 | |
| friendlily and sleepily | 1 | 0 | |
| fronts and toe | 1 | 0 | |
| fruit and nuts | 1 | 1 | |
| fully and truly | 1 | 0 | |
| funny and superficial | 1 | 1 | |
| further and unnecessarily | 1 | 1 | |
| fuzzy and warm | 1 | 0 | |
| garden and lawn | 6 | 0 | |
| gentle and kind | 1 | 0 | |
| gentler and kinder | 1 | 0 | |
| gently and lightheartedly | 1 | 1 | |
| geographical and socio-economic | 1 | 1 | |
| go and vote | 2 | 1 | +I |
| gold and silver | 4 | 1 | +Pw |
| good and right | 1 | 0 | |
| good and thick | 2 | 1 | |
| gradually and smoothly | 1 | 1 | |
| grapefruit and oranges | 1 | 0 | -Pw |
| greatest and latest | 1 | 0 | |
| greens and pinks | 2 | 0 | |
| grow and produce | 1 | 1 | |
| harass and punish | 1 | 1 | +I,-Pw |
| hard and straight | 1 | 0 | |
| head and tail | 1 | 1 | +Pt |
| hear and see | 1 | 0 | -Pt |

| | | | |
|---|---|---|---|
| heavily and slowly | 1 | 0 | |
| hell and peacocks | 1 | 1 | |
| help and serve | 1 | 1 | |
| here and there | 16 | 1 | +Pt |
| hid and knelt | 1 | 0 | -I |
| hid and went | 1 | 0 | |
| high and inside | 1 | 1 | +Pt |
| hit and killed | 1 | 1 | +I |
| hoarsely and quietly | 1 | 1 | |
| honest and stupid | 1 | 1 | +Pt |
| honey and milk | 1 | 0 | |
| hope and pray | 1 | 1 | +R,-Pw |
| hopefully and ingeniously | 1 | 0 | |
| hurtling and plunging | 1 | 1 | |
| ice and snow | 3 | 0.67 | |
| icky and rainy | 1 | 0 | |
| improperly and unfairly | 1 | 1 | |
| in and out | 3 | 1 | +R |
| inaccurate and inappropriate | 1 | 0 | |
| incest and rape | 13 | 0 | |
| incredulity and shock | 1 | 0 | |
| inflame and tear | 1 | 0 | -I |
| informally and often | 1 | 0 | |
| inhumane and terrible | 1 | 0 | -R |
| innately and pathologically | 1 | 1 | |
| insidiously and softly | 1 | 0 | |
| install and make | 1 | 0 | -I |
| install and manufacture | 1 | 0 | -I |
| intellectual and political | 1 | 0 | |
| interest and principal | 7 | 0.29 | -I,-Pw |
| international and public | 1 | 0 | -Pt |
| international and social | 1 | 0 | -Pt |
| irony and satire | 1 | 0 | +R |
| irregularly and slowly | 1 | 0 | |
| irritable and tense | 1 | 0 | |
| ivory and sandalwood | 1 | 1 | |
| jumping and running | 1 | 0 | |
| kind and playful | 1 | 1 | |
| landings and takeoffs | 1 | 0 | -I |
| laptop and notebook | 1 | 1 | |
| laugh and wink | 1 | 1 | |
| left and right | 4 | 0.5 | -Pt |
| lengthily and seriously | 1 | 0 | |
| lighthearted and witty | 1 | 1 | |
| linguist and therapist | 1 | 0 | |
| linguistic and paralinguistic | 1 | 1 | +R |
| linguistics and psychiatry | 1 | 1 | |
| linguists and psychotherapists | 1 | 0 | |
| logically and objectively | 1 | 1 | |
| lost and loved | 1 | 0 | -I |
| lurched and stumbled | 1 | 1 | |
| magazines and newspapers | 3 | 0.67 | |
| maneuvered and raced | 1 | 0 | |
| math and science | 4 | 0.75 | |
| math and sciences | 1 | 1 | |
| mechanically and systematically | 1 | 0 | |
| medicines and yeast | 1 | 1 | +Pw |
| men and women | 15 | 1 | +Pw |
| mental and physical | 4 | 0.5 | -Pt |
| messy and negligent | 1 | 1 | |
| mirrors and smoke | 1 | 0 | |
| modern and new | 1 | 0 | |
| months and years | 3 | 1 | +I |
| morally and spiritually | 1 | 0 | |
| morally and totally | 1 | 0 | |
| nagging and stress | 1 | 0 | |
| neatly and sweetly | 1 | 1 | |
| needlework and sewing | 1 | 0 | |
| needs and wants | 2 | 1 | +Pw |
| newspaper and radio | 1 | 1 | |
| nice and relaxed | 1 | 1 | |
| nice and small | 1 | 1 | |
| nice and sunny | 2 | 1 | |
| nice and toasty | 1 | 1 | |
| nights and weekends | 2 | 1 | +I |
| non-poetry and poetry | 1 | 0 | -R |
| north and south | 5 | 1 | +Pt |
| now and then | 12 | 1 | +R,+Pt |
| obtained and provisioned | 1 | 1 | +I |
| off and on | 7 | 0.86 | -R |
| offbeat and outspoken | 1 | 0 | |
| officially and publicly | 3 | 0.33 | |
| old and ratty | 1 | 1 | |
| older and wiser | 1 | 1 | +Pt |
| open and shut | 2 | 1 | +R |
| operates and owns | 6 | 0 | |
| packages and sells | 1 | 1 | +I |
| parents and students | 1 | 0 | |

| | | | |
|---|---|---|---|
| patients and psychiatrists | 1 | 0 | -Pw |
| people and soils | 1 | 1 | +Pt |
| pepper and salt | 3 | 0 | -Pw |
| perfectly and universally | 1 | 0 | |
| persistent and substantial | 1 | 0 | |
| pies and puddings | 1 | 0 | |
| pineapple and strawberry | 1 | 1 | |
| playbacks and study | 1 | 1 | +I |
| powerfully and tersely | 1 | 0 | |
| pressure and stress | 1 | 0 | |
| pride and recognition | 1 | 1 | |
| printed and sold | 1 | 1 | +I |
| productive and sane | 1 | 0 | |
| proposed and taught | 1 | 1 | +I |
| pull and tug | 1 | 1 | +R |
| push-ups and sit-ups | 1 | 0 | |
| quickly and silently | 1 | 1 | |
| quilting and sewing | 1 | 0 | |
| radically and structurally | 1 | 0 | |
| radio and television | 6 | 0.5 | |
| rapid and sharp | 1 | 0 | |
| real and vibrant | 1 | 0 | |
| realistically and seriously | 1 | 1 | |
| really and truly | 3 | 1 | |
| rebuild and reestablish | 1 | 1 | |
| received and sought | 1 | 0 | -I |
| register and vote | 1 | 1 | +I |
| rent and tuition | 1 | 0 | |
| represents and serves | 1 | 0 | |
| rich and spoiled | 1 | 1 | |
| ridiculous and terrible | 1 | 1 | |
| rise and shine | 1 | 1 | +I |
| roaring and whirling | 1 | 0 | |
| robbed and shot | 1 | 1 | |
| romance and snobbery | 1 | 1 | +Pt |
| rumbles and smolders | 1 | 1 | |
| scabrous and unclean | 1 | 1 | |
| second and third | 4 | 1 | +I |
| see and wait | 4 | 0 | -I |
| semiconductors and supercomputers | 1 | 1 | |
| severable and unconstitutional | 1 | 0 | +R,-I |
| severable and void | 1 | 0 | |
| share and understand | 1 | 0 | -I |
| shots and shouts | 1 | 1 | |
| sing and snap | 1 | 1 | +I |
| sit and wait | 3 | 1 | |
| sitting and staring | 1 | 1 | |
| sitting and watching | 1 | 1 | |
| skillful and startling | 1 | 0 | |
| skirts and sweaters | 1 | 1 | |
| slowed and stopped | 1 | 1 | +I |
| smashed in and torn | 1 | 0 | |
| smiling and winking | 1 | 1 | |
| softly and triumphantly | 1 | 1 | |
| stained and waxed | 1 | 1 | +I |
| successfully and vigorously | 1 | 0 | -I |
| summer and winter | 1 | 1 | |
| talked and wrote | 1 | 1 | |
| telecommunications and transportation | 1 | 0 | |
| three-sevenths and two-sevenths | 1 | 0 | -I |
| trade and transfer | 1 | 1 | |
| tried and true | 3 | 1 | +I |
| troubled and worried | 1 | 0 | -I |
| ungallant and untrue | 1 | 1 | |
| varied and wide | 1 | 0 | |
| voted and went | 1 | 0 | |

Table 1: Individual constraint alignment patterns. Significance: $^{*}$*p*<0.1; †*p*<0.05; ‡*p*<0.025; ♠*p*<0.01; ♥*p*<0.001

| | All binomials | Excl. semantic | Excl. semantic + metrical |
|---|---|---|---|

| | Tokens | | Types | | Tokens | | Types | | Tokens | | Types | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Constraint | N | % | N | % | N | % | N | % | N | % | N | % |
| RelForm | 60 | ♥80 | 32 | ♠78 | | | | | | | | |
| Icon | 128 | ♥98 | 77 | ♥99 | | | | | | | | |
| Power | 72 | †61 | 26 | *69 | | | | | | | | |
| Percept | 151 | ‡59 | 42 | ♥76 | | | | | | | | |
| *ww | 307 | ♥63 | 222 | ♥59 | 180 | ♥67 | 141 | ♥64 | | | | |
| *A>B | 337 | ♥69 | 244 | ♥65 | 196 | ♥71 | 155 | ♥67 | | | | |
| *BStr | 170 | ♥76 | 106 | ♥70 | 95 | ♥83 | 63 | ♥78 | | | | |
| Freq | 667 | ♥55 | 392 | ♥60 | 306 | ♥56 | 232 | ♥56 | 93 | †62 | 73 | ♠68 |
| VPhonemic | 384 | 53 | 211 | 46 | 167 | ♠38 | 127 | *43 | 51 | 23 | 37 | 20 |
| VPhonetic | 598 | 53 | 353 | 48 | 264 | 47 | 204 | 50 | 75 | 45 | 59 | 51 |
| VBackness | 357 | 49 | 208 | †43 | 153 | ♠39 | 130 | ‡41 | 52 | 44 | 37 | 54 |
| VHeight | 491 | ♥42 | 273 | 49 | 199 | 44 | 155 | 48 | 57 | †35 | 41 | 41 |
| CInit | 274 | *45 | 197 | 47 | 134 | 46 | 112 | 46 | 40 | ♠70 | 33 | 64 |
| CFin | 313 | ♥60 | 166 | 48 | 125 | 59 | 99 | 46 | 28 | 57 | 26 | 62 |
| Openness | 276 | ‡43 | 187 | *43 | 133 | ♠33 | 111 | ♠38 | 16 | 56 | 14 | 64 |
| Weight | 421 | 49 | 234 | †43 | 198 | ♥31 | 143 | ♠37 | 59 | 49 | 43 | 60 |
| SonorInit | 433 | ♠42 | 227 | 46 | 159 | 45 | 124 | 46 | 37 | 43 | 34 | 47 |
| SonorFin | 200 | 51 | 120 | 48 | 79 | 56 | 66 | 54 | 14 | 50 | 12 | 42 |
| Alpha | 692 | ♥58 | 411 | 52 | 306 | †44 | 241 | 46 | 96 | 45 | 75 | 49 |

Table 2: Associations between metrical constraints. Values given for entire dataset; results are similar when semantically aligned binomials are excluded. All results are significant at *p*<0.001.

| Odds ratio $\theta$ of activity | | |
|---|---|---|
| | *ww | *BStr |
| *A>B | 36.87 | 22.74 |
| *ww | — | 9.92 |

| Proportion $\pi_{align}$ of like alignment | | |
|---|---|---|
| | *ww | *BStr |
| *A>B | 0.767 | 0.980 |
| *ww | — | 0.989 |

Table 3: Association of Frequency constraint with semantic and metrical constraints. Correlation with iconicity was not significant. (**p*< 0.1; ***p*<0.05; ****p*<0.01; $^{\dagger}p$<0.001; results otherwise insignificant)

| | RelForm | Percept | Power | *A>B | *ww | *BStr |
|---|---|---|---|---|---|---|
| $\theta$ | ∞ | ∞ | 0.22** | 0.51 | 0.66 | 0.60 |
| $\pi_{align}$ | $0.77^{\dagger}$ | $0.68^{***}$ | 0.67* | $0.69^{\dagger}$ | 0.56** | $0.70^{\dagger}$ |

Table 4: Coefficient values for logistic regression models. Type: weighting by input binomial type. Token: weighting by input binomial token.[27]

| Constraint | Type | Token | Constraint | Type | Token |
|---|---:|---:|---|---:|---:|
| Icon | 5.85 | 4.61 | Alpha | 0.03 | 0.34 |
| Percept | 1.55 | 1.31 | Cinit | 0.03 | 0.16 |
| Power | 1.08 | -0.20 | Sfin | -0.05 | 0.40 |
| RelForm | 1.07 | 1.31 | Open | -0.06 | -0.17 |
| BStr | 0.44 | 0.99 | Back | -0.13 | -0.17 |
| *A>B | 0.42 | 0.76 | Sinit | -0.13 | -0.22 |
| Freq | 0.26 | 0.16 | Cfin | -0.18 | 0.44 |
| VLen2 | 0.23 | 0.52 | VLen1 | -0.20 | -0.38 |
| ww | 0.17 | 0.05 | Height | -0.33 | -0.69 |

Table 5: Evaluation for OT, StOT and logistic regression models of binomial ordering

| Evaluation type | Weighting | OT | StOT | Logistic Regression |
|---|---|---:|---:|---:|
| Hard | Type | 76.3% | 72.6% | 76.6% |
| | Token | 71.4% | 74.5% | 79.2% |
| Soft | Type | ∞ | 0.526 | 0.440 |
| | Token | ∞ | 0.507 | 0.396 |

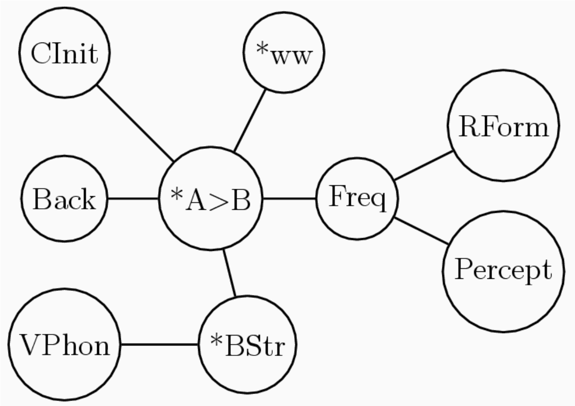

Figure 1: Plausible direct correlations between constraint alignments.

[1] The character Archie Bunker from All in the Family is seen as a sort of working class 'Everyman'.

[2] While there are alterations that do not involve sewing, we assume that most do involve some sewing.

[3] The two semantic constraint/binomial combinations for which we initially disagreed on direction of alignment were Perception for physical and mental, and Power for black and white. After brief discussion we came to agreement that the ordered form physical and mental is aligned with rather than against Perception, as physical is more concrete and mental is more abstract. Black and white was a more difficult judgment. Initially, the first author had classified the form as aligned against Perception (white being the unmarked color of a page and black being the marked color of writing on a page) and with Power (black being a stronger color than white); the second author had classified the form as aligned against Power (on the basis that white is stereotypically associated with institutional power in English-speaking societies). We ultimately decided that black and white were insufficiently asymmetric with respect to the properties of the colors themselves to judge them on this basis. However, we both accepted that white as a social category is more closely associated with institutional power than black, and therefore the binomial should be judged as aligned against Power.

[4] The precise pattern we matched was the uninterrupted phrasal sequence V[^P] CC V[^P] * inside a VP, where V[^P] is any node label starting with a V other than VP (i.e., a lexical verbal node in the Treebank), and * is any node label. The Treebank does not annotate the intermediate lexical V node in a V and V coordination, hence the rule matches are of the form VP -> V and V <Complement>. Note that punctuation was not allowed to intervene between phrases, ruling out sequences such as 'V, and V, NP'.

[5] Veatch provides convincing acoustic and phonological evidence that post-vocalic /r/ should be analyzed as a glide and that it cannot occur after a diphthong in the same syllable. Therefore, we considered words like fire as disyllabic. However, since we have no phonological evidence that /l/ cannot be tautosyllabic with a diphthong, we coded /l/ simply as a tautosyllabic consonant and words like smile and snarled as one syllable.

Since there is no vowel length distinction before /r/ (i.e. [ɪr] has no long equivalent [ir]), we coded pre-/r/ vowels as short. In addition, we included glides and coda /r/s as part of the preceding vowel's length. For example, the vowel in fairly [ɛr] is longer than the vowel in fully [ʊ], and therefore the order of fully and fairly might be accounted for by vowel length.

[6] It is possible that considering more fine-grained distinctions in the categories of vowel height and backness would lead to different results. Pinker and Birdsong (1979) follow Ladefoged (1975) in their coding of first-formant frequency (roughly, height):

> i > ɪ > ɛ > æ > ɑ > ʊ > ɔ > u

and second-formant frequency (roughly, backness):

> i > u > ɪ > ʊ > ɛ > ɔ > æ > a

Speakers may actually be sensitive to these small differences, and future studies might code with this in mind.

[7] We predict that the distinction between ambisyllabicity and tautosyllabicity will have an effect on the ordering of binomials: between items with the same vowels, the more closed syllable will be preferred in the B slot. Although our corpus does not have minimal pairs in which to analyze syllable weight differences like these, future research could test them with stimuli like:

> Which sounds better, A or B?
> A. zinner and zinder
> B. zinder and zinner

[8] Bolinger's exploration of sonority used the following hierarchy (1962:40):

> vowels > voiced continuants > voiced stops and affricates > unvoiced continuants > unvoiced stops and affricates.

However, we used a hierarchy more accepted in phonology. It is possible that coding according to Bolinger's hierarchy would yield different results. Also, we coded affricates to be single consonants. Perhaps considering them two consonants would affect the outcome.

[9] For constraints $C_1$ and $C_2$ with counts as follows (a=active, i=inactive): <$C_1$ inactive, $C_2$ inactive>=$c_{ii}$, <$C_1$ active, $C_2$ inactive>=$c_{ai}$, <$C_1$ inactive, $C_2$ active>=$c_{ia}$, <$C_1$ active, $C_2$ active>=$c_{aa}$, the odds ratio $\theta$ is defined as $(c_{ii}\times c_{aa})/(c_{ia}\times c_{ai})$. If the odds of one constraint being active are a/i when the other constraint is inactive, then the odds are $\theta\times(a/i)$ when the other constraint is active. The odds ratio has not seen frequent use in analysis of linguistic variation, but see Hazen 2002 for one example of its application.

[10] The difference between these alignment ratios for inactive versus active semantic constraints is borderline significant: $p<0.05$ for tokens, $p<0.1$ for types.

[11] Of the 127,042 words in the Carnegie Mellon Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict), 14.7% are vowel-initial, but of its 16,533 monosyllabic words, only 4.2% are vowel-initial.

[12] As noted at the beginning of Section 4, isolated significant results involving only token counts, such as the significant trend against sonority satisfaction for all tokens, are generally spurious, skewed by common frozen binomials.

[13] Logistic regression is commonly used by sociolinguists in the VARBRUL program (Cedergren and Sankoff 1974, Paolillo 2001). Although it is may more frequently be perceived as a tool for statistical analysis, logistic regression can equally be seen (and was originally introduced, e.g., Cedergren 1973) as a grammatical model of variable realization, assigining a probability for each possible output o of a given input i. In fact, it is precisely this formulation as a model of variable realization, combined with the desirable learnability properties mentioned in Section 5.4, that makes logistic so useful a tool for statistical analysis.

Logistic regression is also intimately related with maximum entropy modeling, a state-of-the-art machine learning technique in widespread use in computational linguistics. We take comments in the literature regarding maximum-entropy modeling to apply equally to logistic regression as we use it here.

[14] Although Ernestus and Baayen report that StOT performed better than logistic regression in modeling segment neutralization variation, their StOT models had more free parameters than their logistic regression models. As a result, the direct comparison is not entirely appropriate.

[15] In StOT, the probability that a lower-ranked constraint can outrank a higher-ranked constraint at evaluation time must be less than 1/2, so the probability that any one of a host of n lower-ranked constraints outranks a higher-ranked constraint must be less than $1-(1/2)^n$. In logistic regression, on the other hand, weaker constraints C2 and C3 can outrank a stronger constraint C1 with arbitrarily high probability, as the difference of their weights (β2+β3-β1) can be arbitrarily large (see Equation 1).

[16] In StOT, the constraint component of a grammar consists of a set of constraints $\{C_i\}$ plus real-valued rankings $\{R_i\}$ for each constraint. In addition, there is a fixed noise factor E associated with the grammar. At the time of evaluation, a final constraint ranking $\{R'_i\}$ is determined as follows: for each constraint $C_i$, the output ranking $R'_i$ is determined by sampling from the normal distribution with mean $R_i$ and variance E. Closely-ranked constraints (with respect to E) vary in their post-evaluation order, leading to variability in output.

[17] For a more comprehensive explication of Stochastic Optimality Theory and the Gradual Learning Algorithm, see Boersma and Hayes 2001. As our results in Section 5.3 show, the GLA is not guaranteed to reach a global optimum.

[18] In our experiments we used initial constraint rankings of 0 for all constraints; evaluation noise of 0.1; a constant learning plasticity of 0.001; and 100,000 learning iterations. To determine the probability of binomial orderings after learning, we sampled 1,000 times for each input from post-evaluation constraint rankings and used the sample distribution of output rankings.

[19] We trained logistic regression models using the glm routine of the R statistical software package (R Development Core Team 2004), which fits the model minimizing least squares.

[20] The KL divergence of distribution *q* from distribution *p*, where *q* and *p* are defined over a set *S*, is mathematically defined as

$$D(q \parallel p) \equiv \sum_{s \in S} q(s) \log \frac{q(s)}{p(s)}$$

Intuitively, $D(q||p)$ can be thought of as the penalty incurred for using $p$ to encode $q$. For all $q$ and $p$, $D(q||p) \geq 0$, and $D(q||p)=0$ only if $q=p$. Also, if there is some $s \in S$ such that $p(s)=0$ but $q(s)>0$, $D(q||p)$ becomes infinite. As a result, the KL divergence of our corpus from the traditional OT ranking given in Section 5.1 is infinite.

[21] In the most extreme case, if an input type {A,B} is realized half the time as A and B and half the time as B and A, then any model will have the same hard evaluation accuracy; but the closer the model's predicted output frequency is to 50/50, the better its soft evaluation accuracy.

[22] Jäger and Rosenbach achieved a much closer fit to their dataset, as measured by relative entropy, than we achieved. This is most likely due to the fact that the ratio of constraint violation profiles to constraints is much higher for our data and constraint set, meaning that our model has relatively fewer degrees of freedom with which to fit the data.

[23] Note that minority orderings such as our single token of friends and family, where the majority ordering family and friends is consistent with the highest-ranked constraint, are not candidates for ganging up.

[24] We also attempted the same experiment with the token-trained logistic regression model, which yielded 23 ostensible binomial types with ganging up, including evasive and shifty; but inspection suggested that the results for this model were too badly skewed by inflated weights for non-semantic constraints to draw strong conclusions.

[25] The magnitude of the Height constraint in the overall logistic regression model of Table 4 is larger than that of the clearly important Frequency constraint. As noted in Section 4.4, however, we found no relevant subset of the data in which Height is significantly correlated with binomial ordering. We also investigated the constraint by building a variety of smaller logistic regression models that excluded smaller-magnitude constraints. We were able to use the likelihood-ratio ($G^2$) test to determine that Height never made a significant contribution to any of these smaller models. In addition, the magnitude of the Height constraint tended to increase as we increased the number of constraints in the model. This combination of evidence suggests to us either that the relatively large magnitude of Height in the full logistic regression model of Table 4 is overfitting, or that Height is important in such a narrow subset of our data that we our sample size is too small for us to demonstrate its effects.

[26] Ideally, a corpus of recorded speech could be used to incorporate actual realized vowel lengths into the model, in addition to regional variation and idiosyncracies of pronunciation.

[27] Our pragmatic ordering constraint is never violated in our dataset and therefore receives an arbitrarily high weight in logistic regression. We do not list it in Table 4.