Holistic and Compositional Processing in Multiword Expressions

Wednesday Bushong

September 21, 2015

Abstract

Recent studies have found that high frequency multiword expressions are processed and produced faster than low frequency expressions when unigram frequency is controlled. This evidence is inconsistent with traditional linguistic theories in which there is a sharp distinction between single words and the rules used to combine them into larger strings. Instead, this evidence is in line with usage-based theories which claim that frequent phrases may be explicitly represented as holistic chunks in addition to being built up from single words compositionally. Here, we more directly test the possibility that multiword expressions are stored holistically in a self-paced reading training experiment. We use binomial expressions, phrases of the form "X and Y". They can occur in two word orders, "X and Y" or "Y and X", while preserving lexical items and formal semantic and syntactic structure. We use both frequently attested expressions (such as "alive and well"), and completely novel expressions that are unattested (such as "bishops and seamstresses"). Participants read the expressions in one of the orders in the training phase, and later in the testing phase they saw the same expressions again in either the same or different order. We manipulated binomial type (novel vs. attested) and train order–test order match (match vs. mismatch vs. untrained). We found that highly frequent attested multiword expressions show a training benefit for the match over mismatch condition, but not an overall training benefit for either training condition over the untrained condition, while novel expressions show an overall training benefit, but no benefit for for the match over mismatch condition. This suggests that novel expressions are processed fully compositionally, while attested expressions can be processed as holistic units.

1 Introduction

The frequency of linguistic structures in the world influences processing and production. For example, it is well known that frequent single words are processed and produced faster than infrequent words, and more frequent nouns and verbs are learned earlier by children (Rayner & Duffy, 1986; Jescheniak & Levelt, 1994; Naigles & Hoff-Ginsberg, 1998). Similar effects have been found at other levels of linguistic structure: the overall frequency of a phone predicts its reduction and deletion (Cohen Priva, 2008), and even the frequency of abstract syntactic structures affect production speed and fluency (Gahl & Garnsey, 2004; Jaeger, 2010).

For a long time, linguists have drawn a sharp distinction between the storage of single words in long-term memory, and the storage of knowledge about the abstract operations applied to these single words to form longer utterances (see, e.g., Pinker, 1998). The findings that individual word frequencies and syntactic structure probabilities influence production and processing pose no problem to such a theory: these representations in long-term memroy simply have this information attached to them. However, if multiword frequency effects are found, this poses a crucial problem to the theory since strings of multiple words are supposedly not stored. However, usage-based theories posit that larger strings of words can in fact be stored just like single words (Bybee, 2006).

In fact, recent studies have shown that the frequency of multiword expressions is important for processing and production. For example, four-word phrases such as *don't have to worry* are recognized faster than phrases such as *don't have to wait* when four-gram frequency is different, even when all unigram, bigram, and trigram frequencies are controlled (Arnon & Snider, 2010). Similar effects have been found in production: bigram and trigram frequency modulates speed of production, such that more frequent expressions are produced faster (Jurafsky, Bell, Gregory, & Raymond, 2001; Arnon & Cohen Priva, 2013). Arnon and Cohen Priva (2013) found this effect even when the phrase is not a constituent (e.g., "as far as I"). Bannard and Matthews (2008) found that children made less errors and were quicker to produce single words within more frequent phrase frames. Similar effects have been shown in languages other than English; for example, determiner-noun-adjective utterances are produced faster in a picture naming task with increasing multiword frequency, regardless of unigram frequency (Janssen & Barber, 2012). Together, these findings suggest that humans keep track of the frequency of fairly long strings of words and use these representations during online processing.

However, one alternative explanation of these results is that people process and produce more frequent phrases faster because they communicate more probable events in the world. If *don't have to worry* is a more frequent event than *don't have to wait*, then it is plausible that such phrases might be processed and produced faster, and their corpus frequency might simply be a byproduct of this.

In order to determine whether people actually store longer sequences of words, then, we must find a more unambiguous case where preference cannot be due to other factors. Binomial expressions, phrases of the form "X and Y" (like "alive and well"), are great for this purpose because we can look at both possible orderings of the expression and then lexical items and formal semantic and syntactic structure are identical.

1.1 Binomial Expressions

As with other multiword expressions, frequency effects have been found in binomials as well. Siyanova-Chanturia, Conklin, and van Heuven (2011) found that the more frequent ordering of a binomial is read faster than the less frequent ordering. For convenience, we are going to call the more frequent ordering of a binomial the preferred order, and the less frequent the dispreferred order.

However, there are also other forces that guide word order preference in binomial expressions. Malkiel (1959) was the first to propose that abstract constraints may govern the ordering of the binomials, arguing for semantic, phonological, and lexical constraints. Cooper and Ross (1975) proposed the "Me First" constraint, which dictates that words which are closer to the prototypical

speaker will come first. Benor and Levy (2006) reviewed all of the constraints proposed in the previous literature and a corpus analysis of English binomials. However, the experimental work in this field is sparse.

One of the first studies to experimentally investigate abstract word ordering preferences was Pinker and Birdsong (1979). They gave participants pairs of nonwords in two different orders like “fim-fum" or “fum-fim" and asked them to indicate their preference. They tested a variety of constraints that have been proposed in the literature to be either universal or language-specific. They found that participants reliably had a preference for combinations that aligned with various proposed constraints, which hints that people use these preferences during production of real word sequences.

Recently, Morgan and Levy (submitted) presented a computational model that takes into account a number of these abstract ordering constraints and is trained on highly frequent attested binomials. They found that these seven constraints were the most important for predicting the preferred order of attested expressions:

1. Formal markedness: items that have a more general meaning come first, and superset items come first in binomials which are in a superset-subset relationship. (Example: *linguistic and paralinguistic*)

2. Perceptual markedness: itms which are closer to the “prototypical" speaker (e.g., animate, concrete, etc) come first. (Example: *physical and mental*)

3. Power/intensity: The more powerful and/or intense items should appear first. (Example: *crime and punishment*)

4. Iconic/scalar sequencing: If the two items occur in a sequence, list them in that sequence. (Example: *first and second*)

5. Frequency: the more frequent word comes first.

6. Length: the shorter item comes first.

7. No final stress: if an item has final stress, it should appear first to avoid ending a phrase on a stressed syllable. (Example: *abused and neglected*)

Then, Morgan and Levy (submitted) constructed novel binomials that are unattested across corpora, such as *bishops and seamstresses*, and divided the items into “preferred" and “dispreferred" orders based on model predictions. They found that their model’s predictions for the preferred and dispreferred orders were borne out in a forced-choice preference task. They also found that their model predictions aligned with reading times in a self-paced reading experiment.

However, if abstract ordering constraints were the whole story to binomial expression processing, then we should also see an effect of abstract ordering constraints on the phrasal frequency of highly frequent attested expressions, along with differences in their processing. However, we know from corpus studies that binomial frequency preferences are actually quite variable over time (Mollin, 2013). Furthermore, Morgan and Levy (submitted) looked at highly frequent attested binomials such as *alive and well*, for which we have both model predictions and actual

relative frequency estimates. There were some attested items whose relative frequency aligns with abstract constraint predictions, and some that do not. They found that in these highly attested items, frequency was a much stronger predictor than the model's predictions based on abstract ordering constraints. This suggests that binomial expression processing is dependent on abstract ordering constraints when there is no frequency information, but that frequency becomes a much more important factor as people get more experience with the expression.

1.2 Current Study

The purpose of this study was to determine whether highly frequent multiword expressions are represented as holistic chunks. In order to test this, we use a training paradigm. The experiment was composed of a training phase and a testing phase (although participants were not aware of this). In the training phase, for each item participants were either exposed to it three times, or not at all, and the trained items appeared either in the preferred or dispreferred order. In the testing phase, participants saw all items (including untrained items) in either the preferred or dispreferred order. We looked at both highly frequently attested binomials (such as *alive and well*) completely novel binomial expressions (such as *bishops and seamstresses*).

If attested expressions are represented and processed as holistic chunks, then we should be able to prime exact word order – that is, items should be faster for the order they were trained in than for the order they were not trained in. However, novel expressions should not show such a training benefit.

1.2.1 Predictions

First, we expected to replicate the findings of Siyanova-Chanturia et al. (2011) and Morgan and Levy (submitted) in attested binomials. That is, for our untrained attested items, we should see a preference for the preferred order over the dispreferred order. We should also observe an ordering preference in untrained novel items in line with Morgan and Levy (submitted). In that sense, novel and attested items should behave identically in the untrained condition. However, we predict different behavior in the binomial types in the training conditions.

If attested expressions are represented holistically as well as compositionally, then they should show a training benefit for exact form. That is, they should be read faster in the order they were trained in than the order they were not trained in. We might also expect that single words should contribute less to their processing. We also would expect that trigram frequency will influence the processing of attested expressions. That is, the training benefit should be larger for the items that are trained in the preferred than the dispreferred order in the attested items. If frequency of expression strengthens that item's holistic representation, then we would expect that the exact form training benefit would increase with increasing frequency of the expression.

In novel expressions, we would expect there to be a large overall training benefit due to the priming of the individual words within the phrase. Furthermore, we would expect for unigram frequency to play an important role in their processing. However, we should not see a word order training benefit in the novel items, since there are no existing holistic representations for them.

2 Method

2.1 Participants

207 participants were recruited from Amazon Mechanical Turk and were rewarded $4.00 each for their participation. All participants were native speakers of English and had learned no other languages before age 7.

2.2 Materials

We used 96 binomials expressions: 48 frequently attested, and 48 completely novel. Within attested items, we determined preferred and dispreferred orders based on the relative frequency, such that the order with higher relative frequency was the preferred order. To determine the preferred and dispreferred orderings in the novel binomials, we used a model presented in Morgan and Levy (submitted). The model takes into account seven different constraints that have been claimed to be important in binomial ordering: formal markedness, perceptual markedness, power, frequency, no final stress, and length. See the appendix for a full list of our items.

For each item, we constructed four sentence contexts that did not bias ordering one way or the other. The sentence contexts for each item were constructed to be somewhat different from each other so that participants did not have any extra cues in the sentence to bias them towards expecting one order or another. That is, we made sure that we did not make reference to anything related to either the first or the second item in the binomial. We also always used a fixed sentence context for the test items across participants so that conditions were as comparable as possible. This is an example sentence from our materials:

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

All the words in the sentence up to the first part of the binomial are the introductory region. The binomial expression was the critical region. The three words after the binomial are the spillover region, and the rest of the words in the sentence up to the final word was the postcritical region. In our analyses we generally collapse across the critical, spillover, and postcritical region, although we do have some region analyses below. We exclude final words from analyses because they are generally uninformative about earlier effects in the sentence.

2.3 Procedure

There were two phases in the experiment: training and testing. In the training phase, participants saw each item either three times (trained) or not at all (untrained). Within the trained items, each item was either in the preferred or dispreferred order, and was consistent in ordering within the same item. In the training phase, two-thirds of the sentences, including both fillers and critical sentences, were presented as full sentences rather than self-paced reading, in order to shorten the total length of the experiment. In the testing phase, participants saw all of the items (including the untrained items) in either the preferred or dispreferred order. All sentences in the test phase were self-paced reading. All manipulations were within subjects, so that each participant saw all of the

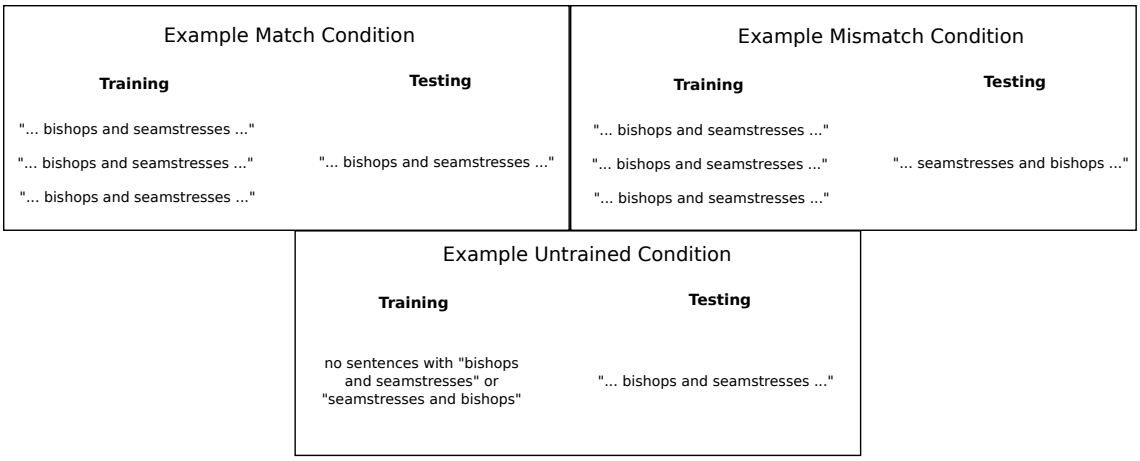


Figure 1: Examples of match, mismatch, and untrained conditions.

Binomial Type	**Train Order**	**Test Order**	**Train-Test Match**
Attested	None	Preferred	Untrained
Attested	None	Dispreferred	Untrained
Attested	Preferred	Preferred	Match
Attested	Dispreferred	Dispreferred	Match
Attested	Preferred	Dispreferred	Mismatch
Attested	Dispreferred	Preferred	Mismatch
Novel	None	Preferred	Untrained
Novel	None	Dispreferred	Untrained
Novel	Preferred	Preferred	Match
Novel	Dispreferred	Dispreferred	Match
Novel	Preferred	Dispreferred	Mismatch
Novel	Dispreferred	Preferred	Mismatch

Table 1: All conditions in the experiment.

conditions 3 times in different items. Although we had 96 items, participants only saw a subset of 36 items, since otherwise the experiment would have been too long. Approximately once every five sentences, there would be a yes/no comprehension question about the content of the preceding sentence. Participants who were below 75% accuracy on these comprehension questions were removed from analyses.

In total, the design was a 2 x 3 x 2 (type x train order x test order). However, we also analyzed a 2 x 3 with type and train-test match as predictors, where a match corresponds to an item being trained and tested in the same order, a mismatch corresponds to an item being trained in the opposite order of testing, and untrained are items that were not seen in training (see Figure 1 for examples of match, mismatch, and untrained conditions, and Table 1 for a table of all conditions in the experiment).

Although we draw a distinction between training and testing periods here, participants were unaware of this and there was no break in the experiment between training and testing. However, since there were no sentences presented as full sentences during the testing period, participants might have noticed that. In any case, it is unclear whether or how this would change their reading behavior.

2.4 Data Analysis

We residualized reading times on word length for each subject relative to RTs on all test sentences We then aggregated the residualized RTs from the first word of the binomial through the penultimate word of the sentence to get an average per-word residualized RT for each trial. We did this because most effects in self-paced reading emerge between the critical region and the immediate spillover region. However, we do present one word-by-word analysis in order to explore the effects more in-depth (see Section 3.2). We removed residual RTs that were greater than three standard deviations away from the overall mean across all test sentences in the experiment. We fit linear mixed-effects models with maximal random effects structures as justified by the design (see Barr et al., 2013). All models were fit using the function in the R package unless otherwise noted (Bates, Maechler, Bolker, & Walker, 2014). Reported p-values are from model comparisons.

3 Results

We present four sets of results here. First, we fit a model with binomial type and train-test match as predictors, to get a feel for the general pattern of results. Train-test match is a factor with three levels: match, mismatch, and untrained, where match corresponds to trained preferred & tested preferred or trained dispreferred & tested dispreferred, a mismatch corresponds to trained preferred & tested dispreferred or trained dispreferred & tested preferred. We then fit a model that used binomial type, train order, and test order results in order to more fully understand what drives the general matching results. Then, we investigate how various different factors modulate these results. Specifically, we look at the effect of unigram frequency of the individual words within the phrase, measures of phrasal frequency in the attested items (overall trigram frequency, specific trigram frequency, and relative frequency), as well as model predictions for ordering preferences.

Predictor	Estimate	Standard Error	t Value
Intercept	8.79287	3.48704	2.522
Binomial Type (Novel)	14.10616	3.53797	3.987
Train-Test Match (Match)	-3.23297	2.29296	-1.410
Train-Test Match (Mismatch)	0.44563	2.28169	0.195
Trial Number	-0.34861	0.05409	-6.445
Binomial Type (Novel) x Train-Test Match (Match)	-6.71990	3.41903	-1.965
Binomial Type (Novel) x Train-Test Match (Mismatch)	-11.28646	3.30006	-3.420

Table 2: Type x Train-Test Match results. Note that this table does not contain information about match vs. mismatch comparisons, since there are three levels of train-test match and the reference level is set at untrained.

Finally, we do some simple analyses of the training phase data.

3.1 Train-Test Match Results

We fit a linear mixed-effects model with a maximal random-effects structure using binomial type and train-test match as predictors[1]. We found a main effect of binomial type, such that attested expressions were read faster than novel expressions ($p<0.007$). There was also a main effect of train-test match ($p<0.001$). There was a 2x3 interaction between binomial type and train-test-match ($p<0.004$). See Table 2 for a summary of results.

To understand the interaction further, we did three 2x2 comparisons for the levels of train-test match. There was a significant interaction between binomial type and train-test match within untrained and mismatch ($p<0.002$); pairwise comparisons revealed that the interaction was driven by novel items being read faster in the mismatch as compared to the untrained condition ($p<0.0001$), while there was no difference in attested items ($p=0.88$). There was also a significant interaction between binomial type and train-test match within untrained and match ($p=0.036$); further pairwise comparisons showed that this interaction was driven by the fact that novel items were read faster in the match as compared to the untrained condition ($p<0.001$), while there was only a nonsignificant trend in attested items ($p=0.097$). Finally, we observed a marginally significant interaction between binomial type and train-test match within the math and mismatch conditions ($p=0.073$); pairwise comparisons showed that attested items were read significantly faster in the match as compared to the mismatch condition ($p=0.057$), while there was no such difference in novel items ($p>0.9$).

3.1.1 Discussion

The main effect of binomial type is unsurprising given that phrasal frequency for attested items is by definition higher than novel items. However, it is very striking that attested items showed no overall training benefit for either of the training conditions as compared to the untrained condition

[1]Full model specification:

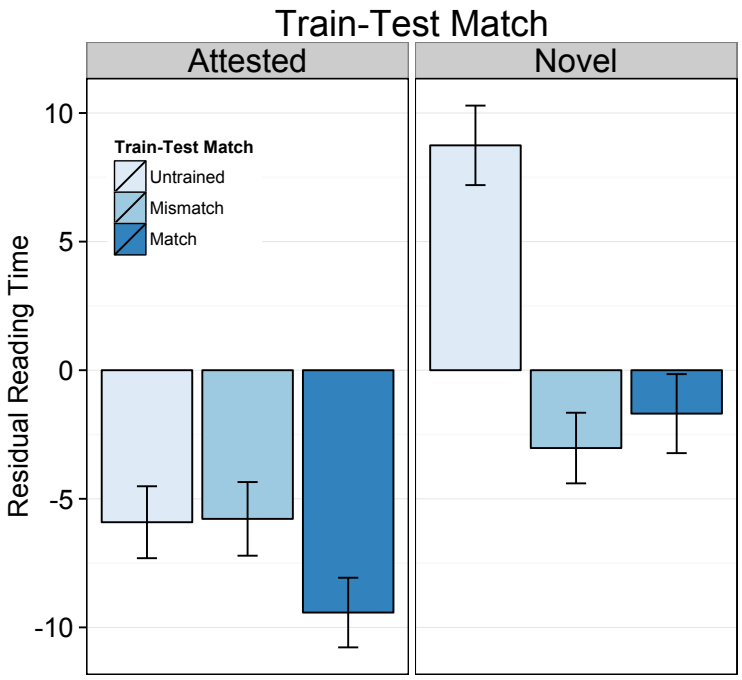


Figure 2: Train-test match results in novel and attested items.

(although the match condition was read numerically faster than the untrained condition). We interpret this as suggesting that the individual words within an attested item are not being activated during training.

One alternative explanation for these results is on the basis of unigram frequency. Lower frequency words tend to show larger priming effects (Scarborough, Cortese, & Scarborough, 1977), and our attested items had on average higher frequency words (attested mean log frequency=10.37, novel=6.43, $p<0.001$). However, it seems unlikely that this effect was driven solely by the unigram frequency differences between the items since attested items should show at least some training benefit if individual words within the expression were being primed. In any case, we decided to do an additional analysis of unigram frequency of the individual words within the expressions and whether this affects novel and attested expressions differently.

Our second main result is that attested expressions showed a word order-specific training benefit, while novel expressions did not. We take this as evidence that comprehenders are activating holistic representations of attested expressions during processing, which causes later processing of the same holistic chunk to be faster. When the comprehender comes across the exact string *alive and well*, they activate their holistic representation of *alive and well*. Later on in the experiment, if they encounter *alive and well* again, they are faster at processing it since the chunk is already activated. However, if they encounter *well and alive*, there should be no processing advantage since *well and alive* is a separate chunk and was not previously activated. By contrast, in

Attested Expressions Word-by-Word					
Match vs. Mismatch		**Mismatch vs. Untrained**		**Match vs. Untrained**	
Region Name	t Value	Region Name	t Value	Region Name	t Value
Word 1	-0.140	Word 1	-0.318	Word 1	-0.378
And	-0.735	And	-2.196*	And	-2.906*
Word 2	-1.723	Word 2	-1.218	Word 2	-2.551*
Spill 1	-1.097	Spill 1	0.260	Spill 1	-0.889
Spill 2	-2.699*	Spill 2	1.492	Spill 2	-1.226
Spill 3	-0.719	Spill 3	-0.781	Spill 3	-1.518

Table 3: Word-by-word results in attested expressions.

Novel Expressions Word-by-Word					
Match vs. Mismatch		**Mismatch vs. Untrained**		**Match vs. Untrained**	
Region Name	t Value	Region Name	t Value	Region Name	t Value
Word 1	1.047	Word 1	-0.732	Word 1	0.174
And	-0.104	And	-1.896	And	-2.166*
Word 2	0.461	Word 2	-4.268**	Word 2	-3.345*
Spill 1	-1.060	Spill 1	-2.780*	Spill 1	-3.886*
Spill 2	-0.589	Spill 2	-2.709*	Spill 2	-3.250*
Spill 3	0.897	Spill 3	-2.439*	Spill 3	-1.491

Table 4: Word-by-word results in novel expressions.

novel items such holistic representations should not exist. Therefore, when a comprehender comes across a string like *bishops and seamstresses*, they activate the individual words *bishops*, *and*, and *seamstresses*. When they encounter *bishops and seamstresses* or *seamstresses and bishops* later, they should be faster for either order since they have previously activated the individual words that make up the phrase.

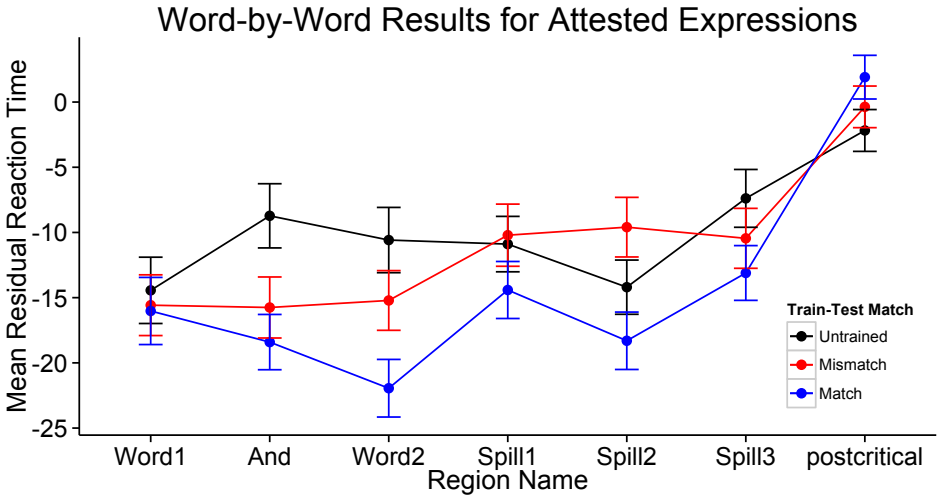
It is possible that there are particular regions that are important for the differences we observe in the novel and attested expressions, so in the next section we do a word-by-word analysis.

3.2 Word-by-Word Train-Test Match Results

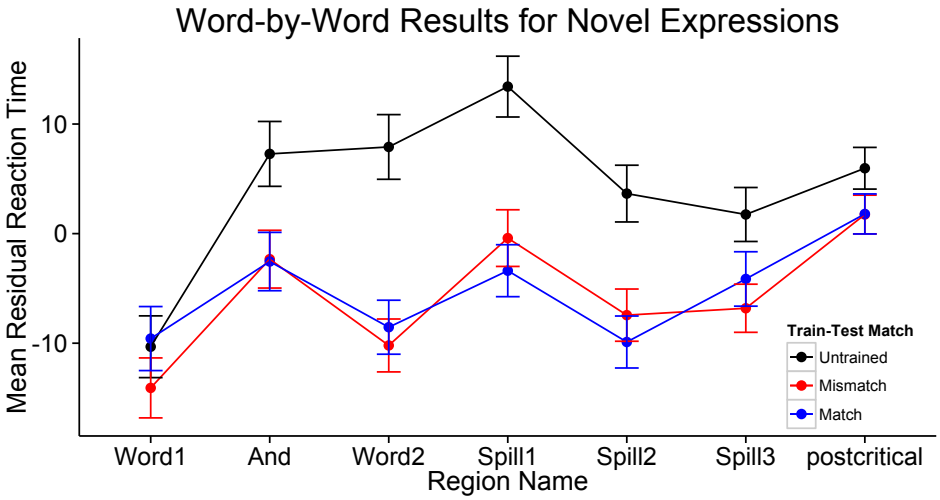
Within attested items, at Word 2, the match condition was faster than the untrained condition and trended toward being faster than the mismatch condition (p=0.014; p=0.09). At Spill 2, the match condition was read faster than the mismatch condition (p=0.009). There was one region, And, where the match and mismatch condition were both read faster than the untrained condition (p=0.004; p=0.028).See Table 3 and Figure 3a for the attested expression results at all regions.

Within novel items, there were no differences between the match and mismatch condition at any region. The match condition was read faster than the untrained condition at And, Word2,

Spill1, and Spill2 (t's<-2, p's<0.05). The mismatch condition was read faster than the untrained condition at Word2, Spill1, Spill2, and Spill3 (t's<-2.4, p's<0.05). See Table 4 and Figure 3b for the novel expression results at all regions.



(a)

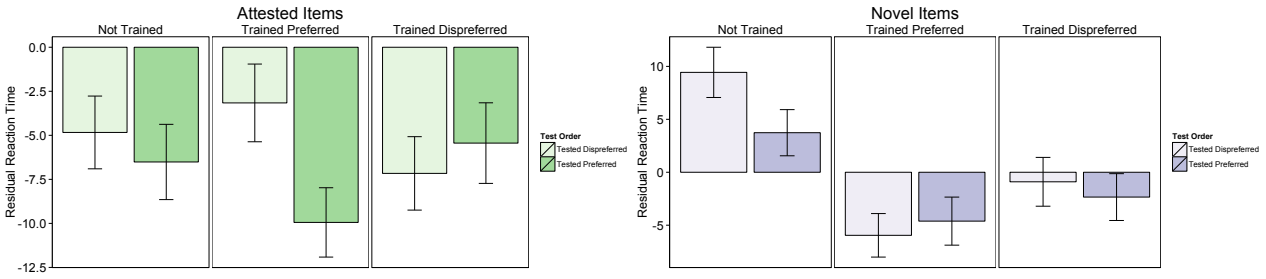


(b)

Figure 3: Word-by-word results for attested and novel expressions.

3.2.1 Discussion

The word-by-word analysis revealed that the benefit for the match over mismatch condition in attested expressions was driven by a difference at Word 2 and Spill 2 (although the difference



(a) Attested items.

(b) Novel items.

Figure 4: Train order x test order results for attested and novel items.

at Word 2 was only marginally significant). Furthermore, it seems that this is due a processing advantage for the match condition at Word 2, since the match condition was faster than both the mismatch and the untrained while there was no difference between the mismatch and untrained conditions. But at Spill 2, it might be a processing disadvantage for the mismatch condition, since the mismatch condition was read more slowly than the match condition and trended toward being read more slowly than the untrained condition, while there was no difference between the match and untrained conditions themselves.

Within the novel items, we found that there were no differences between the match and mismatch conditions in any region, but a large, sustained difference between both training conditions and the untrained condition at similar regions. This suggests that there really is no exact form priming benefit at all in these expressions since there is no difference in the conditions at any region in the sentence. This provides additional evidence that the mismatch and match conditions behave identically in novel expressions.

Next, we present further analyses incorporating the specific train and test orders. We thought that this analysis was important for multiple reasons. First, as noted in the introduction, within the attested items there are large relative frequency differences for each of the orderings. For example, *alive and well* is much more frequent than *well and alive*. If our hypothesis that holistic representations are needed for the word order-specific effect, then we might expect to see different behavior in the dispreferred order as compared to the preferred order. Since the dispreferred order is lower frequency than the preferred order, but higher frequency than the completely novel expressions, we should see some intermediate word order-specific priming behavior in the attested expressions that are trained in the dispreferred order (mean trigram frequency of attested preferred order: 12.95; mean trigram frequency of attested dispreferred order: 9.97; $p<0.001$). Furthermore, if word order-specific priming is wholly dependent on phrase frequency and not on abstract ordering constraints, then we should see no word-order specific priming effects in the novel preferred order. In the section below we present results for this analysis.

Predictor	Mean	95% CI	95% CI	pMCMC
Intercept	10.9510	3.5389	17.7859	0.002 **
Type (Novel)	15.2469	8.0014	23.4002	<0.001 ***
Train Order (Preferred)	0.1545	-6.7709	7.0065	0.972
Train Order (Dispreferred)	-3.1605	-9.6207	3.7214	0.358
Test Order (Preferred)	-3.4588	-8.3453	2.9576	0.246
Trial Number	-0.3630	-0.4596	-0.2570	<0.001 ***
Type (Novel) x Train Order (Pref)	-14.2935	-24.3694	-4.9271	<0.001 ***
Type (Novel) x Train Order (Disp)	-7.9976	-16.6414	1.8236	0.092 .
Type (Novel) x Test Order (Pref)	-1.4590	-9.1788	6.8449	0.716
Train Order (Pref) x Test Order (Pref)	-2.8999	-11.2537	4.0520	0.494
Train Order (Disp) x Test Order (Pref)	3.8965	-3.2311	10.9225	0.312
Type (Novel) x Train Order (Pref) x Test Order (Pref)	7.9874	-3.1062	18.5157	0.142
Type (Novel) x Train Order (Disp) x Test Order (Pref)	-1.2669	-12.2542	8.6469	0.832

Table 5: Model fits for type x train order x test order model. The reference level for train order is "untrained".

3.3 Train Order & Test Order Results

Because we had three different factors with either two or three levels and large random effects structure, we experienced difficulty with model non-convergence using . Therefore, for the results of the initial full model reported here, we used the R package [2] (Hadfield, 2010). Smaller models for 2x2s and pairwise comparisons and were fit using as in the section above.

We found a main effect of binomial type in this parameterization of the model as well ($p<0.001$). We also found two 2x2 interactions between binomial type and train order, such that novel items were read faster if they were trained in the preferred order or dispreferred order compared to untrained than in attested items ($p<0.001$, $p=0.092$), echoing the overall training benefit we found in the alternative parameterization of the model. See Table 5 for the full model fit.

We then broke the model down into separate 2x2 binomial type x train order comparisons within each level of test order. Within items tested in the preferred order, there was a numerical trend toward an interaction between binomial type and train order such that training in the preferred or dispreferred order decreased reading times as compared to the untrained to a greater extent than in attested items ($p=0.174$, $p=0.128$). There was no interaction between binomial type and train order for trained preferred vs. dispreferred ($p=0.66$).

Within items tested in the dispreferred order, there was a significant interaction between binomial type and train order such that training in the preferred order decreased reading times compared to the untrained condition to a greater extent in novel than attested items ($p=0.008$). There was a numerical trend toward a similar interaction for trained dispreferred vs. the untrained condition although this did not reach significance ($p=0.114$). Pairwise comparisons revealed that this was because attested items tested in the dispreferred order were read numerically faster when trained

[2] Full model specification:

in the dispreferred as compared to the preferred order (t=–1.22), while there was a numerical difference in the opposite direction within novel items (t=1.048).

Within attested items, there was no 2x3 interaction between train order and test order (p=0.19). We broke this down into 3 separate 2x2s based on the possible combinations of train order. We found no significant interaction between test order and train order for untrained vs. trained preferred or untrained vs. trained dispreferred (p's>0.1). There was a marginally significant interaction for test order x trained preferred vs. trained dispreferred (p=0.073), such that items trained in the preferred order were read faster in the preferred order as compared to items trained in the dispreferred order.

Within novel items, we found no significant interactions between train order and test order (p's>0.1).

Within attested untrained items, we found no difference between the preferred and dispreferred orders (p=0.258). We found the same results in the novel untrained items (p=0.73).

3.3.1 Discussion

Further investigating the results by train order and test order revealed that the match-mismatch difference within attested items is mainly driven by the items trained in the preferred order. Since the dispreferred order in attested items is in general much less frequent than the preferred order, this makes sense since here we are arguing that phrasal frequency contributes to representations.

Within untrained items, we did not find the expected pattern of the preferred order being read faster than the dispreferred. This might be due to inferences that participants make over the course of the experiment. By the time participants get to the testing period, they have already seen 72 sentences with binomials in them, of which half have been in the preferred and half in the dispreferred order. Participants might implicitly learn that their prior expectations about the distribution of the preferred and dispreferred orders are incorrect in the current situation, and adjust their priors so that they do not expect one order or the other in new items (see Fine, Jaeger, Farmer, & Qian, 2013 and Fine & Jaeger, 2013 for examples of this in syntactic priming). In order to get a less biased measure of participants' original priors, we decided to look at the first exposure to an item in the training period. Another alternative explanation for why we did not see a difference between preferred and dispreferred items in the untrained items is that there might be a difference between the reading times that we can detect using finer-grained frequency measures as opposed to the categorical distinction we have drawn here. Therefore in the next section we investigate whether continuous frequency measures predict reaction times. There are two different types of frequency that we can measure in attested binomials. Firstly, there is relative frequency of an order within a binomial expression, which was what was found to be important in Siyanova-Chanturia et al. (2011). Relative frequency is a value between 0 and 1 calculated from the trigram frequency of a particular ordering divided by the sum of the trigram frequencies of each ordering ("alive and well" / "alive and well" + "well and alive"). There are two other measures of frequency that might be important as well. Firstly, overall trigram frequency, or the sum of the trigram frequencies of each ordering ("well and alive" + "alive and well"). And second, the trigram frequency of a particular ordering (frequency of "alive and well").

In the next section, we investigate the role of unigram and trigram frequency on these effects.

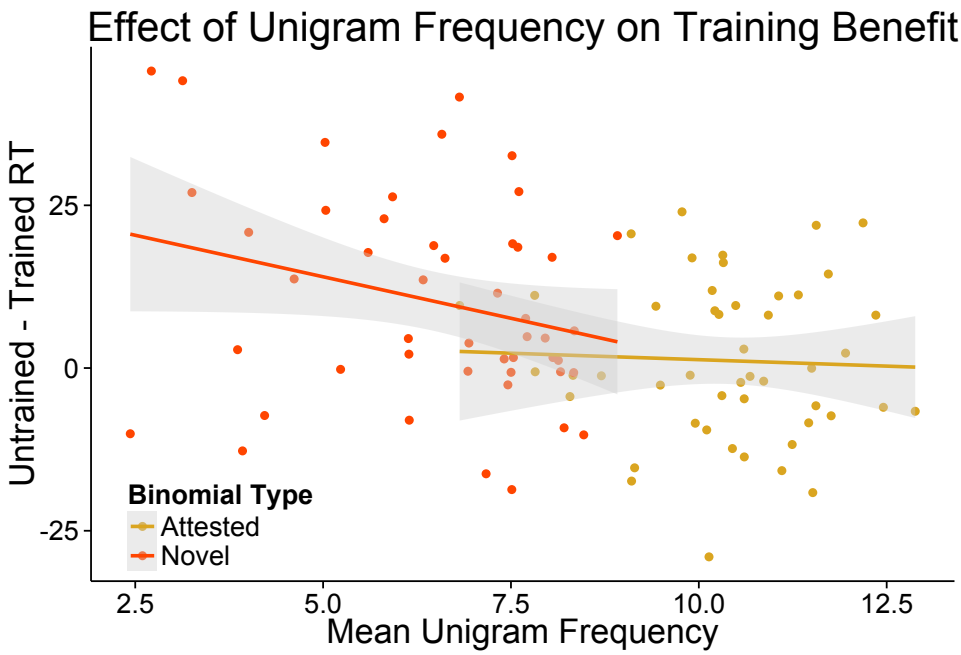


Figure 5: Effect of unigram frequency on overall training benefit in novel and attested items.

3.4 Abstract Ordering Constraints, Unigram and Trigram Frequency Effects

3.4.1 Unigram Frequency

In order to further investigate the difference in overall training benefit between novel and attested expressions discussed in the previous section, we asked whether the average unigram frequency of the words within the expressions influences the overall training benefit to a greater extent in the novel than in the attested expressions[3]. We found a marginal interaction between binomial type and mean unigram frequency (p=0.09), such that mean unigram frequency influenced reading times in novel but not attested items (see Figure 5 and Table **??** for a summary of results). We then looked at novel and attested items separately to further understand how mean unigram frequency influenced both types of expressions.

Within novel expressions, there was a significant interaction between mean unigram frequency and train-test match (p=0.024). Further interactions and pairwise comparisons revealed that mean unigram frequency influenced reading times more in the untrained condition as compared to the match condition (p=0.047), such that reading times trended toward decreasing as mean unigram frequency increased in the untrained condition (p=0.09), while this was not the case in the match condition (p=0.96) or the mismatch condition (p=0.55). There was no interaction between mean unigram frequency and train-test match between the mismatch and match conditions (p=0.54).

Within attested expressions, we found no main effect of mean unigram frequency and no in-

[3]Full model specification:

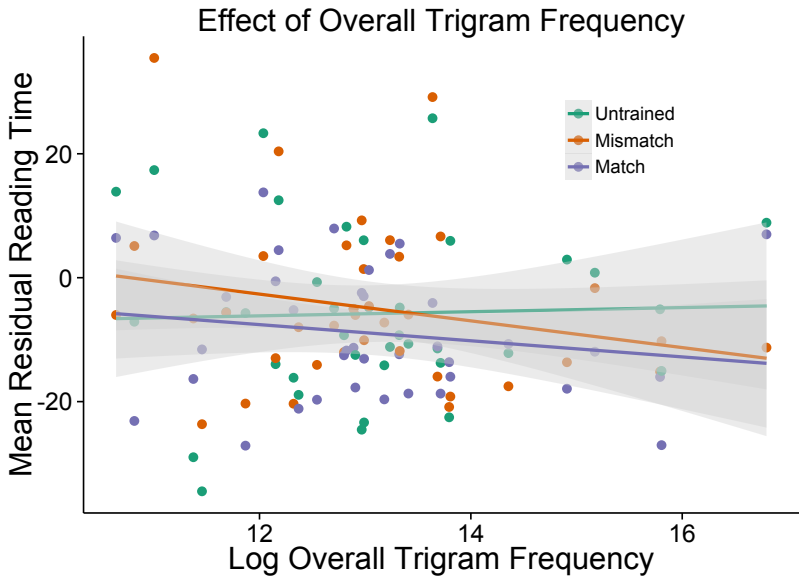


Figure 6: Effect of overall trigram frequency of unordered binomial pair within attested items.

teraction between mean unigram frequency and train-test match (p's>0.4). Within each level of train-test match, there was no effect of mean unigram frequency (p's>0.4 in pairwise comparisons within each condition).

3.4.2 Phrase Frequency in Attested Items

We investigated phrasal frequency within attested items in three ways: first, we looked at overall trigram frequency of the unordered binomial (i.e., trigram frequency of "alive and well" + trigram frequency of "well and alive"). We call this overall trigram frequency. Then, we looked at trigram frequency for each individual ordering in each item (i.e., trigram frequency of "alive and well", excluding trigram frequency of "well and alive"). We call this specific trigram frequency. Finally, we looked at relative frequency of ordering within the binomial item (i.e., trigram frequency of "alive and well" / overall trigram frequency of "alive and well" + "well and alive"). We call this relative frequency. For specific trigram frequency and relative frequency, we also looked at the influence of this factor separately for train order and test order.

Overall Trigram Frequency. We found no main effect of overall trigram frequency (p=0.46). Further comparisons revealed no interaction between overall trigram frequency and train-test match for all pairs of train-test match (match vs. mismatch: p=0.46; untrained vs. match: p=0.59; untrained vs. mismatch: p=0.1). Further pairwise comparisons revealed no effect of overall trigram frequency in any training condition, although there was a nonsignificant trend in the untrained items for more frequent expressions to be read faster (match: p=0.39; mismatch: p=0.85; untrained: p=0.16).

Specific Trigram Frequency of Train Order. We found a marginally significant interaction be-

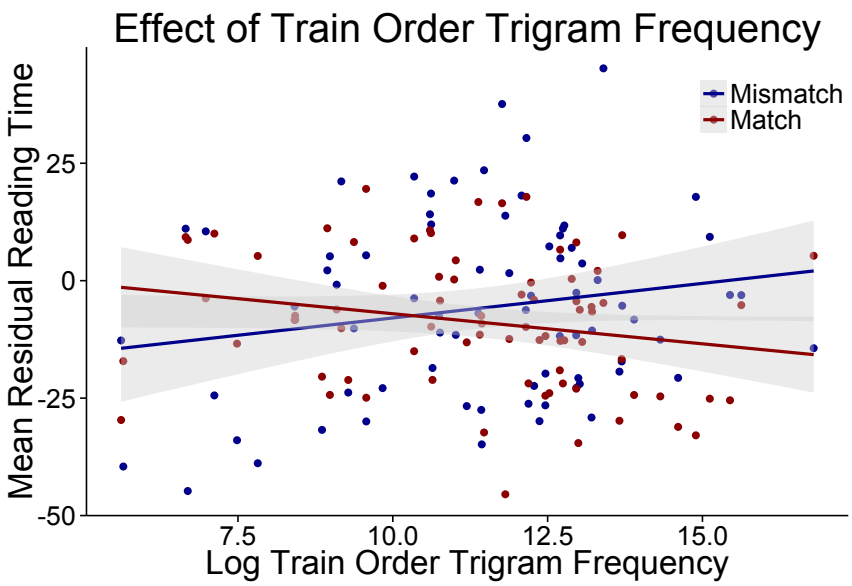


Figure 7: Effect of train order trigram frequency on the match and mismatch conditions.

tween specific trigram frequency of the training order and train-test match within the match and mismatch conditions (p=0.09), such that the reading time difference between the match and mismatch conditions grew larger as trigram frequency increased. Pairwise comparisons revealed that there was a numerical trend for the reading times in the mismatch condition to increase as trigram frequency increased, but this was not significant (p=0.16). There was no modulation of reading times by trigram frequency in the match condition (p=0.64) (see Figure 7).

Specific Trigram Frequency of Test Order. We found a main effect of specific trigram frequency of testing order, such that more frequent orders were read faster (p=0.026). There was no interaction between train-test match and specific trigram frequency (p=0.8), nor were there any interactions when train-test match was broken down into each possible combination (all p's>0.4) (see Figure 8).

Relative Frequency of Train Order. There was no main effect of relative frequency of training order (p=0.155). We also found no interaction between relative frequency of training order and train-test match (p=0.76). There was no effect of relative frequency of training order in either of the train-test match conditions (mismatch: p=0.28; match: p=0.33) (see Figure 9).

Relative Frequency of Test Order. We found a main effect of relative frequency of testing order, such that more relatively frequent orders were read faster (p=0.034). We found no interaction between relative frequency and train-test match (p=0.85), and no interaction between any relative test order frequency and any combination of train-test match (p's>0.6). However, we did find a marginal effect of relative frequency of test order within untrained items such that untrained items trended toward being read faster when the relative test order frequency was higher (p=0.07) (see Figure 10).

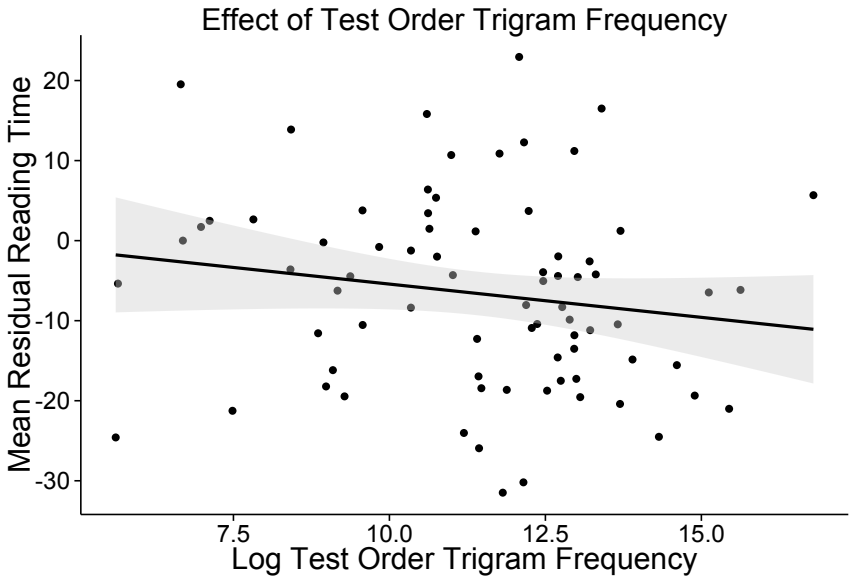


Figure 8: Effect of test order trigram frequency on reading times. The effect was uniform across training conditions.

3.4.3 Model Predictions for Abstract Ordering Constraints

Within both novel and attested untrained items, we did not observe and effect of model prediction for test order (p's>0.3), nor any interactions between model predictions and train-test match (p's>0.2).

3.4.4 Discussion

The unigram frequency analysis revealed that in novel binomials, mean unigram frequency was most important for reading times in the untrained order, and did not play a role in the reading times of items in the other conditions. However, mean unigram frequency did not play a role in reading times of untrained attested items. This suggests that the frequency of individual words within a multiword expression matters less for more frequent phrases. This is consistent with the view that frequent phrases are processed at least somewhat noncompositionally.

Within attested items, we looked at a number of phrasal frequency measures: overall trigram frequency of the unordered pair of words, trigram frequency of the specific order (for both train order & test order), and relative frequency of an ordering within a binomial (for both train order & test order). We found no effects of overall trigram frequency or relative frequency. Instead, the most important factor was the trigram frequency of the specific order. Specific trigram frequency of the training order significantly interacted with train-test match, such that higher-frequency training orders resulted in a larger difference between the mismatch and match conditions at test. This suggests that higher frequency phrases are stored as holistic chunks, with lower frequency orders being activated less.

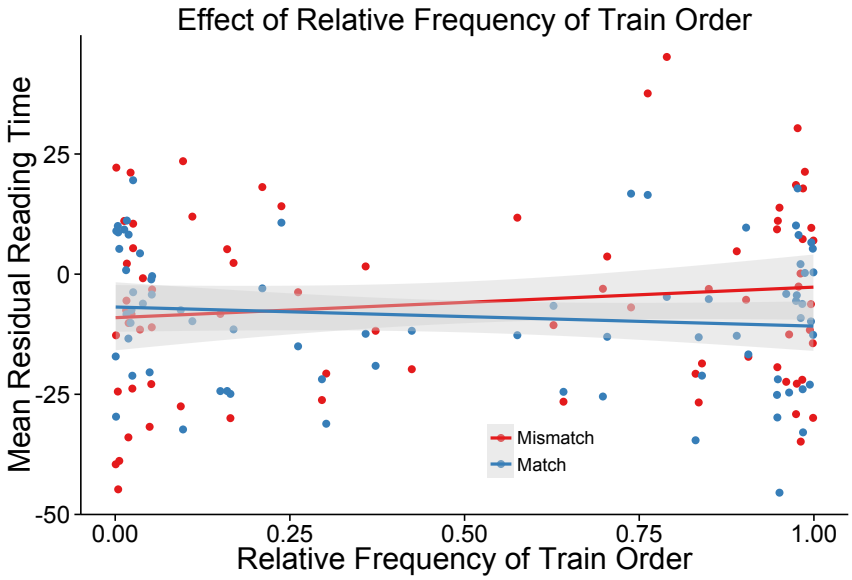


Figure 9: Effect of relative frequency of train order in attested items.

Previous experiments (Siyanova-Chanturia et al., 2011, Morgan & Levy, submitted) found that the more relative frequent order of a binomial is read faster. In line with this, we found a significant effect of relative frequency and trigram frequency of test order on reading times. However, we did not find the effect just in the group means (where preferred is relative frequency >0.5 and dispreferred is relative frequency <0.5). However, we did not find the analogous effect of model predictions in the untrained novel expressions, although we did observe a numerical difference. It is possible that we did not have enough data, or a strong enough method to pick up the differences in the novel items.

Finally, we turn to an analysis of results from the training period of the experiment.

3.5 Training Phase Results

We additionally analyzed the portion of the training data that was self-paced reading. We found a main effect of binomial type ($p<0.004$). Within attested items, there was no effect of order ($p=0.6$), which was also the case for novel items ($p=0.48$). This was also the case when we restricted analysis to only the first exposure to an item in the training phase (attested: $p=0.55$; novel: $p=0.34$).

3.5.1 Discussion

We did not find an effect of ordering preference in the training phase data, failing to replicate previous findings (Siyanova-Chanturia et al., 2011, Morgan & Levy, submitted). However, remember that two-thirds of the trials in the training period were presented as full sentences, and here we only analyzed trials that were self-paced reading, so we had much less data than we probably need

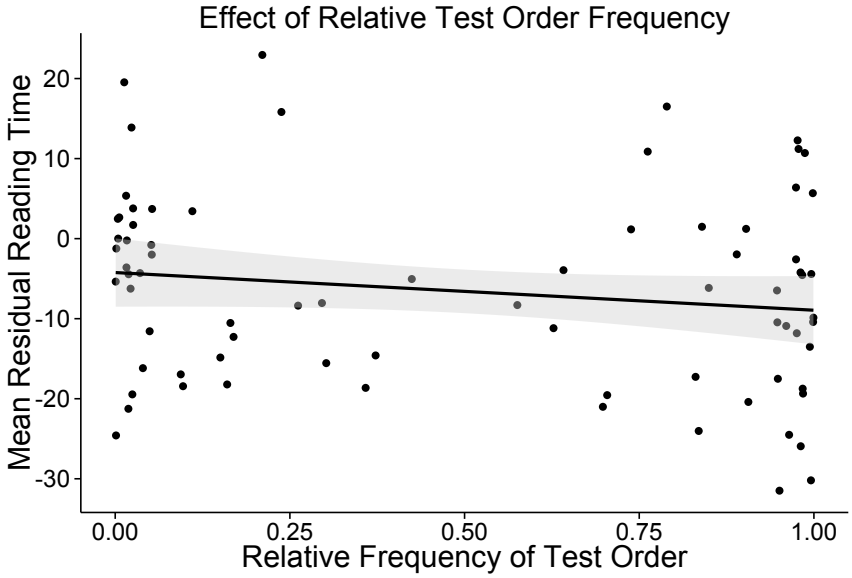
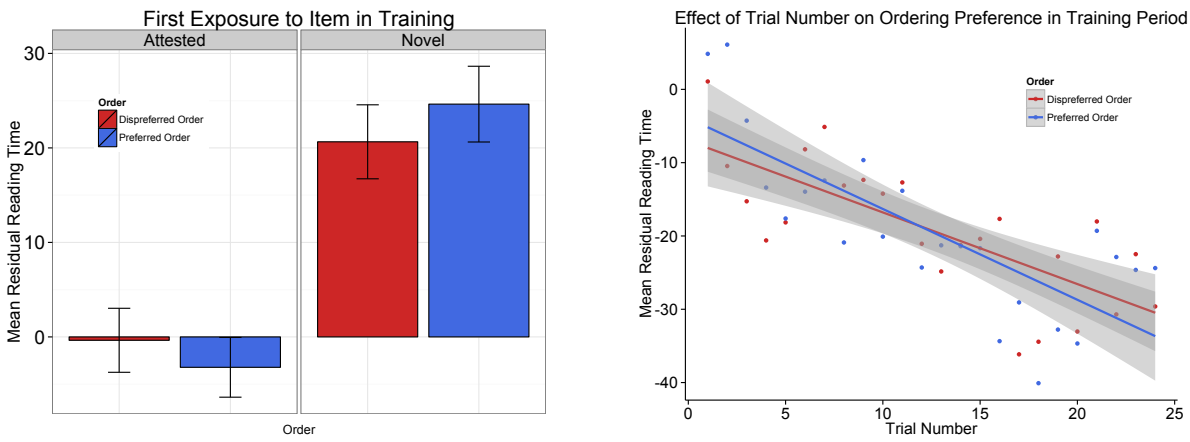


Figure 10: Effect of relative frequency of test order in attested items.

to replicate this effect. Furthermore, it is possible that by the time someone is first exposed to a particular item, they might have already seen another item more than once, influencing their expectations about items they have not yet seen in the experiment.

4 General Discussion

We found that attested multiword expressions show word order-specific priming effects, while novel expressions do not. Furthermore, the lack of word order-specific priming was consistent across all regions of the sentence in the novel items, suggesting that there is no priming effect at all in novels. This suggests that humans are activating holistic representations of attested expressions. When one order is presented in training, participants activate that exact holistic representation of that word order. So in the mismatch condition, the order they see at test is a different representation than the representation they activated during training, so there is no such priming benefit. In contrast, people have no such holistic representations of the novel items. Instead, during training people might be simply activating representations of the individual words in the item that provides a general boost at testing, but not a word order-specific effect. A crucial prediction this interpretation makes is that as representations of multiword expressions become more strong (i.e., the expressions are more frequent in the language), the word order-specific priming benefit should become stronger. Additional analyses were suggestive that this might be the case: as overall trigram frequency of the training order increases, the difference between the mismatch and match conditions increased (although this was just a marginal interaction). However, it's important to note that this was just a post-hoc analysis and is far from conclusive. We need to do follow-up studies that



(a) (b)

Figure 11: Training period results. (a) shows that there was no ordering preference in items during their first exposure in training for novel or attested items. (b) shows that participants significantly speed up over the course of the training trials, but that there is no difference between preferred and dispreferred orders at any point in training.

specifically test expressions of a wide variety of frequencies.

This interpretation is supported by our other main result, namely that novel expressions show a robust overall training benefit in both of the training conditions as compared to the untrained condition, a benefit that was not present in attested expressions. We interpret this as a lack of individual word priming in attested expressions. If attested expressions are represented holistically, then priming should activate those holistic representations, and not necessarily the representations of individual words, which is consistent with out results here. However, there is a concern is that this effect might arise solely from the fact that the mean unigram frequency of the words within our attested expressions were higher than in our novel expressions. Low-frequency words show larger priming effects (Scarborough et al., 1977), so it is possible that the priming effects are simply smaller in the attested expressions. However, that would not explain why there would be no individual word priming effect at all in the attested expressions. Furthermore, additional analyses revealed that unigram frequency influenced residual reading times more in the novel than attested items. However, this was just a post-hoc analysis, and further investigation is needed to confirm this hypothesis.

Another important factor to consider when interpreting these results is the difference in semantic relations within the novel and attested items. Our attested items mostly contained words that are more semantically related than in the novel items (e.g., “read and write" vs. “newts and litter"). It is possible that this somehow plays into how easy it is for people to recall exact form of an expression. Maybe participants use a strategy where they pay more attention to the individual words within a multiword expression when they are less semantically related, which would predict the overall training benefit we observed in our novel expression data. However, we do not have enough

items in each bin of semantic relatedness to settle this concern here. In particular, it is very difficult to find high-frequency phrases that contain semantically unrelated words, since people tend to talk about things that are semantically related.

Finally, it is important to consider how the timescale is important to interpreting these results. At first it seems counterintuitive that novel expressions would not show word order-specific priming. After all, attested expressions were novel at one point, and over time people get experience with them, leading to the preferred >dispreferred expectation that we observe. Therefore, on some level, people must be storing information about the novel binomials' word order, since we are adding to their direct experience with the item during training. Perhaps we did not detect this because we did not expose people to the items enough times (each trained item was presented 3 times in training and 1 time in testing, for a total of 4 exposures in that participant's lifetime). However, it is also possible that we did not observe a word order-specific priming effect because we are testing on a very short timescale. It is possible that people were learning the frequency of novel binomials across the experiment, but had not yet "added" that information to their long-term memory to act as a prior.

Interestingly, we did not replicate the findings of Siyanova-Chanturia et al. (2011) and Morgan and Levy (submitted) with respect to the reading time advantage for the preferred order in untrained expressions in the experiment. Although there was a trend for the novel binomials to be read faster in the preferred than dispreferred order if they were untrained, there was no such difference in the attested items. However, we did find an effect of relative frequency of test order, such that more relatively frequent orders were read faster. So although we did not find an overall difference between preferred (relative frequency >0.5) and dispreferred (relative frequency <0.5) group means, this correlation is suggestive that we were able to detect a small difference in relative frequency with reading times. One possibility for why we did not find a difference in the group means is that people are updating their expectations for word orders. By the time participants reached the testing period, they had already had 72 exposures to binomials from 24 different binomial items, and half of these were presented in the preferred and half of them were presented in the dispreferred. Since people should expect to see many more preferred orders than dispreferred orders naturally, then perhaps they picked up on the unnatural distribution within the experiment. Then perhaps people updated their priors to have a lower expectation for encountering preferred orders. Then when a new item is presented, people are not using their prior expectations as heavily during processing, leading to a smaller relative frequency effect. One way to counteract this effect in future experiments would be to have many more items that appear in the preferred order in training, to mimic the natural distribution of these items in the language. Conversely, we could show all of the training items in the dispreferred order and then see if people read new items faster in the dispreferred order in testing.

In order to investigate this further, we analyzed the first exposure to each item in the training phase. However, here we also found no difference between preferred and dispreferred orders in either the novel or attested items. However, it is possible that we did not have enough data in the training trials to detect an ordering preference in processing speed; indeed, remember that two-thirds of the training trials were presented as full sentences, and we did not analyze the full-sentence data, so we did not have a lot of training trial data. Furthermore, the training trials were

fully randomized, so it is possible that some items were seen for the first time after other items had been seen more than once, which might somehow affect how people are biased when they encounter new items. In order to control for this, in future experiments we should make sure that all first exposures occur before second exposures, etc.

5 Future Work

There are a few different follow-up experiments that would be informative based on concerns we raise in the previous section.

Firstly, in the experiment reported here we did not have a large range of frequencies in our attested items. It is possible that we do not have enough data on middle-frequency attested expressions to fully understand the role of frequency in the training effects we found. Since our interpretation of our data crucially predicts that word order-specific training effects should become stronger with increasing trigram frequency, it would be useful to do an experiment which explicitly incorporates this into the design. Instead, we want to do an experiment much like the one presented here, except with only attested items and across a large range of frequencies.

Another concern with our items in this experiment is that the unigram frequencies are not controlled across conditions. In order to more thoroughly investigate how unigram frequency affects processing of high frequency and low frequency phrases, we should control unigram frequency while varying phrase frequency. Adjective-noun pairs would be a great test case here; it is easy to construct expressions across a wide range of frequencies while controlling for unigram frequency. In fact, there is already some evidence that unigram frequency behaves differently in high frequency and low frequency phrases in adjective-noun pairs, at least in memory tasks (Jacobs, Dell, Benjamin, & Bannard, under revision). One interesting way to investigate how much unigram frequency plays a role in processing of phrases would be to do a more explicit priming experiment. For example, subjects could read a sentence containing the phrase, and then after the trial make a lexical decision on a target word that is semantically related to one of the words in the phrase. If the processing of high-frequency phrases is less influenced by the individual words within the phrase, then there should be a smaller priming effect in high-frequency phrases than in low-frequency phrases.

It seems obvious that people must learn ordering preferences eventually for novel expressions given enough experience. Perhaps our experiment is not designed in such a way as to be conducive for testing what people have learned, and instead we are just testing what people have adapted to in a short period of time (some sort of intermediate timescale between priming and learning). At present what our results suggest is that novel expressions are not able to be primed for exact form within the course of about half an hour to an hour of experience (the average amount of time it took our participants to complete the study). However, we might predict the opposite outcome if we instead had training and testing take place on separate days. In the attested expressions, during the experiment people are biased to expect one order over the other based on their training. However, as time goes on after the experiment, people's expectations should decay back to their priors, and we should observe no training effect if we test on the next day. However, in the novel binomials, there are no prior expectations from direct experience. So people have the same expectations in

all conditions during the experiment. However, since during the experiment we have given people direct experience with the novel binomials, they should start forming priors about them. At a longer time delay, these priors might become incorporated and people should have expectations based on the order they were trained in during the previous time.

6 Conclusion

We show that highly frequent attested multiword expressions show a training benefit for exact form, but not an overall training benefit, while novel expressions show an overall training benefit, but no benefit for exact form. Furthermore, we show that mean unigram frequency of the individual words within a binomial only affects processing of novel expressions, not attested expressions. We also show that the overall trigram frequency of the training order within attested expressions affects the training benefit for exact form. Together, these results suggest that novel expressions are processed fully compositionally, while attested expressions can be represented and processed as holistic, chunks, in line with usage-based theories.

References

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

Bannard, C., & 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), 241–8.

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

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). *lme4: Linear mixed-effects models using eigen and s4.* Retrieved from (ArXiv e-print; submitted to *Journal of Statistical Software*)

Benor, S., & Levy, R. (2006). The chicken or the egg? a probabilistic analysis of english binomials. *Language*, 233–278.

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

Cohen Priva, U. (2008). Using information content to predict phone deletion. In *Proceedings of the 27th west coast conference on formal linguistics* (pp. 90–98).

Cooper, W., & Ross, J. (1975). World order. In R. Grossman, L. San, & T. Vance (Eds.), *Papers from the parasession on functionalism* (p. 63-111). Chicago: Chicago Linguistic Society.

Fine, A. B., & Jaeger, T. F. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, *37*(3), 578–591.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS One*, *8*(10), 1-18.

Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 748–775.

Hadfield, J. D. (2010). Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1–22. Retrieved from

Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (under revision). *Part and whole linguistic experience affect recognition memory for multiword sequences.*

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, *61*(1), 23–62.

Janssen, N., & Barber, H. a. (2012). Phrase frequency effects in language production. *PloS One*, *7*(3), 1-11.

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (p. 229-254). Amsterdam: John Benjamins.

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

Mollin, S. (2013). Pathways of Change in the Diachronic Development of Binomial Reversibility in Late Modern American English. *Journal of English Linguistics*, *41*(2), 168–203.

Morgan, E., & Levy, R. (submitted). *Abstract knowledge versus direct experience in processing of binomial expressions.*

Naigles, L. R., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? effects of input frequency and structure on children's early verb use. *Journal of Child Language*, *25*(01), 95–120.

Pinker, S. (1998). Words and rules. *Lingua*, *106*(1), 219–242.

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

Rayner, K., & 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.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human perception and performance*, *3*(1), 1–17.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. (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.

A Appendix: Experimental Items

Novel Binomial Expressions			
Word 1	Word 2	Word 1	Word 2
first	ninety-eighth	chauffeurs	stewardesses
discontent	tearfulness	teal	annoying
masculine	undignified	mad	fuming
actresses	lumberjacks	loons	horses
hesitate	readjust	flowers	zinnias
robots	bicycles	coroners	senators
quails	felines	dreary	disheveled
vegetables	kale	tigers	puppies
superintendent	groundskeeper	purple	bitter
wildfires	campfires	currant	pomegranate
candy	bacteria	lard	gelatin
allergic	unaccustomed	vowels	vocabulary
rats	sharks	vacations	therapy
nurses	patriarchs	beautiful	stinky
happily	rudely	kittens	blankets
lanky	lankier	chickens	fences
abashed	sorry	provides	donates
marooned	missing	ducks	peanuts
chanting	enchanting	boards	two-by-fours
murdered	deposed	catholics	non-catholics
llamas	cherries	farms	hayfields
determined	forgettable	linguistics	psychiatry
bishops	seamstresses	stakes	barriers
newts	litter	stress	nagging

Table 6: Novel expressions used in the experiment.

Attested Binomial Expressions			
Word 1	Word 2	Word 1	Word 2
alive	well	black	white
bride	groom	brothers	sisters
intents	purposes	backwards	forwards
king	queen	mind	body
crime	punishment	trial	error
mix	match	supply	demand
sweet	sour	past	present
stocks	shares	east	west
arts	sciences	family	friends
heart	soul	men	women
mother	child	radio	television
pain	suffering	flora	fauna
safe	sound	read	write
buy	sell	gold	silver
church	state	boys	girls
war	peace	goods	services
newspapers	magazines	time	place
profit	loss	oil	gas
right	wrong	temperature	pressure
food	drink	hands	arms
husband	wife	vitamins	minerals
name	address	religion	politics
research	development	face	body
knife	fork	maps	charts

Table 7: Attested expressions used in the experiment.