Abstract knowledge versus direct experience in processing of binomial expressions

Emily Morgan[a,b,*], Roger Levy[a,c]

[a]*Department of Linguistics, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0108, United States*
[b]*Department of Psychology, Tufts University, 490 Boston Ave, Medford, MA 02155, United States*
[c]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, United States*

Abstract

We ask whether word order preferences for *binomial expressions* of the form *A and B* (e.g. *bread and butter*) are driven by abstract linguistic knowledge of ordering constraints referencing the semantic, phonological, and lexical properties of the constituent words, or by prior direct experience with the specific items in questions. Using forced-choice and self-paced reading tasks, we demonstrate that online processing of never-before-seen binomials is influenced by abstract knowledge of ordering constraints, which we estimate with a probabilistic model. In contrast, online processing of highly frequent binomials is primarily driven by direct experience, which we estimate from corpus frequency counts. We propose a trade-off wherein processing of novel expressions relies upon abstract knowledge, while reliance upon direct experience increases with increased exposure to an expression. Our findings support theories of language processing in which both compositional generation and direct, holistic reuse of multi-word expressions play crucial roles.

Keywords: binomial expressions, word order, sentence processing, comprehension, frequency

1. Introduction

When we encounter common expressions like *I don't know* or *bread and butter*, do we process them word-by-word or do we treat them as holistic chunks? Research on sentence processing has largely focused on how single words are combined into larger utterances, but intuitively it seems that high frequency multi-word expressions might be processed holistically, even if they could in principle be treated compositionally. Recent research has thus questioned what possible sizes of combinatory units should be considered as the building blocks of sentence processing: Must all multi-word expressions be generated compositionally each time they are used, or can the mental lexicon contain holistic representations of some multi-word units?

The primary diagnostic for this question is whether the frequency of occurrence of multi-word expressions is predictive of their behavior in language processing. Such frequency effects are well documented at the level of individual words: more frequent words are faster to read (Inhoff and Rayner, 1986; Rayner and Duffy, 1986; Rayner et al., 1996), more likely to be skipped in reading (Rayner et al., 1996; Rayner and Well, 1996), and more susceptible to phonetic reduction (Bybee, 1999; Gregory et al., 1999). But do comparable frequency effects exist for multi-word expressions, when the frequency of their component words is controlled for? If the frequency of a given expression is being mentally stored, this implies that there is a mental representation of the expression as a whole. In contrast, if there are no frequency effects at the level of multi-word expressions, this is evidence against them having holistic representations akin to those of individual words.

A traditional view of grammar does not include holistic representations of multi-word expressions. According to this view, there is a strict separation between the individual words of a language and the rules

[*] Corresponding author: *emily.morgan@tufts.edu*; (617) 627-3523

for combining them. Pinker (2000), for example, describes a "traditional words-and-rules theory" in which "there are two tricks, words and rules. They work by different principles, are learned and used in different ways, and may even reside in different parts of the brain." (See also Ullman, 2001; Ullman et al., 2005.) One tenet of this theory is that forms which can be generated compositionally are not stored: for instance, in the case of the English past tense, irregular forms are stored, while regular forms are generated anew using the *-ed* suffix each time they are used (Pinker, 1991). It remains possible within this theory that some regular forms—particular extremely high frequency ones—may be stored as well, but this is not the general method for dealing with such forms. As Pinker (2000) explains, one key motivation for this theory is memory constraints on the representation of language knowledge: it is more efficient to store a single, widely applicable rule than to store each regular form individually.

In a similar vein, this theory predicts that multi-word expressions should not be stored holistically because they can be generated compositionally, except in the case of non-compositional exceptions such as idioms (Swinney and Cutler, 1979). Again, as with regularly inflected wordforms, some exceptions may exist, but the exponentially larger number of multi-word expressions with which people have experience makes it even less likely that these expressions would be stored holistically, given the motivating concern with storage efficiency. The words-and-rules theory thus does not predict that the processing of a multi-word expression will be affected by the frequency of the expression as a whole, though it can be affected by the frequencies of the individual words making up the expression.[1]

In contrast, there exists a growing movement of grammatical theories that do not draw a sharp distinction between the lexicon and the combinatory rules (e.g. Langacker, 1987; Johnson, 1997, 2006; Pierrehumbert, 2000; Bybee, 2001, 2006; Goldberg, 2003; Gahl and Yu, 2006; Hay and Bresnan, 2006; Baayen et al., 2011; van den Bosch and Daelemans, 2013). Rather than conceiving of rules as static entities dissociated from the lexicon, these *usage-based* approaches instead conceive of rules as dynamically generated generalizations over one's linguistic experience. In particular, many of these approaches (notably Bybee, 2001; Hay and Bresnan, 2006, among others) claim that people mentally store *exemplars*, or tokens of linguistic experience, which can be larger than single words. Language users then form generalizations from exemplars at multiple levels of granularity (e.g., morpheme, word, or phrase) simultaneously, and the resulting network of generalizations constitutes our grammatical knowledge. Single words and multi-word expressions are thus on an equal footing: both are possible units that can be inferred from exemplars, and frequencies of multi-word expressions are predicted to be stored and tracked just as frequencies of single words are.

Similar claims are made by exemplar-based computational models, which, like the exemplar-based grammatical theories, can incorporate combinatorial units of varying sizes from morphemes to sentences (e.g. Bod, 1998; Bod et al., 2003; Bod, 2008; Pierrehumbert, 2000; Johnson et al., 2007; O'Donnell et al., 2011; Post and Gildea, 2013). Within these models, the process of learning a grammar is explicitly one of deciding what sizes of units are most applicable or probable to explain the available language data. Under the learned grammars, many utterances can be parsed in multiple ways, either as combinations of individual words, or as holistic expressions, or various combinations thereof.

Evidence for these usage-based theories in the domain of multi-word expressions comes in large part from previous demonstrations of phrase-level frequency effects. Bybee (2006) reviews numerous corpus analyses demonstrating that the frequency of multi-word expressions is predictive of phonological reduction, grammaticalization, and other properties of usage, with a focus on highly frequent expressions such as *I don't know* or *going to*. Frequency effects for multi-word expressions have also been demonstrated in a controlled experimental setting: in a phrasal-decision task (analogous to a lexical decision task), Arnon and Snider (2010) found that more frequent phrases—e.g. *Don't have to worry*—were judged to be sensible phrases of English faster than less frequent phrases matched for word and substring frequencies—e.g. *Don't have to wait*. They further demonstrate that these effects exist across a wide range of frequencies, not just at the highest end of the frequency spectrum. (For a comparable finding using phonetic duration in corpus

[1] It may be possible to accommodate frequency effects for multi-word expressions under this theory, depending upon further details of the parser. In particular, processing of later words in an expression could be conditioned upon earlier words, thus creating an overall frequency difference. But this is not a direct prediction of the words-and-rules theory.

data, see Arnon and Cohen Priva, 2013. Similar frequency effects have also been found in child language acquisition; see Bannard and Matthews, 2008.)

The exemplar-based approach also accords with more recent work on idioms, which challenges the traditional notion of idioms as strictly non-compositional. Gibbs (1990) and Nunberg et al. (1994) argue that many idioms can be seen as conventionalized metaphoric extensions of their literal meanings, and thus need not be treated as exceptions to the prevailing rules. (Similarly, see Holsinger, 2013.) On the whole, we thus see a broad shift towards recognizing that many expressions reside in a grey zone between entirely compositional and entirely non-compositional, and furthermore that an expression may be conventionalized while still being at least somewhat compositional.

But there remain open questions regarding these exemplar-based approaches and the interpretation of frequency effects for multi-word expressions. One limitation in the work to date is that it is difficult to differentiate the effects of language experience per se from the effects of real-world knowledge. Bybee (2006), for example, stresses the importance of language experience:

> As is shown here, certain facets of linguistic experience, such as the frequency of use of particular instances of constructions, have an impact on representation that we can see evidenced in various ways. . .

However, much of her cited evidence conflates linguistic experience with real-world experience. For example, in the phonological reduction of extremely frequent phrases such as *I don't know*, is this reduction due to the frequency of the linguistic expression per se, or is it due to the frequency of the event of not knowing something? Similarly, in the case of Arnon and Snider's contrast between phrases such as *Don't have to worry* and *Don't have to wait*, there could be a difference in the real-world likelihood of the events described by these expressions, which causes faster processing due to the difference in conceptual predictability, as opposed to linguistic predictability.[2] In general, this confound between linguistic experience and real-world knowledge exists whenever one compares expressions describing different real-world events.

Another outstanding question is how to empirically measure the trade-off between the reuse of stored multi-word expressions and the compositional generation of expressions. In the case of novel or infrequently attested expressions, we assume that such expressions must be processed compositionally using abstract linguistic knowledge—that is, generalized knowledge that is not bound to specific lexical items or expressions. In the case of frequently attested expressions, two potential processing strategies exist: compositional generation or reuse of stored holistic representations. Previous experimental work has primarily focused on the question of whether there is *any* reuse of stored multi-word expressions, and has suggested that there is at least some, but it remains possible that even very frequent and conventionalized multi-word expressions could in part or at times also be generated anew using abstract knowledge. Thus the major question now is *to what extent* both holistic reuse and compositional generation play a role in language processing (Wiechmann et al., 2013). As mentioned above, computational models have attempted to address this question by simulating what combination of linguistic units of varying sizes most parsimoniously predict corpus data (Bod et al., 2003; O'Donnell et al., 2011; Post and Gildea, 2013). But there has been no attempt so far to directly measure the competing influences of reuse and generation via behavioral experimentation.

Our work here does just that: we will quantify the extent to which people's processing of attested expressions is influenced by their frequency of direct experience with those specific expressions versus by the abstract linguistic knowledge that allows them to generate such expressions compositionally. To do so, we need to investigate a linguistic construction for which we can independently estimate people's frequency of direct experience and their abstract knowledge of its composition. Moreover, we want a construction with wide variation in how frequently attested specific instances of the construction are, so that we can measure how the influence of these competing explanations changes as a function of the overall frequency of an expression. For these reasons, an ideal construction is *binomial expressions*.

[2] Arnon and Snider did attempt to control for this real-world likelihood difference by collecting plausibility ratings for their materials, which they demonstrated did not differ in plausibility between conditions. However, plausibility in all conditions was very high, so extent differences may not have been detected due to ceiling effects.

1.1. Binomial expressions

In this paper, we will address the generation and reuse of multi-word expressions by focusing on *binomial expressions* of the form *A and B*, such as *bread and butter* or *sweet and sour*. We include in our definition of binomial expressions all potential items with this form, including unattested expressions (e.g. *bishops and seamstresses*). Although binomial expressions are sometimes taken to include expressions with other conjunctions (e.g. *or*), here for simplicity we consider only expressions joined with *and*. Many binomial expressions have a preferred order (e.g. not *butter and bread* or *sour and sweet*), but binomials vary in how strong these ordering preferences are: some binomials are entirely fixed in order, or *frozen* (e.g. *safe and sound/*sound and safe*), while others are quite free (e.g. *television and radio/radio and television*). Binomial expressions are thus a case of multi-word expressions that vary along two dimensions: how frequent they are, and how conventionalized their order is.

What causes binomial ordering preferences? One possibility is that preferences arise from abstract linguistic constraints that reference phonological, semantic, or other lexical properties of the elements in a binomial (e.g. the shorter word should come first). An alternate possibility is that preferences are driven by direct experience with the specific binomials in question: an order is preferred because it has been experienced more often.

Binomial expressions thus allow us to study the trade-off between abstract knowledge and direct experience. Specifically, we ask whether ordering preferences for binomials expressions are driven by direct experience with these expressions or by abstract constraints on the order of their elements. Moreover, we ask whether the influence of these two knowledge sources changes as a function of the frequency of an expression.

Additionally, binomial expressions are particularly suitable for studying effects of language experience per se, as opposed to real-world knowledge or other confounds, because the formal syntactic and semantic properties of these expressions are preserved regardless of ordering. Binomial expressions thus have an inherent control condition, unlike Bybee's (2006) investigation of high frequency expressions—whose other potentially relevant linguistic properties (e.g. unigram word frequencies) are not explicitly controlled—or the use of control expressions describing different real-world events by Arnon and Snider (2010, e.g., *Don't have to worry* vs. *Don't have to wait*). We can thus study the effects of direct linguistic experience on binomial expressions by manipulating their ordering while minimizing confounds.

1.1.1. Previous work on binomial ordering preferences

Siyanova-Chanturia et al. (2011) demonstrated online effects of binomial ordering preferences: In an eye-tracking study, participants read common binomial expressions in either their preferred or dispreferred order, embedded in sentence contexts, e.g.:

(1) John showed me pictures of the *bride and groom* both dressed in blue.

(2) John showed me pictures of the *groom and bride* both dressed in blue.[3]

Expressions were read faster in their preferred order. Is this reading time difference due to the frequency of people's direct experience with these specific expressions or to their abstract knowledge of constraints on binomial ordering?

It has long been known that at least in certain contexts, binomial ordering preferences are sensitive to a variety of semantic, phonological, and lexical constraints, but the degree to which these constraints apply in online processing remains unclear. Early work portrayed these constraints as contributing to the diachronic longevity of expressions, while more recent work has suggested, albeit inconclusively, that such constraints play a role online as well.

Much of the existing work on binomial ordering preferences relies upon corpus analyses or analyses of hand-selected examples. Malkiel (1959) was the first to propose that the relationship between words in a binomial could contribute to the prominence or longevity of the expression. Based on hand-selected examples of frozen binomials, he proposes a number of constraints on ordering, both semantic and phonological, as well

[3] Binomial expressions are italicized here for clarity but were not italicized in the experiment.

as discussing other possible relationships between words (e.g., rhyming and alliteration). A more extensive study of binomial ordering preferences was carried out by Cooper and Ross (1975), whose work focuses on demonstrating a *Me First* constraint, which posits that "first conjuncts refer to those factors which describe the prototypical speaker." (This prototypical speaker is later described as "Here, Now, Adult, Male, Positive, Singular, Living, Friendly, Solid, Agentive, Powerful, At Home, and Patriotic, among other things.") They further introduce a number of phonological constraints on ordering, noting that the various constraints seem to differ in strength and may interact with each other, but they do not attempt to quantify these strengths or their interactions. Their investigation is based on a hand-selected sample of common binomial expressions, and they explicitly frame their discussion in terms of constraints that contribute to the diachronic longevity of an expression. Fenk-Oczlon (1989) introduced the idea that these constraints might apply to online processing as well as diachronic language change, arguing that most of Cooper and Ross's proposed constraints could be subsumed under the constraint that "the more frequent and therefore informationally poorer elements tend to occupy initial position" and that this new constraint is motivated by cognitive principles. His argument is supported by corpus data, but he does not provide any evidence from online processing measures. Similarly, Sobkowiak (1993), again based on corpus data, suggests that most of the previously proposed constraints can be subsumed under a principle of "unmarked-before-marked", which he relates to the information structure principle of "given before new".

More recent work has stopped attempting to unify disparate constraints and has instead focused on determining the relative rankings or weights of different constraints. In particular, Benor and Levy (2006) surveyed a large number of proposed constraints on ordering preferences from the previous literature, and considered a variety of probabilistic modeling frameworks for combining them. They found that a logistic regression model best predicts ordering preferences for a large selection of binomial expressions randomly selected from a corpus. Similarly, Mollin (2012) inferred a hierarchy of constraints from corpus data and found comparable rankings to those found by Benor and Levy.

While the existence of binomial ordering constraints in corpus data is well demonstrated, it is unclear whether these constraints apply only diachronically or whether they have synchronic cognitive status. Offline experimental tasks have suggested the synchronic cognitive reality of some constraints, mostly phonological. Using a forced-choice preference task in which subjects choose between possible orders of a binomial expressions, Bolinger (1962) demonstrated a preference to avoid having two stressed syllables in a row, comparable to findings in other domains of grammatical encoding (Jaeger, 2006; Lee and Gibbons, 2007). Pinker and Birdsong (1979) used a rating task with nonce words to argue for four phonological constraints, including "Panini's Law" (the shorter word, measured in syllables, should come first; named after a 4th Century B.C. Sanskrit linguist), as well as constraints on vowel quality, vowel length, and initial consonant obstruency. Wright et al. (2005) used a forced-choice preference task to demonstrate that male names preferentially precede female names, even when phonology and frequency are controlled for. Moreover, they showed that male names tend to have "first-position" phonological properties and are on average more frequent than female names. These offline tasks demonstrate that at least some abstract constraints on ordering are synchronically cognitively active, but they do not demonstrate whether these constraints are available during real-time language processing or whether they are available only upon later reflection.

Prior to Siyanova-Chanturia et al.'s work, a small number of online investigations used recall tasks to simulate language production, with mixed results regarding whether abstract ordering constraints are active in online production. Bock and Warren (1985) did not find effects of concreteness in ordering preferences, although the number of subjects and items in their task is small relative to the numbers we will use. Kelly et al. (1986) and Onishi et al. (2008) did find effects of prototypicality. McDonald et al. (1993) found effects of animacy and prosody, but—in contrast to Pinker and Birdsong—not word length. Thus the previous work provides weak evidence for some effects of abstract ordering constraints in production. The existence of such effects in comprehension has yet to be tested.

So based on our current knowledge, it is unclear whether to attribute the processing differences found by Siyanova-Chanturia et al. to the frequency of people's direct experience with these specific expressions or to their abstract knowledge of constraints on binomial ordering. Here we adopt a two-pronged approach to address this question. We look for effects of abstract ordering constraints on novel binomial expressions, thus establishing a baseline for such effects in the absence of direct experience with the binomials in question. Additionally, we compare the processing of these novel expressions with Siyanova-Chanturia et al.'s frequently attested expressions, allowing us to assess the relative roles of abstract knowledge and direct experience in the processing of attested expressions.

1.2. Our approach and its predictions

In this section, we describe in more detail the theoretical and methodological approach that we will take to studying binomial expressions. We begin by identifying three variables whose potential effects on processing we want to consider and determining how to quantify each one.

1.2.1. Independent variables of interest

For a word pair (A, B), the first variable we consider is the *overall frequency* of binomial expressions containing these elements—in other words, the combined frequency of the expressions *"A and B"* and *"B and A"*. To estimate the overall frequency of people's experience with these expressions, we can obtain frequency estimates from large corpora. Frequency can thus be analyzed as a continuous variable (generally measured in occurrences per million words), although in the current work we will treat it dichotomously (unattested versus frequently attested).

The next variable we consider is the *relative frequency*, or proportion of occurrences, of each order. Again, we can estimate this from corpus frequencies. The relative frequency of *"A and B"* is the number of occurrences of *"A and B"* divided by the overall frequency of (A, B) binomial expressions. It is thus a real number between 0 and 1, inclusive. The relative frequency of *"B and A"* is one minus the relative frequency of *"A and B"*.

The final variable we consider is the ordering preference due to people's *abstract knowledge* of binomial ordering constraints. For a given order *"A and B"*, we want a value between 0 and 1 corresponding to the probability of someone producing that order based on their knowledge of the abstract constraints governing binomial ordering. Unlike the previous two variables, we cannot directly estimate people's abstract knowledge from corpus frequencies. Instead, we will build a probabilistic model based on that of Benor and Levy (2006) to give us these estimates. In this paper, we make the simplifying assumption that abstract ordering preferences are fixed for a given expression; that is, they do not depend on the local context, linguistic or otherwise. This assumption would not always hold in a more naturalistic setting: in a separate corpus analysis (Morgan, 2016, chapter 3), we find that ordering preferences for 4% of tokens are directly influenced by the local linguistic context, e.g. because one element in the pair was previously mentioned. However, our experimental materials (described in Section 3) will as much as possible avoid local contexts that would influence expression order, so we consider this a reasonable simplification for the present work.

Of these variables, the two that directly compete to explain binomial ordering preferences in online processing are relative frequency and abstract knowledge. Crucially, although these two variables may be correlated, we assume that they are not equivalent, as relative frequency can be influenced by factors beyond abstract knowledge such as conventionalization and idiomaticity, famous quotations, or language change that interacts with abstract ordering constraints (e.g. changes in word meaning or pronunciation). For example, although abstract knowledge includes a strong constraint to put men before women, *ladies and gentlemen* is strongly preferred to *gentlemen and ladies* due to its conventionalized use in formal addresses. Discrepancies between abstract knowledge and relative frequency are not necessarily limited to such extreme cases as *ladies and gentlemen* but may exist in subtler ways for many expressions in the language.

We further note that the roles of relative frequency and abstract knowledge in determining ordering preferences may change depending on the overall frequency of an expression: in the most extreme case, a never-before-encountered binomial by definition cannot be influenced by its relative frequency in previous experience. Our goal is therefore to measure the relative contributions of abstract knowledge and relative frequency to binomial ordering preferences, and to determine whether and how these change as a function of overall frequency.

1.2.2. Dependent variables of interest

We consider two measures of people's processing of binomial expressions. First, we carry out a forced-choice preference experiment in which people see both possible orders of a binomial expression and choose which they prefer. For each expression, we can then calculate the proportion of people who prefer a given order. Next, we measure reading times for expressions in each order as an online measure of processing difficulty. We thus obtain two measures indexing degree of human preference for one order over other. We can then test which combination of our proposed independent variables—overall frequency, relative frequency, and abstract knowledge—best predict the human data.

1.2.3. Predictions

Let us consider possible combinations of independent variables and what effects they might have on the behavioral data data.

Abstract knowledge only. One possibility is that only abstract knowledge of ordering constraints influences processing. This would be the case if a) there are no effects of direct experience with specific binomial orders (in line with a words-and-rules theory of language processing), and b) there *are* online effects of ordering constraints. In this case, we predict that abstract knowledge but not relative frequency will have predictive power. More specifically, this theory predicts that abstract knowledge will be the best predictor of the behavioral data, and that its predictive power should not change as a function of relative or overall frequency.

Relative frequency only. If, as predicted by exemplar-based theories, there are effects of direct experience with specific binomial orders, then relative frequency should influence behavior for expressions that people have experience with, i.e. expressions with nonzero overall frequency. If, furthermore, abstract ordering constraints are *not* active in online processing, then only relative frequency should play a role. In this case we predict that novel binomial expressions will show no ordering preferences because people have no experience with them, but that relative frequency will be predictive of the behavioral data for all attested binomials. Under such a theory, relative frequency may improve as a predictor with increased overall frequency, but this would be due to having more robust estimates of relative frequency with increased overall frequency, not due to any change in the role of abstract knowledge.

Both abstract knowledge and relative frequency. If exemplar-based theories are correct that there are effects of direct experience, and moreover if abstract ordering constraints are active in online processing, then we predict that both relative frequency and abstract knowledge will be predictive of the behavioral data. For novel binomial expressions, with which people lack direct experience, abstract knowledge will be predictive. For attested expressions, some combination of abstract knowledge and relative frequency will be the best predictor (as predicted by Bod et al., 2003; O'Donnell et al., 2011; Post and Gildea, 2013).

To summarize, we investigate the roles of abstract knowledge and direct linguistic experience in the processing of both novel and frequently attested binomial expressions. We estimate people's direct experience with expressions in each possible order using corpus frequencies, and we estimate their abstract knowledge of ordering preferences using a probabilistic model. We evaluate which combination of these best predicts behavioral data in a forced-choice preference task and a self-paced reading task.

The organization of the remainder of this paper is as follows: In Section 2, we introduce the probabilistic model used to estimate abstract knowledge of binomial ordering preferences. In Section 3, we describe the experimental materials used in our behavioral experiments. In Sections 4 and 5, we discuss two experiments. Section 6 gives a general discussion.

2. Probabilistic model of ordering preferences

We begin by developing a probabilistic model of binomial ordering preferences. This model integrates the constraints on ordering that have been discussed in the previous literature (as summarized by Benor and Levy, 2006), allowing us to approximate a native English speaker's abstract of knowledge of ordering preferences for a given binomial expression, independent of their direct experience with tokens of the expression.

We develop a logistic regression model following Benor and Levy. For a given word pair (A, B), this model predicts the probability that a binomial expression will be realized as *A and B*. We train our model on Benor and Levy's dataset, a random selection of binomial expressions drawn from a collection of corpora.[4] As Benor and Levy note, conclusions drawn from token counts rather than type counts may be skewed by the presence of a small number of very frequently attested frozen expressions (e.g. *back and forth*, with a token count of 49). We thus train our model on binomial types rather than tokens. This necessitated excluding expressions that appeared in both orders (15 word pairs), leaving us with 379 binomial expression types.

Benor and Levy coded their dataset for twenty potential constraints on ordering based on a thorough review of the previous literature. A constraint is said to be *active* for a given word pair if it favors one order over another; not all constraints are active for all word pairs. When constraints are active, they are binary-valued, favoring either word A first or word B first. Specifically, constraints are coded as 1 when they favor alphabetic order, -1 when they favor non-alphabetic order, and 0 when they are inactive. Outcomes are coded as 1 if the binomial expression appears in alphabetical order and 0 otherwise.

Benor and Levy did not do any model selection to determine which of their constraints were good predictors, although their results show that some, particularly among the nonmetrical phonological constraints, are very poor predictors. For our model, we use a subset of their constraints. Our goal is to develop the best possible model of binomial expression preferences that is nonetheless reasonably parsimonious (in particular, does not include those constraints that are clearly poor predictors), but it is not our goal to conclusively demonstrate that particular constraints are significant predictors of preferences: rather, our goal is to develop an effective predictive model that can be used to investigate the link between abstract knowledge of binomial ordering preferences and behavioral responses in offline and online processing tasks. We thus adopt relatively lenient criteria for inclusion of constraints in our final model. From Benor and Levy's twenty constraints, we begin by excluding two constraints that are rarely active in the dataset, and all expressions in which they are active: the Absolute Formal Markedness constraint (the two elements do not share a derivation, but one element is structurally more simple—i.e. contains fewer morphemes; active once) and the Pragmatic constraint (ordering is directly influenced by the local linguistic context; active thrice). With the remaining constraints, we fit a logistic regression model using the `glm` function in R (R Core Team, 2014). Each constraint was entered as a predictor, with no interactions between constraints. We performed backwards model selection, excluding constraints one at a time based on their Wald z statistic, until all remaining constraints had $p < 0.15$.[5,6]

Our final model contains seven constraints. All affected the model's predicted ordering preference in the direction expected by Benor and Levy or by the sources who first proposed the constraint. See Table 1 for details of the constraint weightings. The constraints included in our final model are (with examples of binomials that satisfy each constraint drawn from the training data):

Formal markedness The word with more general meaning or broader distribution comes first. For example, in *boards and two-by-fours*, *boards* are a broader class of which *two-by-fours* is one member.

Perceptual markedness Elements that are more closely connected to the speaker come first. This constraint encompasses Cooper and Ross's (1975) 'Me First' constraint and includes numerous subcon-

[4] For reasons that could not be determined, the version of the dataset we had access to contained 689 binomial tokens, three tokens fewer than stated in Benor and Levy.

[5] We made one exception by keeping the Iconic Sequencing constraint in our model, although it had a high p value. This constraint was never violated in our dataset, and estimation of the Wald z statistic is unreliable in cases such as this with large estimated coefficients, due to inflated standard error estimates (Agresti, 2002; Menard, 2002). A likelihood ratio test supports our keeping this constraint in the model. (See Table 1.)

[6] Backwards model selection is anti-conservative (Harrell, 2001), but this is not a problem in light of the desire for leniency discussed above, as we are not attempting to draw strong conclusions about which particular constraints influence preferences. In terms of the effects on later results, including irrelevant constraints in our predictive model would add noise to our abstract ordering preference estimates, making it harder to detect effects of abstract knowledge. As we do ultimately find effects of abstract knowledge, any noise introduced at this stage was apparently not substantial enough to counteract these findings.

Constraint	*Regression coeff.*	*Std. Error*	*z value*	*p value (z)*	*Log-lik ratio*	*p value (χ^2)*
Formal Markedness	1.39	0.56	2.49	0.01	3.85	0.006
Perceptual Markedness	1.72	0.51	3.40	0.0007	7.77	0.00008
Power	1.03	0.57	1.81	0.07	1.81	0.06
Iconic Sequencing	18.62[a]	709.22	0.026	0.98	53.47	$<2\text{x}10^{-16}$
No Final Stress	0.50	0.33	1.50	0.13	1.16	0.13
Frequency	0.32	0.14	2.35	0.02	2.76	0.02
Length	0.43	0.21	2.07	0.04	2.18	0.04

[a] This coefficient is effectively infinity, as this constraint is never violated in the training data. See Footnote 5 regarding the standard error and z statistic in this case.

Table 1: Constraint weights in our probabilistic model. In addition to reporting the Wald z statistic and p-values based on it (columns 3–4) , we report results of a likelihood-ratio test comparing versions of the model differing only in whether they include the constraint in question (and containing all other constraints; columns 5–6).

straints, e.g.: animates precede inanimates; concrete words precede abstract words. For example, in *deer and trees*, *deer* are animate while *trees* are inanimate.

Power The more powerful or culturally prioritized word comes first. For example, in *clergymen and parishioners*, *clergymen* have higher rank within the church.

Iconic/scalar sequencing Elements that exist in sequence should be ordered in sequence. For example, in *achieved and maintained*, a state must be *achieved* before it can then be *maintained*.

No final stress The final syllable of the second word should not be stressed. For example, in *abused and neglected*, *abused* has final stress and should therefore not be in the second position.

Frequency The more frequent word comes first, e.g. *bride and groom*.

Length The shorter word (measured in syllables) comes first, e.g. *abused and neglected*.

The dataset on which we originally trained our model contained seven binomial expressions that were also included in Siyanova-Chanturia et al.'s (2011) items, which we later use as test items. Therefore, after doing model selection on the original dataset, we retrained our model, excluding these seven items from the training data. All results, beginning with Table 1, are reported based on the retrained model.

2.1. Model validation

We validate the model by testing its predictions on the training data and on the 42 attested binomials used by Siyanova-Chanturia et al. (2011). Constraint values for the Siyanova-Chanturia et al. binomials were hand-coded as described in Section 3. The model correctly predicts the ordering preferences for 287/372 (77%) of the training data and 30/42 (71%) of Siyanova-Chanturia et al.'s items, both significantly greater than chance (50%) in a two-tailed binomial test ($p < 0.001$ and $p < 0.01$).

3. Experimental materials

Using our probabilistic model, we develop the linguistic stimuli used in both experiments. Our stimuli consisted of 84 word pairs, with each pair producing two possible binomial expressions (*A and B* or *B and A*). 42 of our items, taken directly from Siyanova-Chanturia et al. (2011), are frequently attested. They range from almost completely frozen (e.g. *bread and butter*) to relatively flexible (e.g. *radio and television/television and radio*).

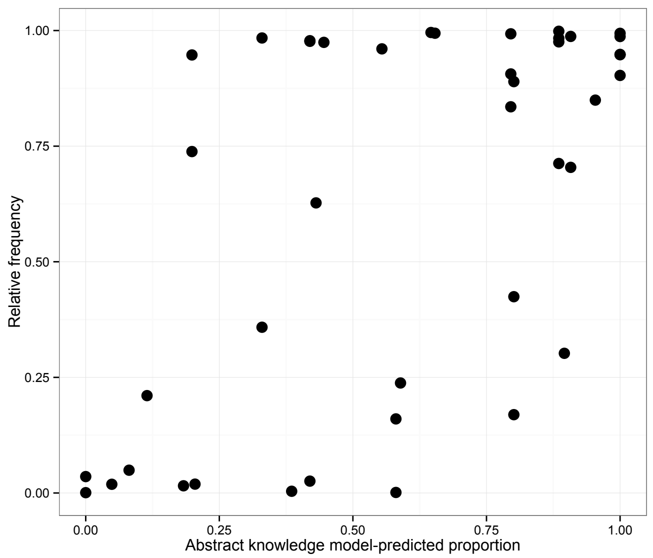

Figure 1: Abstract knowledge model-predicted proportion and empirical relative frequency of each attested binomial appearing in alphabetical order. Abstract knowledge and relative frequency are significantly but not perfectly correlated.

We further created 42 novel items which our model predicts to have strong ordering preferences (e.g. *bishops and seamstresses/seamstresses and bishops*). To ensure that speakers have no prior experience with these expressions, we consult the nearly 500-billion-word Google books n-gram corpus (Lin et al., 2012). Our novel binomials are not included in this corpus in either order.[7]

Our probabilistic model gives us an estimate of the direction and strength of ordering preference for each item based on abstract ordering constraints. To generate model predictions for these items, we must code them for the seven constraints described in Section 2. Final Stress and Length were coded by either the first author or a trained research assistant, both native speakers of American English. Frequency estimates were obtained from the HAL database via the English Lexicon Project (Balota et al., 2007).[8] Coding the remaining four constraints requires real-world knowledge, and so they were coded twice, independently, by the first author and a trained research assistant. Conflicting judgments were resolved through discussion; with discussion, the two coders were always able to reach agreement.

As predicted in Section 1.2.1, our attested items show a significant but not perfect correlation between model-predicted abstract ordering preference and relative frequencies (computed from the Google n-grams corpus; Brants and Franz, 2006): $r(40) = 0.59; p < 0.0001$. This relationship is visualized in Figure 1.

For our novel binomials, we chose expressions that our model predicts to have strong ordering preferences, with values less than 0.3 or greater than 0.7. As much as possible, we chose expressions that minimized the correlations between constraints (e.g. to dissociate length and frequency). A comparison of the profiles of constraint activity for novel and attested items is given in Appendix B.

[7] Levy et al. (2012) estimate that college-age English speakers have been exposed to no more than 350 million words of English in their lifetimes. To be included in the Google books corpus, an n-gram must have appeared at least 40 times in their 468,491,999,592 word corpus. Thus our binomials can have appeared at most 39 times in this corpus, and there is at most a roughly 3% chance that a college-age speaker would have heard any given one of these expressions. Although our participants are on average slightly older than college-age, we believe there is still an exceedingly small chance that they will have substantial experience with any of these expressions.

[8] On three occasions, one word in a pair was not in the English Lexicon Project database (*groundskeeper, ninety-eighth*, and *wildfires*). In these cases, the non-included word was assumed to be the less frequent.

For all items, both novel and attested, we constructed a sentence context for the binomial expression, e.g.:

(3) There were many bishops and seamstresses in the small town where I grew up.

(4) There were many seamstresses and bishops in the small town where I grew up.

Sentence structure was unrestricted, but the binomial expression was never in the first two or the last four words of the sentence. Sentences were designed not to introduce pragmatic constraints on binomial ordering: in particular, neither binomial element (nor any word related to exclusively or primarily to only one of the elements) was mentioned in the sentence before the binomial occurred.

With these materials, we carried out two behavioral experiments, a forced-choice preference experiment and a self-paced reading experiment.

4. Experiment 1: Forced-choice preference

4.1. Method

4.1.1. Participants

75 native English speakers (mean age=36 years; sd=14) participated. Participants were recruited through Amazon Mechanical Turk, restricted to people connecting to the website from within the United States, and were paid 50 cents. Participants were asked to report their "Native language (what you learned to speak with your mother as a child)". Those who did not report English among their native languages were excluded.

4.1.2. Procedure

The Amazon Mechanical Turk instructions directed participants to an external website, where our experiment was presented using WebExp (Keller et al., 2009). Participants first filled out a demographic questionnaire, then continued to the main experiment. On each trial, participants saw one item embedded in sentence context, in both possible orders, e.g.:

- There were many bishops and seamstresses in the small town where I grew up.
- There were many seamstresses and bishops in the small town where I grew up.

Participants were asked to choose which order "sounds more natural". Each participant saw all 84 items. Which expression order was listed first was counterbalanced across participants. Order of item presentation was randomized separately for each participant. The experiment typically took 10-15 minutes.

4.2. Results

Before proceeding with our main multiple regression analysis of the effects of abstract knowledge and direct experience on ordering preference, we present a striking overall difference between the distributions of preference strengths for attested versus novel binomials. Figure 2 shows that ordering preferences are more polarized for attested than for novel binomials (despite the fact that we selected our novel binomials to have extreme preferences); in other words, preferences are more consistent across subjects for the attested expressions. We define a measure of extremity for each item as the difference between its experimentally determined preference strength (i.e. proportion of times preferred in alphabetical order) and 0.5. In a t-test, the attested items are significantly more extreme than the novel ($t = 8.31, p < 0.001$). We discuss this issue further in Sections 4.3 and 6.3.

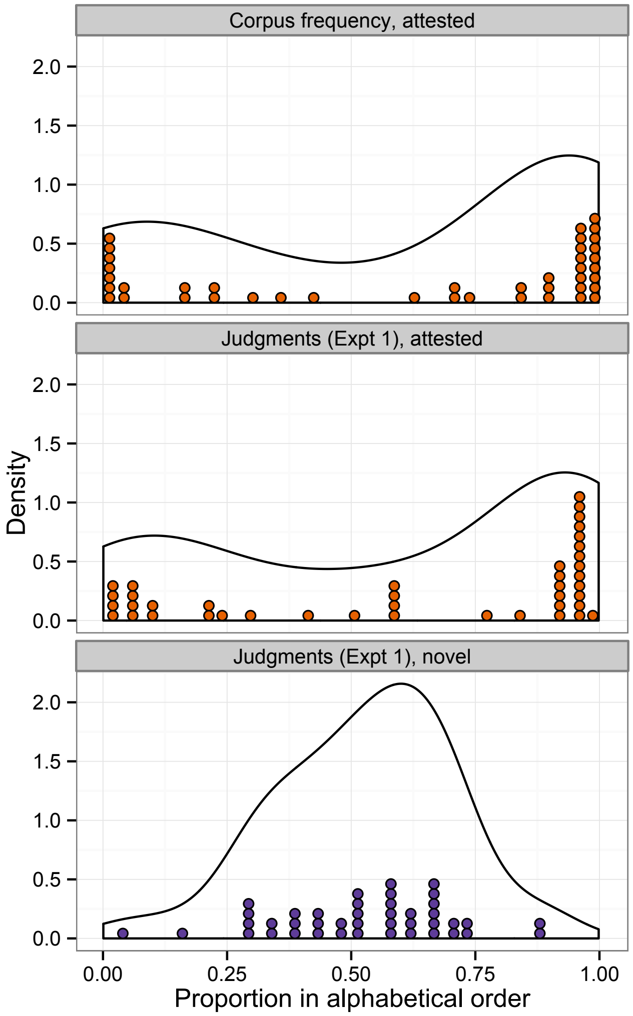

Figure 2: Results of Experiment 1: Proportion of binomials occurring in alphabetical order in Google n-grams corpus frequency (top) and subjects' forced-choice preference judgments (middle/bottom). Dots show individual binomial types, while lines show density estimates. In judgments, attested binomials have more extreme preferences (i.e. more consistent across subjects) than novel binomials, demonstrating a qualitatively similar distribution to corpus frequencies.

4.2.1. Multiple regression analysis

Next we analyze our data using mixed-effects logistic regression (Jaeger, 2008). Our dependent variable is the preferred order, coded as alphabetical or non-alphabetical: alphabetical order is used as a neutral order because results of our initial model selection—see Section 2—indicate that alphabetical order is not a significant predictor of ordering preference. Our independent (fixed-effect) predictors are:

- **Type** (attested/novel) is treatment coded with "attested" as the reference level, i.e. the `Intercept` value applies to attested items, and this value is adjusted by the `Type:novel` value for novel binomials. We predict no significant intercept (i.e. attested binomials are not significantly more likely to be preferred in alphabetical or non-alphabetical order, absent other factors), and no significant effect of type (i.e. novel binomials are not significantly more or less likely to be preferred in alphabetical order

	Estimate	*Std. Error*	*z value*	*p value*
Intercept	-0.14	0.15	-0.98	0.33
Type: novel	0.25	0.19	1.32	0.19
Abs know (Type: attested)	2.32	0.56	4.12	.00004***
Abs know (Type: novel)	1.45	0.35	4.11	.00004***
Rel freq	6.18	0.49	12.55	$<2\text{x}10^{-16***}$

Table 2: Model fit for results of Experiment 1. All VIF < 1.2.

than attested binomials).

- **Abstract knowledge** is operationalized as our model's predicted probability (between 0 and 1) of the expression occurring in alphabetical order. We center this predictor around 0.5. We nest the abstract knowledge predictor within type, i.e. we fit separate parameters for the effect of abstract knowledge for novel and attested binomials, allowing us to consider the effects of abstract knowledge on each type independently. For each type, if abstract ordering constraints are active in influencing offline judgments, then we predict a significant effect of abstract knowledge.

- **Relative frequency** estimates are computed for attested binomials using the Google n-grams corpus (Brants and Franz, 2006) as the frequency of *"A and B"* divided by the frequency of *"A and B"* plus *"B and A"* (resulting in a value between 0 and 1), and centered around 0.5. Relative frequency for all novel binomials is set to 0 after centering. (Thus no interaction of relative frequency with type is necessary.) If direct experience with attested expressions influences offline judgments, then we predict a significant effect of relative frequency.

Following Barr et al. (2013), we use the maximal random effects structures for subjects and items justified by the experimental design: by-subject and by-item intercepts, and by-subject slopes for type, abstract knowledge, their interaction, and frequency.

Model results are given in Table 2.[9] Significance levels for effects are reported using the Wald z statistic and are confirmed using likelihood ratio tests. We see a significant effect of abstract knowledge for both novel and attested expressions, demonstrating that abstract ordering constraints are active in determining forced-choice preferences for both binomial types. In a likelihood ratio test comparing this model to a model with only an additive (non-nested) fixed effect of abstract knowledge, we find no significant difference ($\chi^2(1) = 1.63, p = 0.20$); in other words, the effect of abstract knowledge does not differ significantly between novel and attested expressions. The effect of abstract knowledge for novel binomials is displayed in Figure 3.

We also see a significant effect of relative frequency, demonstrating that direct experience also plays a role in determining preferences for attested expressions. We note that relative frequency is a stronger predictor than abstract knowledge, measured in terms of larger regression coefficient estimate, larger z value, and larger change in likelihood when removed from the model. The strong predictive power of relative frequency is displayed in Figure 4.

[9] The model presented here includes all the fixed-effect predictors and interactions that are of crucial theoretical interest for the hypotheses we set out to test. In order to explore possible further interactions between predictors, as well as possible changes in behavior over the course of the experiment, we fit a mixed-effects logistic regression including as predictors all the previous predictors, a trial order predictor, and all two-way interactions, using the `MCMCglmm` package in `R` (Hadfield, 2010). (The trial order predictor was not included in the original model presented here because a main effect of trial order is implausible, as it would indicate a changing probability of prefering binomials in alphabetical order over the course of the experiment. However, its interaction with other predictors—in particular, abstract knowledge and relative frequency—is potentially of interest.) No further interactions (beyond the type x abstract knowledge interaction included in the original model) reached significance.

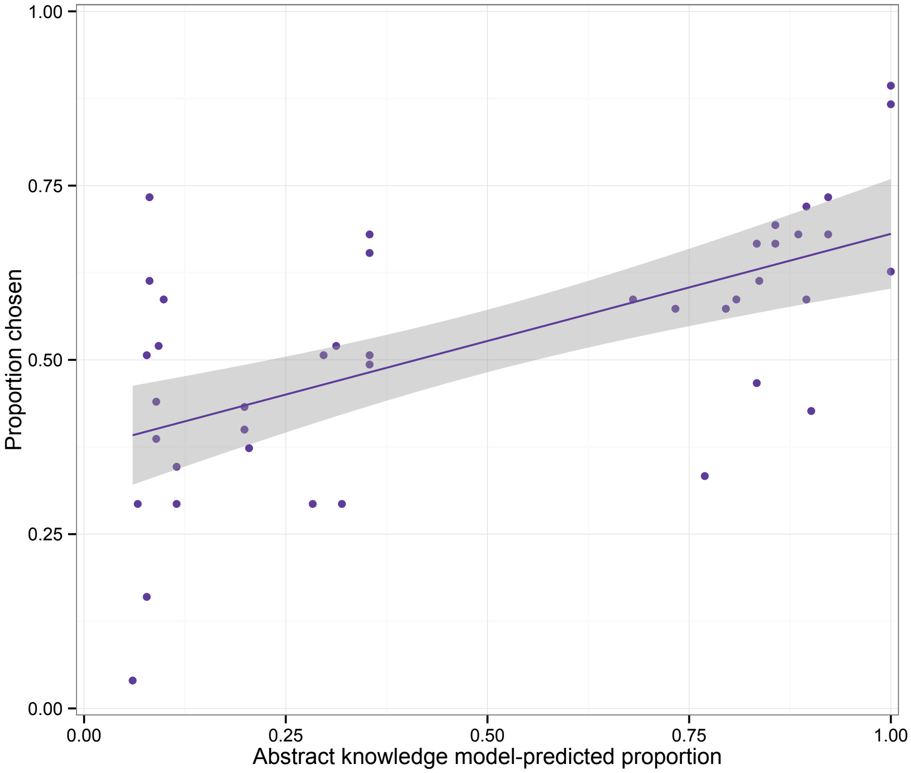

Figure 3: Results of Experiment 1 (novel items): Ordering preferences for novel binomials by model-predicted abstract knowledge. Each point represents an item. x values are the abstract knowledge model's prediction for how often the item will appear in alphabetical order. y values are how often the item was preferred in that order. Line shows best linear fit on the by-items aggregated data. Abstract knowledge is a significant predictor of preferences for novel expressions.

4.3. Discussion

In this experiment, we set out to test whether abstract knowledge and direct experience (specifically, relative frequency) predict ordering preferences in a forced-choice preference task for both novel and frequently attested binomial expressions. We demonstrate that preferences of both attested and novel expressions are affected by abstract knowledge and that preferences of attested expressions are also strongly predicted by relative frequency. This pattern of results supports a theory wherein both abstract knowledge and direct experience play a role in processing. Moreover, for attested expressions, we find that relatively frequency is a stronger predictor of preferences than abstract knowledge, suggesting that processing of these expressions relies more heavily upon direct experience than upon abstract knowledge.

Although the effect of abstract knowledge does not differ significantly across binomial types, we do not think it is justified to draw strong theoretical conclusions from this null result. As we will see in Section 5.2.2, abstract knowledge does interact significantly with binomial type in Experiment 2. We defer further discussion of this issue until Section 5.3.

We additionally find that forced-choice preferences are more extreme for attested than for novel expressions; that is, attested expressions are more consistently preferred in one direction than novel expressions. Taken at face value, this finding suggests that increased overall frequency of an expression exaggerates or solidifies people's preferences. Another possibility, however, is that preferences for novel expressions are underlying equally as extreme as those of the attested expressions, but that the forced-choice judgement process

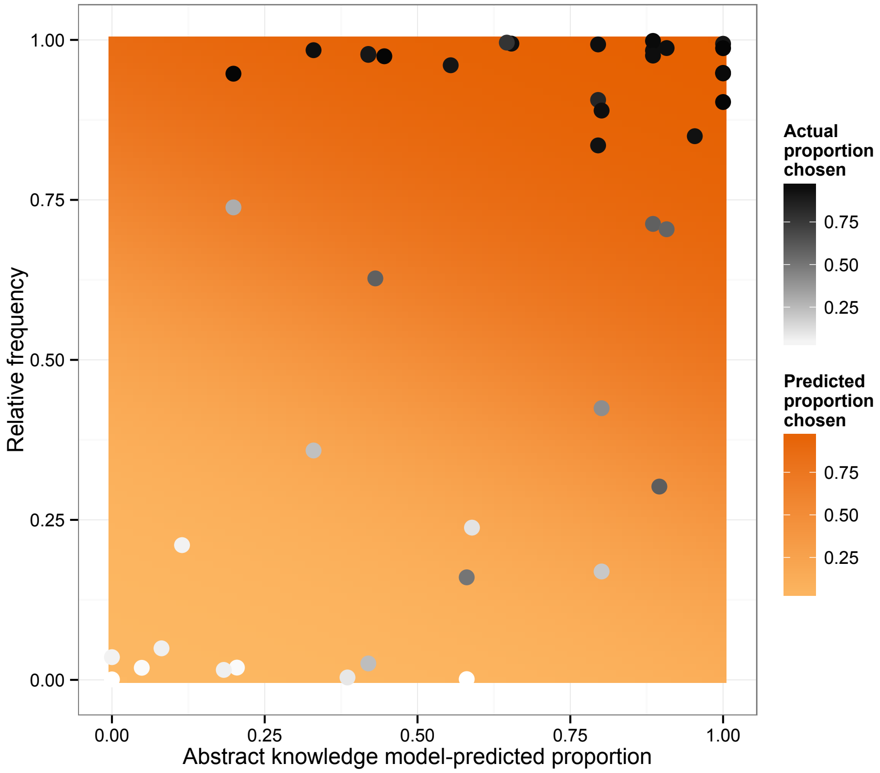

Figure 4: Results of Experiment 1 (attested items), visualized as colors overlaid on Figure 1. Each point represents an item. x values are the abstract knowledge model's prediction for how often the item will appear in alphabetical order. y values are the item's relative frequency of appearing in that order. Points' shading (white to black) shows often the item was preferred in that order. Background shading (light to dark orange) shows the best-fit model (Table 2) prediction for how often the item was preferred in that order. Both relative frequency and abstract knowledge predict true preferences, as depicted by the diagonal background gradient but relative frequency is the stronger predictor, as depicted by the stronger vertical than horizontal gradient.

for these items is noisier,[10] making the resulting preferences for novel expressions appear less extreme than they truly are. We will return to this question in the general discussion.

One potential confound mentioned earlier is the role of local sentence context on binomial order preferences. Although we tried to avoid biasing contexts in designing our materials, it is always possible that some bias unintentially slipped through. However, even if such bias does exist within individual sentences—i.e. the sentence context favors one order more than another, relative to the binomials' intrinsic ordering preference in a hypothetical neutral context—it would not confound the results presented here. Specifically, because our dependent variable is an alphabetical versus non-alphabetical preference, in order to bias our results the local context biases would need to be systematically correlated with the alphabetical/non-alphabetical preferences as given by our predictors of interest (abstract knowledge and relative frequency). Since we have no reason to expect this to be the case, any unintentional effects of local context will merely add noise to our estimates of ordering preferences.

[10] There are many reasons why this could be the case. For instance, when judging attested items, participants may believe that there is a "right" answer and take care to give that answer, whereas when judging novel items, they may put in less effort.

	novel	attested
preferred	0.97	0.97
non-preferred	0.97	0.96

Table 3: Comprehension question accuracy for Experiment 2. Novel and attested items are divided into preferred/non-preferred order according to abstract knowledge model predictions.

In the next experiment, we ask whether the patterns found in our forced-choice preference experiment likewise hold in an online reading experiment.

5. Experiment 2: Self-paced reading

5.1. Method

5.1.1. Participants

400 native English speakers (mean age=34 years; sd=12) participated. Experiment 2 required substantially more participants than Experiment 1 because the self-paced reading data are noisier than the forced-choice data and because, as described in Section 5.1.2, each subject saw approximately half the items in Experiment 2, compared to all the items in Experiment 1. Participant recruitment was identical to Experiment 1, except that participants were paid $1.00.

5.1.2. Procedure

The experiment was presented within Amazon Mechanical Turk using flexspr (Tily, 2012; previously used by Bergen et al., 2012; Linzen and Jaeger, 2015; Singh et al., 2015). Using this method online allows for collection of more data than would be possible in a laboratory setting, and previous work has replicated multiple in-the-lab results with web-based self-paced reading (Enochson and Culbertson, 2015). Participants first filled out a demographic questionnaire, then read sentences in a self-paced reading paradigm: sentences were presented one word at a time, and participants pressed a button to advance to the next word. Reading times for each word were recorded. Participants read three practice sentences, then continued to the main experiment.

Our materials consisted of the same 84 binomial expressions in sentence context as used in Experiment 1, plus 84 unrelated filler sentences. Two stimulus lists were constructed with items rotated and counterbalanced between lists so that each participant only saw a given binomial in one of its two possible orders. Due to a programming error, out of the 168 items in each list, each participant saw a random selection of 80 items. Order of presentation was randomized separately for each participant.

Presentation of each sentence was followed by a yes/no comprehension question. Answers did not depend on the order of the binomial expression. The experiment typically took about 30 minutes.

5.2. Results

5.2.1. Comprehension question accuracy

Comprehension question accuracy is extremely high across all conditions. See Table 3.

5.2.2. Multiple regression analysis

We use regression analysis to compare abstract knowledge and relative frequency as predictors of reading times, analagous to our analysis in Experiment 1.

We divide our experimental items into regions of analysis as shown below:

Prelim	*Word1*	*And*	*Word2*	*Spill1*	*Spill2*	*Spill3*	
There were many	bishops seamstresses	and	seamstresses bishops	in	the	small	. . .

The *Prelim* region encompasses the entire beginning of the sentence up to the binomial expression; all further regions are a single word. We analyze reading time data for each trial summed over a six-word region spanning from Word1 through Spill3. By summing across reading times for these regions, we take advantage of the controlled properties of our stimuli: regardless of order of binomial presentation across conditions, participants will have read the same group of words within the region being analyzed. (For more direct comparison with the previous literature, we present word-by-word analyses of reading times in Appendix C.)

Specifically, we computed a summed reading time measure for each trial as follows: we excluded all trials in which the reading time for any word was less than 100ms or greater than 5000ms. To account for influences of word length, as described by Ferreira and Clifton (1986), we then computed subject-specific residualized reading times (regressed against word length) for each word from the Word1 through Spill3 regions, using data from all non-sentence-final words in non-practice trials.[11] Summing the residuals for this six-word region gives us a residual reading time for each trial. We performed outlier removal without regard to item type or condition: we computed a grand mean and standard deviation and exclude trials with summed times more than 2.5 standard deviations above or below the mean, resulting in a loss of 1.7% of data.

We analyze the data using a mixed-effects linear regression similar to that used in Experiment 1. Our dependent variable is summed residual reading time. Our independent (fixed-effect) predictors and their interpretations are identical to those used in Experiment 1 (Section 4.2.1) with one addition:

- **Trial order** is the position in the experiment in which the given trial occurred. As is common in reading experiments (e.g. Hofmeister et al., 2011; Fine et al., 2013 and many others), we expect that subjects will read faster later in the experiment due to practice effects.

In addition to our hypotheses regarding possible influences of abstract knowledge and direct experience on reading times (which are the same as in Experiment 1), we additionally anticipate a possible statistically significant but theoretically uninteresting main effect of binomial type because the two types contain different words in different sentence frames, and thus one type may be read faster than the other. Following Barr et al. (2013), we use the maximal random effects structure for subjects as justified by the experimental design, namely an intercept and slopes for type, abstract knowledge, their interaction, and relative frequency. We also include a by-subjects random slope for trial order. For items, defined as unordered word pairs, we include a random intercept, a random slope for trial order, and (in place of random slopes for both abstract knowledge and relative frequency) a random slope for a binary alphabetical/non-alphabetical factor, thus allowing for arbitrary item-specific ordering preferences.

Model results are given in Table 4.[12] Effects with $t \geq 2$ are taken to be significant. Positive coefficients indicate slower reading. We see a significant main effect of type with novel expressions read slower, which we attribute to these expressions containing less frequent words on average, in addition to being less frequent expressions overall.

We do not find a significant effect of abstract knowledge for attested expressions, suggesting that abstract ordering constraints are not active in the online processing of these expressions. However, we do find a significant effect of abstract knowledge for novel expressions. In a likelihood ratio test comparing this model to a model with only an additive (non-nested) effect of abstract knowledge, we find a significant difference ($\chi^2(1) = 4.24, p < 0.04$); in other words, the effect of abstract knowledge differs significantly between novel and attested expressions, playing a significant role in online processing for novel expressions only. We additionally find a significant effect of relative frequency, demonstrating that higher relative frequency leads to faster reading in the online processing of attested expressions.

Finally, we find a significant effect of trial order, with faster reading later in the experiment. Results are visualized in Figures 5 and 6.

[11] For analyses using raw reading times, see Appendix D.

[12] The model presented here includes all the fixed effect predictors and interactions that are of crucial theoretical interest for the hypotheses we set out to test. In order to explore possible further interactions between predictors, we fit a mixed-effects linear regression including as predictors all these fixed-effect predictors and all two-way interactions using the `MCMCglmm` package in `R` (Hadfield, 2010). No further interactions (beyond the type x abstract knowledge interaction included in the original model) reached significance.

	Estimate	*Std. Error*	*t value*
Intercept	196.34	26.04	7.54
Type: novel	195.17	25.77	7.57
Abs know (Type: attested)	13.88	23.14	0.60
Abs know (Type: novel)	-48.73	18.02	-2.70
Rel freq	-59.25	18.42	-3.22
Trial order	-8.35	0.39	-21.24

Table 4: Model fit for results of Experiment 2. Effects with $t > 2$ are taken to be significant. All VIF < 1.7.

5.3. Discussion

We demonstrate for the first time that novel binomial expressions show online effects of abstract ordering preferences. In contrast, reading times for frequently attested binomial expressions are only influenced by relative frequency. These findings imply a trade-off in online processing between reliance on abstract knowledge and direct experience, where novel expressions must be processed on the basis of abstract knowledge only, but highly frequent attested expressions can be processed primarily with reference to previous direct experience.

Here we found a significant interaction of abstract knowledge with binomial type, such that abstract knowledge was significantly less active in determining reading times for attested binomials than for novel binomials. In contrast, in Experiment 1, we found no such significant interaction. What is consistent across these two experiments is that processing of attested expressions is more strongly influenced by direct experience than by abstract knowledge. However, given the inconsistent results concerning the interaction of abstract knowledge and binomial type, we cannot state with confidence whether abstract knowledge is differentially active between novel and attested binomials.

6. General discussion

We set out to investigate the roles of abstract knowledge and direct experience in the processing of binomial expressions, asking whether binomial ordering preferences are driven by constraints on the semantic, phonological, and lexical properties of words in an expression, or by prior experience with the specific expression in question. Our key findings are as follows. First, we demonstrated that abstract ordering constraints are active in the comprehension of novel expressions in both an offline forced-choice task and a online self-paced reading task. Second, we demonstrated that for frequently attested expressions, effects of direct experience largely overwhelm abstract knowledge in predicting behavioral data, both in the offline task and especially in the online task.

Our results support exemplar- or usage-based theories of language, which allow for the storage and reuse of multi-word expressions. Specifically, our finding that ordering preferences for attested binomial expressions are primarily driven by relative frequency is evidence that the processing of these expressions makes use of holistic multi-word mental representations. In contrast, a traditional words-and-rules theory would predict that these expressions are generated compositionally each time they are encountered, and that the ordering preferences of attested expressions, like those of novel expressions, should stem from abstract ordering constraints rather than relative frequency of direct experience.

Of the predictions made in Section 1.2.3, our results indicate that both abstract knowledge and relative frequency play a role in the processing of binomial expressions. Many patterns are possible for the manner in which these two knowledge sources trade off as a function of the overall frequency of an expression: In one extreme, abstract knowledge could apply only for expressions that have never before been encountered, with relative frequency taking over as soon as any direct experience exists. In the other extreme, abstract knowledge could apply in the vast majority of cases, with relative frequency limited to playing a role only for the highest frequency items, such as those used in our experiments. A middle ground position proposes a gradual switch from reliance on abstract knowledge to reliance on relative frequency as overall frequency increases.

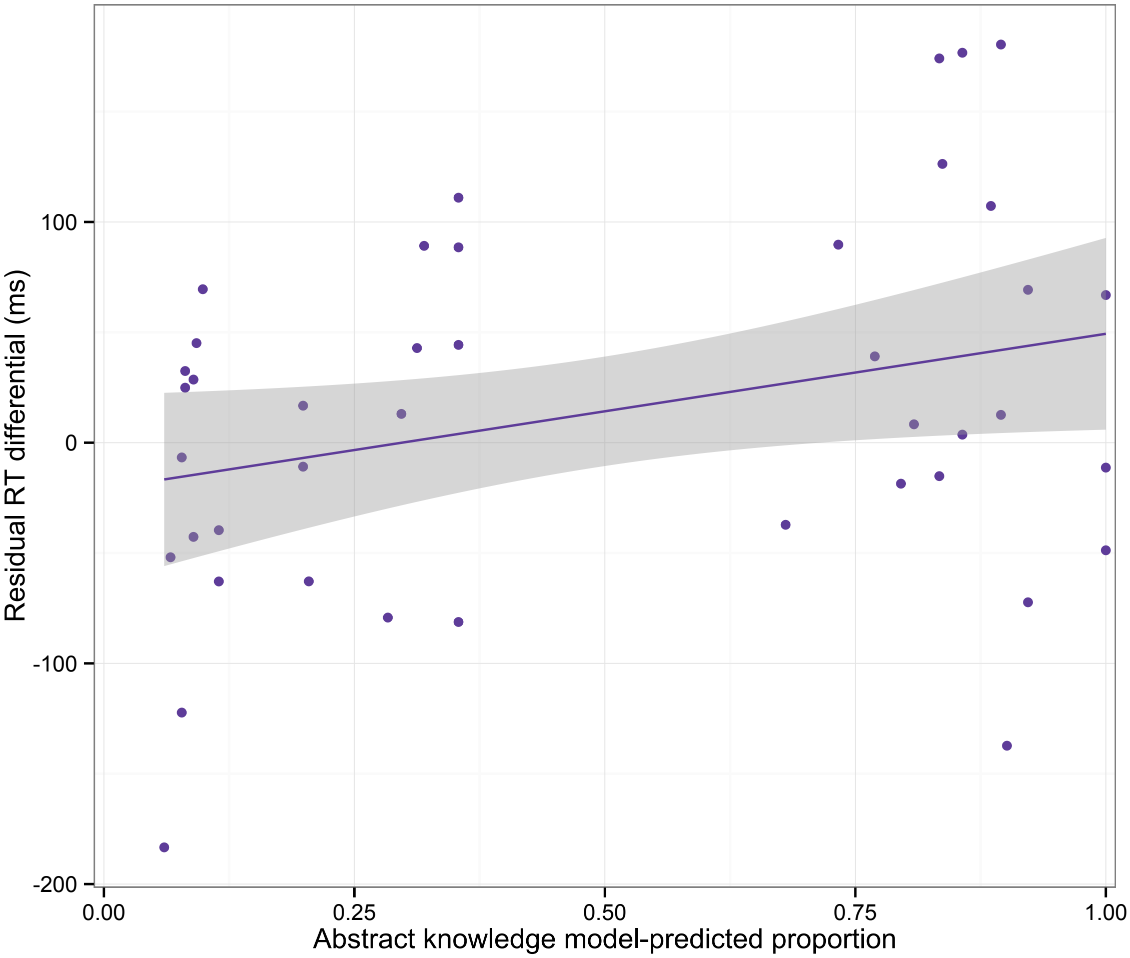

Figure 5: Results of Experiment 2 (novel items): Reading time differentials for novel binomials by model-predicted abstract knowledge. Each point represents an item. x values are abstract knowledge model's predictions for how often the item will appear in alphabetical order. y values are the differences between average summed residual reading times for the non-alphabetical and alphabetical orders. Line shows best linear fit on the by-items aggregated data. Abstract knowledge is a significant predictor of reading times.

We propose that both extremes are unlikely and that the middle position of a gradual trade-off is the most likely. The first extreme is counterintuitive, since a single encounter with an expression seems insufficient to thoroughly trump abstract knowledge. The second extreme has been argued against by Arnon and Snider (2010), who found frequency effects for multi-word expressions across a wide range of frequencies. Their finding of frequency effects for low-to-medium frequency items would not be predicted by a theory in which direct experience applies only to the processing of extremely high frequency items. The gradual trade-off theory, on the other hand, is supported by a wide variety of computational models.

6.1. Convergent evidence from computational models

6.1.1. Connectionist models

A similar trade-off has been demonstrated in connectionist models of language learning in domains such as past-tense formation (Rumelhart and McClelland, 1986) and grammatical structure (Elman, 2003), which

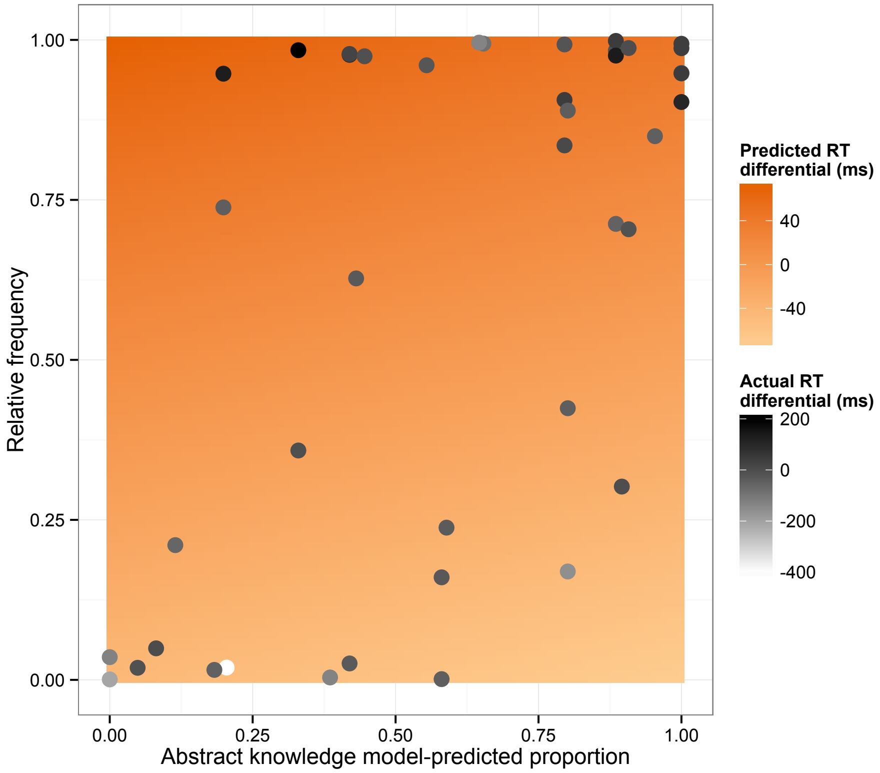

Figure 6: Results of Experiment 2 (attested items), visualized as colors overlaid on Figure 1. Each point represents an item. x values are the abstract knowledge model's prediction for how often the item will appear in alphabetical order. y values are the item's relative frequency of appearing in that order. Points' shading (white to black) shows the item's true average RT differential. Background shading (light to dark orange) shows the best-fit model (Table 4) prediction for RT differential. Only relative frequency is a significant predictor of reading times, as depicted by the strong vertical background gradient.

learn both generalized patterns and specific exceptions. These models learn to predict patterns within their training data (e.g. Form the past tense by adding *-ed*). When new items are introduced, they are at first treated accorded to the general patterns, but with further training, the model can learn to treat certain items as exceptions.

These models have primarily been conceived as models of early language acquisition and tested on frequent items (e.g. common verbs), where it can assumed that by adulthood, most native speakers will have extensive experience with all the items in questions, and will thus consistently recognize certain words as exceptions to the general rules. However, their behavior on new items straightforwardly generalizes to low frequency items that even adult native speakers would have relatively little direct experience with, such as attested but low frequency binomial expressions, making the prediction that these items could occupy a middle ground of partial reliance upon both general patterns (i.e. abstract knowledge) and direct experience, even in a fully developed adult grammar.

6.1.2. Exemplar-based computational models

A gradual trade-off is also predicted by a particular class of exemplar-based computational models of language: namely, those that incorporate representations of fragments varying in size, and that allow not only for holistic reuse of the largest fragments but also for rule-based composition of smaller fragments (e.g. Bod, 1998; Bod et al., 2003; Bod, 2008; Johnson et al., 2007; Demberg, 2010; O'Donnell et al., 2011). Within these models, multi-word expressions can thus be parsed both through direct reuse and through compositional generation. The probabilities assigned to these units—the holistic expressions, the individual words, and the compositional rules—will collectively determine the relative likelihoods of reuse versus regeneration. For more frequent expressions, the probability of reusing a holistic unit will be higher, while for less frequent expressions, the probability of compositional generation will be higher. These probabilities change gradually depending on the frequency of a given expression as well as the frequencies of similar expressions. These models thus also predict a gradual trade-off between reliance on abstract knowledge for infrequent items and reliance upon direct experience for frequent items.

6.1.3. Nonparametric Bayesian models

The gradual trade-off theory is also supported by a nonparametric Bayesian perspective (e.g. Goldwater et al., 2009; Xu and Tenenbaum, 2007), in which expectations are influenced by both a prior probability and the incoming data. In a Bayesian model, when little data has been seen, expectations are driven by the prior probability. As more data is seen, the data becomes increasingly influential, asymptotically approaching complete dominance. For binomial expressions, abstract knowledge can be thought of as a prior probability for ordering preferences, absent any direct experience with a given expression, and each direct encounter with an expression constitutes further data. Under the Bayesian perspective, when one has little experience with an expression, expectations will be governed by abstract knowledge, but with increasing experience, the relative frequency of ordering within the experienced data will be increasingly dominant in determining expectations.

6.2. Advantages of our approach

While numerous models support our conclusions, the experiments presented here crucially advance the state of our understanding beyond what was previously known by providing a novel approach for using *behavioral evidence*, in conjunction with modern corpora and statistical techniques, to quantify the contributions of abstract knowledge and direct experience. Our probabilistic model provides quantitative estimates for the effects of abstract knowledge, while corpus frequencies provide estimates for direct experience. Using multiple regression modeling, we can directly compare the predictive strength of these two influences on behavioral data such as the results of our forced-choice and self-paced reading tasks. This approach allows us to move beyond the previous modeling-based approaches, which focused on predicting corpus data or language-wide trends. We can now investigate the trade-off between abstract knowledge and direct experience using behavioral evidence.

Additionally, the statistical techniques employed here allow us to make quantitative claims about the strength of reliance on both abstract knowledge and direct knowledge. We have seen this in a limited way so far, as we demonstrated that processing of frequently attested binomials is driven primarily by relative frequency, and only to a lesser degree by abstract knowledge. We have also predicted that there should be a gradual shift from reliance upon abstract knowledge to reliance upon relative frequency estimates as overall frequency increases; however, we cannot conclude this directly from our current data because overall frequency has only been explored as a dichotomous variable: either entirely novel or very frequent. In future work, we plan to look at an in-between zone of attested but not highly frequent expressions, e.g. *sunglasses and sunscreen/sunscreen and sunglasses* (1/1000th the frequency of the average attested expression in the current study). We predict that these expressions should show noticeable effects of both abstract constraints and relative frequency. Moreover, looking over a range of overall frequencies, we predict that we will see a quantitative trade-off between reliance on abstract knowledge and reliance on direct experience.

This approach to studying the trade-off between abstract knowledge and direct experience generalizes beyond the study of binomial expression ordering preferences. The cornerstone of this approach is that we are able to independently quantify the contributions of direct experience with specific expressions and abstract knowledge in the absence of direct experience. We propose that a combination of corpus frequencies and probabilistic modeling can provide such estimates for a wide range of linguistic constructions (e.g. the dative alternation [Bresnan et al., 2007] and adjective ordering [Dixon, 1982; Truswell, 2009]) allowing us to ask broad questions about the trade-off between compositional generation and the reuse of stored expressions in linguistic processing. For example, to what extent are adjective ordering preferences due to abstract rules (e.g. shape before color) versus to known collocations of highly frequent adjective sequences? The methods we have developed here make these questions accessible for future research.

6.3. Further predictions about language structure

Our results additionally lead to predictions about language structure. Our gradual trade-off theory predicts that items with higher overall frequency will be more likely to have relative frequency preferences that contradict abstract knowledge preferences. This prediction is analogous to the finding that more frequent verbs are more likely to be irregular (Bybee, 1985; Lieberman et al., 2007): in the case of high overall-frequency items, people have enough exposure to learn idiosyncratic or abstract-knowledge-violating preferences, but in the case of low overall-frequency items, people have insufficient exposure to overcome their abstract knowledge. A further prediction follows from the results of Experiment 1, in which we found that preferences for attested items were more extreme, or polarized, than preferences for novel items. Assuming that preferences for attested items are driven primarily by relative frequency, this result predicts that as overall frequency increases, relative frequencies will become more polarized.[13]

In related work, we found that these predictions were borne out in a corpus analysis (Morgan and Levy, 2015), which demonstrated that binomial expressions with higher overall frequency have relative frequencies that deviate more from abstract knowledge—in particular, by being more polarized. This finding in turn leads to further questions about the historical trajectories of binomial expression ordering preferences, and the dual roles of individuals' language processing and cultural transmission in shaping language structure (Kirby et al., 2007; Morgan and Levy, 2016). Thus the results presented here additionally open the door to further investigation of the mutually constraining processes of synchronic language processing and diachronic language change.

Acknowledgements

This work has benefited from presentation and discussion at a variety of venues, including the 26th annual CUNY Sentence Processing Conference and the 2014 annual meeting of the Linguistic Society of America. Thank you to Inbal Arnon, Grant Goodall, and an anonymous reviewer for helpful discussion, to Mike Brooks for assistance with preparation of experimental materials and data coding, and to Harry J. Tily for providing us with flexspr. The research presented here was funded by research grants NSF 0953870 and NICHD R01HD065829 and fellowships from the Alfred P. Sloan Foundation and the Center for Advanced Study in the Behavioral Sciences to Roger Levy, and NIH T32-DC000041 to the UCSD Center for Research in Language.

Appendix A. Experimental materials

Comprehension questions are used only in Experiment 2.

Appendix A.1. Novel expressions

1. He was **abashed** and **sorry** about his horrible behavior.
    - Did he defend his behavior?
2. This bar is popular among the **actresses** and **lumberjacks** who live in the neighborhood.
    - Do the lumberjacks hate the bar?
3. Because Jim was **allergic** and **unaccustomed** to elderberries, he was careful to avoid them.

[13] We did not see the analog of this result in Experiment 2: reading time were not slower in the dispreferred order and faster in the preferred order for attested than for novel expressions. Based on the results of Morgan and Levy (2015), we conclude that this is due either to noise or to floor/ceiling effects on reading times.

   - Did Jim like to eat elderberries?

4. My cousin's new talking and singing toy is **annoying** and **teal** according to my aunt.
   - Does my cousin have a new toy?
5. The dentist told Sally that **bacteria** and **candy** would rot her teeth.
   - Did the dentist recommend eating candy?
6. The elephants at the zoo were **beautiful** and **stinky** so the children loved them.
   - Were there elephants at the zoo?
7. The engineer specialized in making **bicycles** and **robots** when he worked for the company.
   - Did the engineer specialize in destroying things?
8. There were many **bishops** and **seamstresses** in the small town where I grew up.
   - Did I grow up in a small town?
9. The berries were **bitter** and **purple** when I ate them this morning.
   - Did I eat berries this morning?
10. Seth told me that there are **blankets** and **kittens** in that box over there.
    - Were there blankets in the box?
11. The rangers seemed to act like **campfires** and **wildfires** were the same thing.
    - Did I hear about fires from a policeman?
12. At the wizard school, **chanting** and **enchanting** were very common occurrences.
    - Did the wizards ride broomsticks frequently?
13. When I met many **chauffeurs** and **stewardesses** at a party, I started questioning my job.
    - Did I go to a party?
14. The third grade class saw **cherries** and **llamas** at the state fair.
    - Did the class go to the state fair?
15. There was nothing but **chickens** and **fences** in the field behind the house.
    - Was the field behind the house?
16. His uncles were all **coroners** and **senators** in their day jobs, but they all wanted to get into the movie industry.
    - Did he have uncles?
17. The drink flavored with **currant** and **pomegranate** was delicious according to Kim.
    - Did Kim like the drink?
18. The dictator was **deposed** and **murdered** by his military adviser.
    - Did the dictator survive?
19. I talked with my boss about whether to hire the **determined** and **forgettable** job candidate that we interviewed.
    - Did I discuss something with my boss?
20. The doctor said that **discontent** and **tearfulness** are signs of depression.
    - Did the doctor talk about flu symptoms?
21. Luke always looked so **disheveled** and **dreary** but he was my best friend.
    - Was Luke my best friend?
22. The kind minister **donates** and **provides** a lot of food to the charity.

- Was the minister kind?

23. My favorite animals have been **felines** and **quails** ever since I was a kid.
    - Have I always hated animals?
24. The finalists in the tennis championship were ranked **first** and **ninety-eighth** in the world prior to the tournament.
    - Was there a golf champtionship?
25. In the spring, Julie will plant **flowers** and **zinnias** in her new garden.
    - Does Julie have a garden?
26. The store owner was **fuming** and **mad** when he found out what was stolen.
    - Was something stolen?
27. As a vegetarian, **gelatin** and **lard** are difficult to avoid.
    - Do vegetarians have a hard time?
28. Laura heard that the school's **groundskeeper** and **superintendent** got married over the summer.
    - Did Laura hear about a divorce?
29. His mother didn't hear when when Nate **happily** and **rudely** told his sister to shut up.
    - Did his mother hear what Nate said?
30. As Joe carried a tall stack of boxes, he had to **hesitate** and **readjust** before he could go further.
    - Was the worker carrying barrels?
31. At the zoo we saw **horses** and **loons** in their natural habitats.
    - Did we go to the zoo?
32. I need to grab my **jacket** and **phone** before I leave the house.
    - Do I have everything I need in order to leave?
33. Sarah likes to buy **kale** and **vegetables** at the famer's market.
    - Does Sarah only buy meat?
34. My cousins were all **lankier** and **lanky** but were surprisingly strong.
    - Were my cousins weak?
35. The pet store was full of **litter** and **newts** when Martha visited on Saturday.
    - Did Martha go to the pet store?
36. Peter met a man who was **masculine** and **undignified** at the conference he went to last month.
    - Did Peter go to the conference last year?
37. The pirate was **marooned** and **missing** for nearly five months.
    - Was the pirate stranded for a year?
38. My grandparents were all **nurses** and **patriarchs** when they were alive.
    - Were some of my grandparents teachers?
39. In my dream, I had **puppies** and **tigers** that I kept as pets.
    - Was I dreaming?
40. Jenny was interested in **rats** and **sharks** as a young child.
    - Was Jenny interested in kittens?
41. Maria could use **therapy** and **vacations** to feel less stressed.
    - Is Maria stressed?
42. Irena had trouble with **vocabulary** and **vowels** while she was learning English.
    - Did Irena have trouble with vowels?

Appendix A.2. Attested expressions

1. The clerk asked for Melissa's **address** and **name** in order to complete the form.
   - Did the clerk help Melissa complete the form?
2. Sarah was relieved to find that her friends were **alive** and **well** after the car crash.
   - Were Sarah's friends alright?
3. Most universities have programs in the **arts** and **sciences** in addition to having various professional schools.
   - Do most university have programs about law?
4. Soccer players practice running both **backwards** and **forwards** in order to stay nimble.
   - Do soccer players practice running sideways?
5. Hunter dislikes reading **black** and **white** text off a computer screen so he uses an unusual color scheme.
   - Does Hunter like the standard color scheme?
6. Learning to strengthen your **body** and **mind** is one of main purposes of doing yoga.
   - Does yoga improve your strength?
7. George always brings **bread** and **butter** with him when he goes camping.
   - Does George always bring hot chocolate when he goes camping?
8. John showed me pictures of the **bride** and **groom** both dressed in blue.
   - Did the couple wear green?
9. I always love seeing my **brothers** and **sisters** when I go home for the holidays.
   - Do I enjoy going home?
10. Caleb likes to **buy** and **sell** electronics on eBay as a hobby.
    - Does Caleb work with eBay professionally?
11. I watched the **cat** and **mouse** run frantically around the barn.
    - Was there a dog in the barn?
12. It can be difficult to determine the **cause** and **effect** of weather patterns over the ocean.
    - Are ocean weather patterns hard to predict?
13. Clarissa found the painting of a **child** and **mother** to be very moving.
    - Did Clarissa see a painting?
14. Catherine was not surprised that tensions between **church** and **state** ran high during the election season.
    - Was there tension during the election season?
15. Peter studied the laws concerning **crime** and **punishment** in Ancient Greece and Rome.
    - Did Peter study what happened in Ancient Greece?
16. Jesse felt like he had worked **day** and **night** on the project but he only got a B on it.
    - Did Jesse get an A?
17. The economist became famous for studying the way **demand** and **supply** affect the steel industry.
    - Did the economist study oil companies?
18. Mark finds working on **development** and **research** for the marketing company to be a very satisfying career.
    - Does Mark want to change jobs?

19. Although some **drink** and **food** were provided at the reception, there was not enough to go around.
    - Was there something to eat at the reception?
20. Diane wrote a book about her travels **east** and **west** around the globe for a year.
    - Did Diane write a book about living in Paris?
21. Sometimes it feels like **error** and **trial** is the only way to learn.
    - Do you sometimes need to learn by trying things?
22. Heather invited her **family** and **friends** to her annual holiday party.
    - Does Heather have a holiday party every year?
23. It is important to study both the **fauna** and **flora** in a region in order to fully understand the ecosystem.
    - Can studying plant life tell you everything you need to know about an ecosystem?
24. Many children find eating with a **fork** and **knife** to be a difficult skill to learn.
    - Do some children have trouble with eating utensils?
25. Keith marveled at the **gold** and **silver** decorations on the walls of the palace.
    - Were the walls dull?
26. Excercising regularly is important for your **heart** and **soul** according to my mother.
    - Did I receive advice from my aunt?
27. Michelle was surprised to learn that the **husband** and **wife** were getting a divorce.
    - Was the couple celebrating their anniversary?
28. I could not guess the **intents** and **purposes** of the confusing new regulations.
    - Were the regulations confusing?
29. Everyone bowed as the **king** and **queen** entered the throne room.
    - Did a jester enter the room?
30. Learning to forecast **loss** and **profit** was a topic in Brian's business skills class.
    - Did Brian take a class on business skills?
31. Paul primarily got his news through **magazines** and **newspapers** rather than through television.
    - Does Paul read the news?
32. I like to **match** and **mix** my clothing to create new outfits.
    - Do I like to always wear the same thing?
33. Jen thought that the **men** and **women** in her dance class were all very talented.
    - Did Jen think that some of her classmates were untalented?
34. Blake dislikes seeing all the **pain** and **suffering** in the world when he watches the news.
    - Does Blake enjoy watching the news?
35. The anthropologist studied the way different cultures conceived of **peace** and **war** during the Middle Ages.
    - Did the anthropologist study dinosaurs?
36. By comparing the **past** and **present** we can learn about universal human tendencies.
    - Does history help us understand humanity?
37. Seth follows both **radio** and **television** broadcasts to stay informed about current events.
    - Does Seth like to follow current events?
38. Some children enjoy learning to **read** and **write** but others dislike it.

	Form	Power	Icon	Percept	Length	Freq	Stress
Form	1.00	0.22	0.01	0.01	0.01	−0.34	0.02
Power	0.22	1.00	0.01	0.06	−0.05	0.05	−0.06
Icon	0.01	0.01	1.00	−0.11	0.44	0.26	0.30
Percept	0.01	0.06	−0.11	1.00	−0.14	−0.10	−0.17
Length	0.01	−0.05	0.44	−0.14	1.00	0.30	0.83
Freq	−0.34	0.05	0.26	−0.10	0.30	1.00	0.21
Stress	0.02	−0.06	0.30	−0.17	0.83	0.21	1.00

Table B.5: Correlations of constraint activity for attested binomials.

	Form	Power	Icon	Percept	Length	Freq	Stress
Form	1.00	0.09	0.02	−0.01	0.15	0.44	0.18
Power	0.09	1.00	0.05	−0.02	0.10	0.13	0.11
Icon	0.02	0.05	1.00	0.03	0.13	0.03	0.04
Percept	−0.01	−0.02	0.03	1.00	0.24	0.22	0.11
Length	0.15	0.10	0.13	0.24	1.00	0.10	0.50
Freq	0.44	0.13	0.03	0.22	0.10	1.00	−0.28
Stress	0.18	0.11	0.04	0.11	0.50	−0.28	1.00

Table B.6: Correlations of constraint activity for novel binomials.

- Do some children enjoy reading more than others?

39. Teaching children what is **right** and **wrong** is a difficult task for parents.
    - Is it easy to teach children morals?
40. After the storm, Haley was glad to hear that her grandparents were **safe** and **sound** in their country home.
    - Was there a storm?
41. The broker bought some risky **shares** and **stocks** without knowing it and only discovered it later.
    - Was the broker originally unaware of what he did?
42. Susan disliked the **sour** and **sweet** soup at the fancy restaurant.
    - Was the restaurant fancy?

Appendix B. Constraint activity profiles

Figure B.7 shows the proportion of items for which each constraint is active (recalling that each constraint can be active or inactive for a given expression). As we can see, constraints are active approximately equally often in each group. Tables B.5 and B.6 show correlations between constraints: constraint activity is coded as 1 if it predicts that an expression should occur in alphabetical order and -1 if it predicts that an expression should occur in non-alphabetical order, or 0 for inactive constraints. We see that, for both novel and attested expressions, most constraints are not highly correlated. One noteworthy exception is Length and Final Stress, which are highly correlated because single-syllable words are as short as possible (hence should come first according to Length) and necessarily have final stress (hence should come first according to Final Stress).

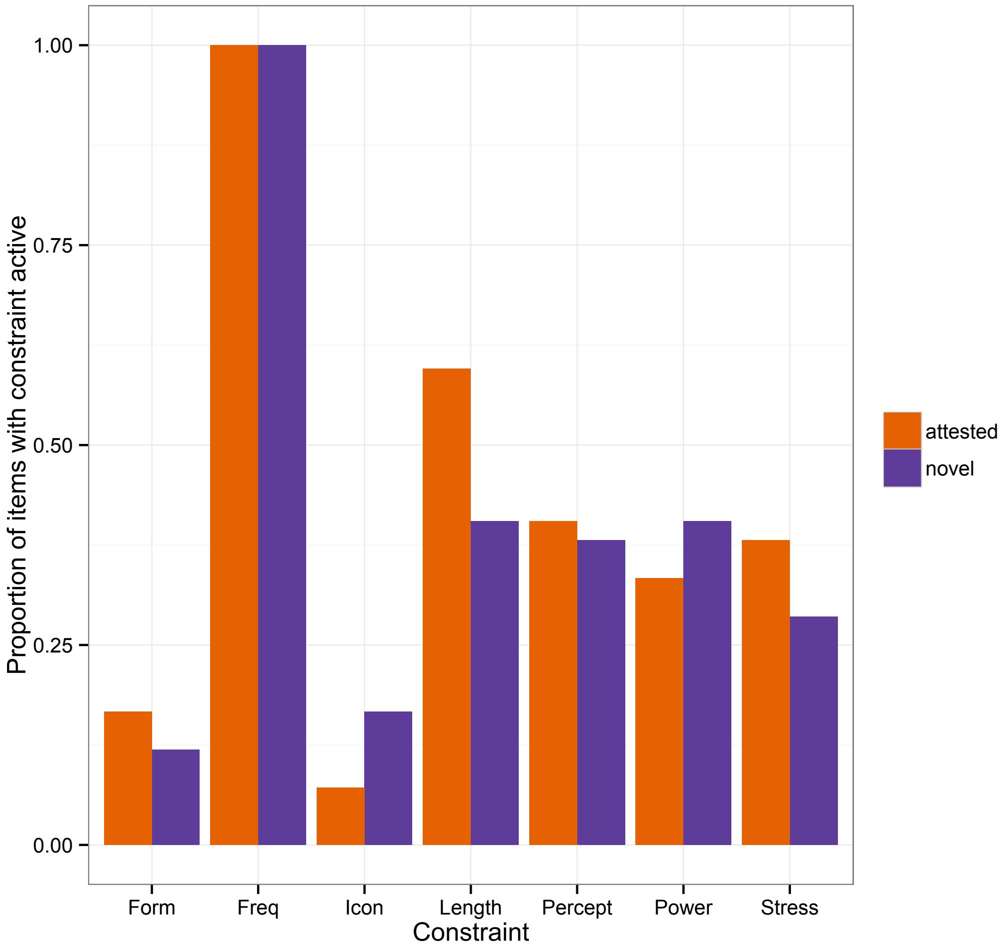

Figure B.7: Proportion of binomial expressions for which each constraint is active.

Appendix C. Experiment 2 region-by-region analyses

Here we present region-by-region analyses of the self-paced reading data from Experiment 2. Our goals in these analyses are to replicate the results of Siyanova-Chanturia et al. (2011) that attested binomial expressions are read faster in their preferred order, and to demonstrate that this finding extends to novel expressions when categorized into preferred/dispreferred orders on the basis of abstract knowledge. Specifically, we analyze reading times by dichotomizing binomials into preferred/dispreferred conditions, rather than using continous abstract knowledge and relative frequency predictors as in Section 5.2.2. For simplicity of presentation, and because we are not concerned here with comparisons across binomial types, we analyze each type (attested/novel) separately.

Residualization on word length and outlier removal are identical to that reported in Section 5.2.2, except that outlier removal was done for each region and each binomial type separately (because each region within each type is analyzed separately in this section).

For each binomial type and region, we fit a linear mixed-effects regression model with residualized reading times (in milliseconds) as the dependent variable. Our independent predictor of interest is a dichotomous preferred/non-preferred variable (treatment coded with “preferred” as the reference level). Details of how preferred order is assessed vary between binomial types and are discussed in more details below. Trial order

Table C.7: Means, standard errors, and t values for the estimated coefficient of the preferred/dispreferred predictor in the region-by-region analyses of the self-paced reading experiment. t values greater than 2 are taken to be significant.

		Prelim	Word1	And	Word2	Spill1	Spill2	Spill3
Novel	Mean (SE)	3.02 (2.14)	11.61 (5.16)	8.21 (6.05)	11.18 (4.79)	-1.74 (4.49)	1.03 (2.41)	2.01 (2.71)
	t value	1.41	2.25	1.36	2.34	-0.39	0.43	0.74
Attested	Mean (SE)	-2.05 (1.87)	-3.55 (3.34)	7.44 (2.59)	15.26 (3.37)	8.82 (2.40)	2.94 (2.75)	2.90 (1.91)
(corpus freq)	t value	-1.10	-1.06	2.87	4.53	3.67	1.07	1.52
Attested	Mean (SE)	-0.12 (1.91)	-3.39 (3.32)	-1.79 (2.78)	5.87 (3.93)	6.12 (2.64)	-2.29 (2.75)	1.84 (2.01)
(model)	t value	-0.06	-1.02	-0.64	1.49	2.32	-0.83	0.92

is also included as a predictor. Following Barr et al. (2013), we use the maximal random effects structure for subjects justified by the experimental design, namely an intercept and a slope for preferred/non-preferred order. We also include a random by-subjects slope for trial order. For items, defined as unordered word pairs, we use an intercept and a slope for a binary alphabetical/non-alphabetical factor (comparable to that used in Section 5.2.2). Results for the predictor of interest are shown in Table C.7.

Novel expressions. For novel expressions, we assign each expression a preferred and non-preferred order on the basis of our abstract knowledge model's prediction for ordering preferences. Results are shown in Figure C.8. As seen in Table C.7, we find significant effects of order at the Word1 and Word2 regions, with preferred read faster than non-preferred.

Attested expressions. For attested expressions, we consider two ways to sort expressions into preferred and non-preferred order: we can use corpus frequencies, replicating Siyanova-Chanturia et al. (2011), or we can use abstract knowledge model predictions for a more direct comparison with the novel expressions. We will show results sorted both ways.

We begin by showing results with preferred/non-preferred determined by corpus frequencies as reported by Siyanova-Chanturia et al.[14] Results are shown in Figure C.9. We find significant effects of order at the And, Word2, and Spill1 regions, with preferred read faster than non-preferred.[15]

Next we analyze our attested expressions as sorted by abstract knowledge model predictions. Results are shown in Figure C.10. We find a significant effect of order at the Spill1 region, with preferred read faster than non-preferred.

Discussion. We replicate Siyanova-Chanturia et al.'s (2011) finding that attested binomial expressions are read faster in their preferred order. We also demonstrate for the first time that novel binomials show online effects of abstract constraints on ordering, with faster reading times in our model's predicted preferred direction.

We do not present a region-by-region version of the multiple regression analyses presented in Section 5.2.2 because we do not expect the results seen there to hold at each region individually. As noted in Section 5.2.2, the analyses presented there took advantage of the fact that within the six-word region analyzed, participants read the same set of words regardless of order of binomial presentation. Within the word-by-word analyses presented here, however, words differ across conditions: Word1 in the preferred condition becomes Word2 in the dispreferred condition, and vice versa (e.g. "bishops and seamstresses" versus "seamstresses and bishops"). Moreover, recall that effects of lexical frequency are one component of abstract knowledge (Section 2), such that binomials in preferred order on average have a more frequent word preceeding a less

[14] Siyanova-Chanturia et al.'s reported preferences differ from the Google n-gram preferences for one item, *family and friends*.

[15] Siyanova-Chanturia et al. only report aggregate reading times, not word by word reading times, so we cannot say whether our results directly replicate exactly where in the sentence these effects appear.

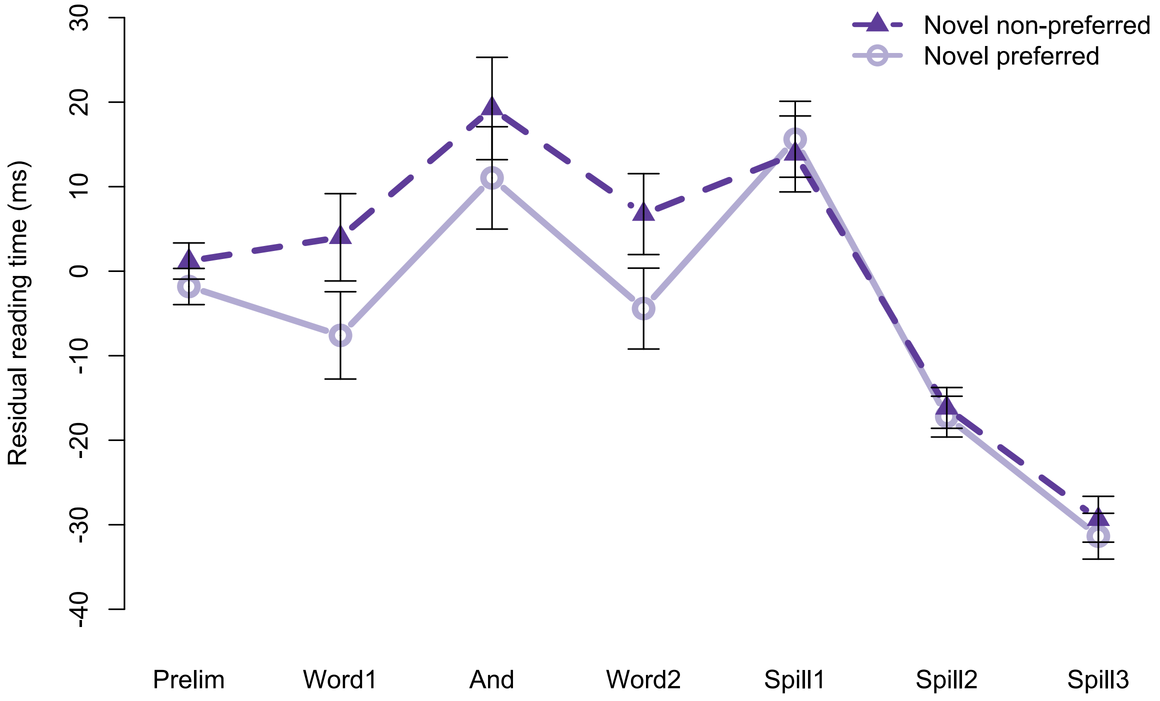

Figure C.8: Self-paced reading times for novel expressions. Error bars show standard errors for the predictor of interest (Table C.7).

frequent word, while binomials in dispreferred order on average have a less frequent word proceeding a more frequent word. Thus, on the basis of lexical frequency alone, we would expect to see the preferred order read faster around Word1 (or shortly thereafter, due to spillover), and the dispreferred order read faster around Word2 (or shortly thereafter). In other words, on the basis of lexical frequency alone, we would expect to see a local reversal of the effect of abstract knowledge around Word2 (although we expect this reversal to be smaller in magnitude than the overall benefit of conforming to abstract knowledge across the binomial as a whole). This prediction is born out numerically in the Spill1 region for novel binomials, although it does not approach significance.

Appendix D. Experiment 2 results with raw reading times

Here we replicate the analyses presented in Section 5.2.2 with raw rather than word-length-residualized reading times. Model results are given in Table D.8. Crucial effects are very similar to those seen in Table 4. In a likelihood ratio test comparing this model to a model with only an additive (non-nested) effect of abstract knowledge, we find a marginally significant difference ($\chi^2(1) = 3.12, p = 0.08$). We attribute the lower significant level here compared to that presented in in Section 5.2.2 to presence of extra noise in the raw compared to the residualized reading time data.

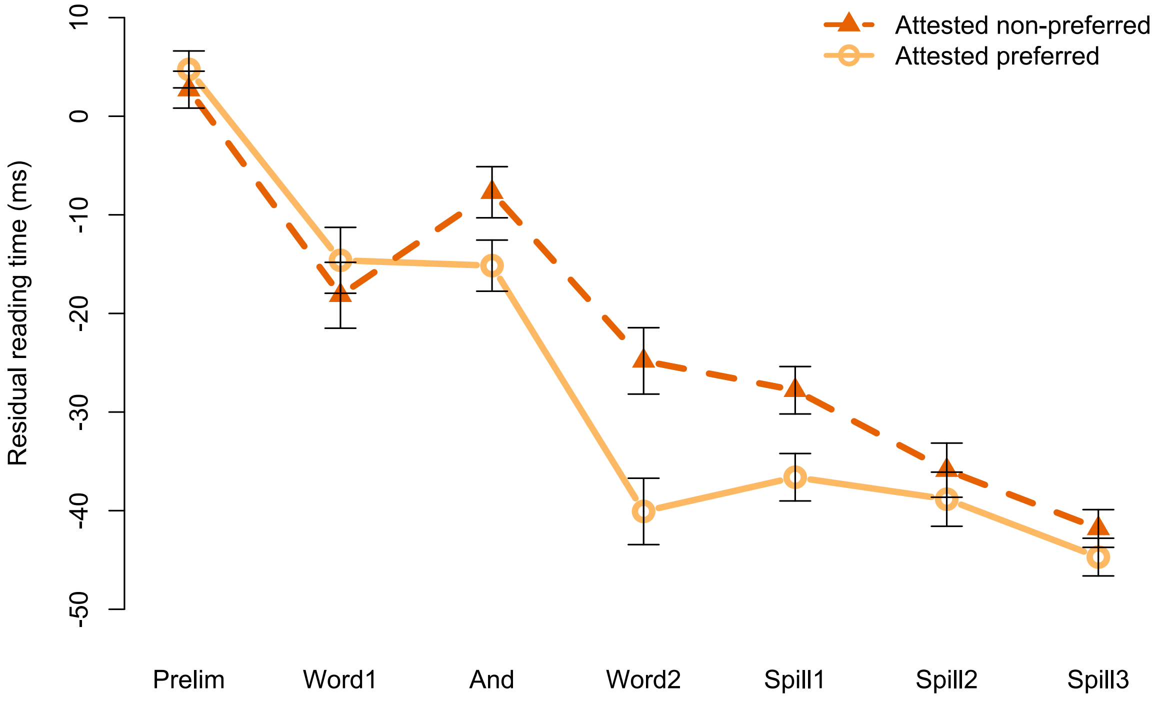

Figure C.9: Self-paced reading times for attested expressions with preferred direction determined by corpus frequency. Error bars show standard errors for the predictor of interest (Table C.7).

Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience.

Arnon, I. and Cohen Priva, U. (2013). More than Words: The Effect of Multi-word Frequency and Constituency on Phonetic Duration. *Language and Speech*, 56(3):349–371.

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1):67–82.

Baayen, R. H., Milin, P., Durdevic, D. F., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3):438–481.

Balota, D. A., Yap, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39:445–459.

Bannard, C. and Matthews, D. (2008). Stored Word Sequences in Language Learning: The Effect of Familiarity on Children's Repetition of Four-Word Combinations. *Psychological Science*, 19(3):241–248.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

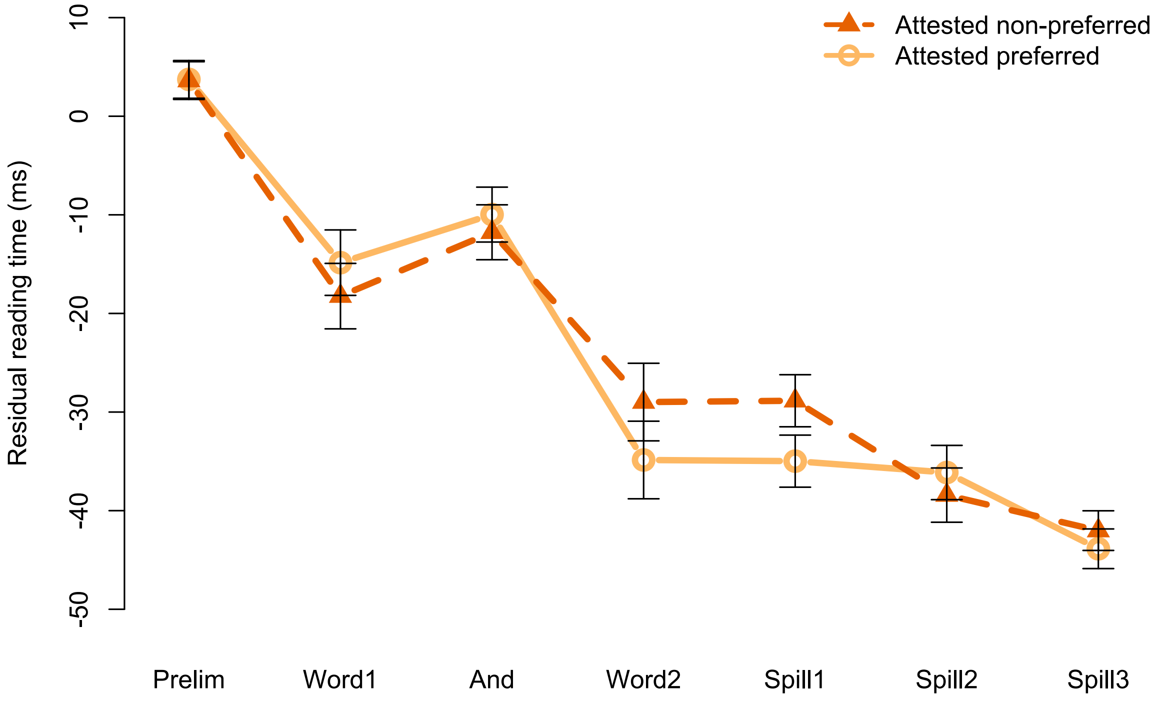

Figure C.10: Self-paced reading times for attested expressions with preferred direction determined by model predictions. Error bars show standard errors for the predictor of interest (Table C.7).

Benor, S. and Levy, R. (2006). The Chicken or the Egg? A Probabilistic Analysis of English Binomials. *Language*, 82(2):233–278.

Bergen, L., Levy, R., and Gibson, E. (2012). Verb omission errors: Evidence of rational processing of noisy language inputs. *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society*, pages 1320–1325.

Bock, J. K. and Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21(1):47–67.

Bod, R. (1998). *Beyond Grammar*. An Experience-Based Theory of Language. Center for the Study of Language and Information, Stanford University.

Bod, R. (2008). The Data-Oriented Parsing approach: Theory and application. *Studies in Computational Intelligence (SCI)*, 115:307–348.

Bod, R., Scha, R., and Sima'an, K., editors (2003). *Data-Oriented Parsing*. University of Chicago Press.

Bolinger, D. L. (1962). Binomials and pitch accent. *Lingua*, 11:34–44.

Brants, T. and Franz, A. (2006). *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia.

	Estimate	*Std. Error*	*t value*
Intercept	2246.62	41.66	53.93
Type: novel	204.18	28.51	7.16
Abs know (Type: attested)	0.72	24.97	0.03
Abs know (Type: novel)	-56.81	19.88	-2.86
Rel freq	-57.46	19.98	-2.88
Trial order	-198.11	9.46	-20.95

Table D.8: Model fit for results of Experiment 2 using raw reading times. Effects with $t > 2$ are taken to be significant. All VIF < 1.6.

Bresnan, J., Cueni, A., Nikitina, T., and Baayen, R. H. (2007). Predicting the dative alternation. *Cognitive Foundations of Interpretation*, pages 69–94.

Bybee, J. (1985). *Morphology.* A Study of the Relation between Meaning and Form. John Benjamins Publishing.

Bybee, J. (1999). Usage-based Phonology. In Darnell, M., Moravcsik, E. A., Noonan, M., Newmeyer, F. J., and Wheatley, K., editors, *Functionalism and Formalism in Linguistics: General papers*, pages 211–242. John Benjamins, Amsterdam.

Bybee, J. (2001). *Phonology and Language Use.* Cambridge University Press.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 82(4):711–733.

Cooper, W. E. and Ross, J. R. (1975). World Order. In Grossman, R. E., San, L. J., and Vance, T. J., editors, *Papers from the Parasession on Functionalism*, pages 63–111. Chicago Linguistics Society, Chicago.

Demberg, V. (2010). *A Broad-Coverage Model of Prediction in Human Sentence Processing.* PhD thesis, The University of Edinburgh.

Dixon, R. M. W. (1982). *Where have All the Adjectives Gone?: And Other Essays in Semantics and Syntax.* Walter de Gruyter.

Elman, J. L. (2003). Generalization from Sparse Input. *Proceedings of the 38th Annual Meeting of the Chicago Linguistic Society.*

Enochson, K. and Culbertson, J. (2015). Collecting Psycholinguistic Response Time Data Using Amazon Mechanical Turk. *PLoS ONE*, 10(3):e0116946–17.

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics*, 27(3):517–556.

Ferreira, F. and Clifton, Jr, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, 25(3):348–368.

Fine, A. B., Jaeger, T. F., Farmer, T. A., and Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, 8(10):e77661.

Gahl, S. and Yu, A. C. L. (2006). Introduction to the special issue on exemplar-based models in linguistics. *The Linguistic Review*, 23(3):213–216.

Gibbs, R. W. (1990). Psycholinguistic studies on the conceptual basis of idiomaticity. *Cognitive Linguistics*, 1(4):417–451.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *TRENDs in Cognitive Sciences*, 7(5):219–224.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., and Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society*, 35:151–166.

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models. *Journal of Statistical Software*, 33(2):1–22.

Harrell, F. E. (2001). *Regression Modeling Strategies.* Springer Series in Statistics. Springer New York, New York, NY.

Hay, J. and Bresnan, J. (2006). Spoken syntax: The phonetics of giving a hand in New Zealand English. *The Linguistic Review*, 23(3):1–29.

Hofmeister, P., Jaeger, T. F., Arnon, I., Sag, I. A., and Snider, N. (2011). The source ambiguity problem. *Language and Cognitive Processes.*

Holsinger, E. (2013). Representing Idioms: Syntactic and Contextual Effects on Idiom Processing. *Language and Speech*, 56(3):373–394.

Inhoff, A. W. and Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, 40(6):431–439.

Jaeger, T. F. (2006). Phonological Optimization and Syntactic Variation: The Case of Optional *that*. *Proceedings of the 32nd Annual Meeting of the Berkeley Linguistics Society BLS.*

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4):434–446.

Johnson, K. (1997). The auditory/perceptual basis for speech segmentation. *OSU Working Papers in Linguistics*, 50:101–113.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, 34(4):485–499.

Johnson, M., Griffiths, T. L., and Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *Advances in Neural Information Processing Systems*, 19:641–648.

Keller, F., Gunasekharan, S., Mayo, N., and Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, 41(1):1–12.

Kelly, M., Bock, J. K., and Keil, F. C. (1986). Prototypicality in a Linguistic Context. *Journal of Memory and Language*, 25:59–74.

Kirby, S., Dowman, M., and Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12):5241–5245.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar*, volume 1 of *Theoretical Prerequisites.* Stanford University Press, Stanford.

Lee, M.-W. and Gibbons, J. (2007). Rhythmic alternation and the optional complementiser in English: New evidence of phonological influence on grammatical encoding. *Cognition*, 105(2):446–456.

Levy, R., Fedorenko, E., Breen, M., and Gibson, E. (2012). The processing of extraposed structures in English. *Cognition*, 122(1):12–36.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., and Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, 449(7163):713–716.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic Annotations for the Google Books Ngram Corpus. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 169–174.

Linzen, T. and Jaeger, T. F. (2015). Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. *Cognitive Science*, pages 1–30.

Malkiel, Y. (1959). Studies in irreversible binomials. *Lingua*, 8:113–160.

McDonald, J., Bock, K., and Kelly, M. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*, 25:188–230.

Menard, S. (2002). *Applied Logistic Regression Analysis.* SAGE Publications, Incorporated.

Mollin, S. (2012). Revisiting binomial order in English. *English Language and Linguistics*, 16(1):81–103.

Morgan, E. (2016). *Generative and Item-Specific Knowledge of Language.* PhD thesis, University of California, San Diego.

Morgan, E. and Levy, R. (2015). Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure. *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society*, pages 1649–1654.

Morgan, E. and Levy, R. (2016). Frequency-Dependent Regularization in Iterated Learning. In Roberts, S. G., Cuskley, C., McCrohon, L., Barceló-Coblijn, L., Fehér, O., and Verhoef, T., editors, *The Evolution of Language: Proceedings of the 11th International Conference EVOLANG.*

Nunberg, G., Sag, I. A., and Wasow, T. (1994). Idioms. *Language*, 70(3):491–538.

O'Donnell, T., Snedeker, J., Tenenbaum, J. B., and Goodman, N. (2011). Productivity and Reuse in Language. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 1613–1618.

Onishi, K. H., Murphy, G. L., and Bock, K. (2008). Prototypicality in sentence production. *Cognitive Psychology*, 56(2):103–141.

Pierrehumbert, J. (2000). Exemplar dynamics: Word frequency, lenition and contrast. In Bybee, J. and Hopper, P., editors, *Frequency and the Emergence of Linguistic Structure*, pages 137–157. John Benjamins, Amsterdam.

Pinker, S. (1991). Rules of Language. *Science, New Series*, 253(5019):530–535.

Pinker, S. (2000). *Words and Rules.* The Ingredients of Language. Harper Perennial, New York.

Pinker, S. and Birdsong, D. (1979). Speakers' sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior*, 18(4):497–508.

Post, M. and Gildea, D. (2013). Bayesian Tree Substitution Grammars as a Usage-based Approach. *Language and Speech*, 56(3):291–308.

R Core Team (2014). *R: A language and environment for statistical computing.* Vienna, Austria.

Rayner, K. and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

Rayner, K., Sereno, S. C., and Raney, G. E. (1996). Eye movement control in reading. *Journal of Experimental Psychology: Human Perception and Performance*, 22(5):1188–1200.

Rayner, K. and Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, 3(4):504–509.

Rumelhart, D. E. and McClelland, J. L. (1986). On Learning the Past Tenses of English Verbs. In McClelland, J. L., Rumelhart, D. E., and the PDP research group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA.

Singh, R., Fedorenko, E., Mahowald, K., and Gibson, E. (2015). Accommodating presuppositions is inappropriate in implausible contexts. Technical report.

Siyanova-Chanturia, A., Conklin, K., and van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3):776–784.

Sobkowiak, W. (1993). Unmarked-before-marked as a Freezing Principle. *Language and Speech*, 36(4):393–414.

Swinney, D. A. and Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, 18(5):523–534.

Tily, H. (2012). flexspr. *Personal communication*.

Truswell, R. (2009). Attributive Adjectives and Nominal Templates. *Linguistic Inquiry*, 40(3):525–533.

Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience*, 2(10):717–726.

Ullman, M. T., Pancheva, R., Love, T., Yee, E., Swinney, D., and Hickok, G. (2005). Neural correlates of lexicon and grammar: Evidence from the production, reading, and judgment of inflection in aphasia. *Brain and Language*, 93(2):185–238.

van den Bosch, A. and Daelemans, W. (2013). Implicit Schemata and Categories in Memory-based Language Processing. *Language and Speech*, 56(3):309–328.

Wiechmann, D., Kerz, E., Snider, N., and Jaeger, T. (2013). Introduction to the Special Issue: Parsimony and Redundancy in Models of Language. *Language and Speech*, 56(3):257–264.

Wright, S. K., Hay, J., and Bent, T. (2005). Ladies first? Phonology, frequency, and the naming conspiracy. *Linguistics*, 43(3):531–561.

Xu, F. and Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2):245–272.