Cognitive Perspectives On English Word Order

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Arts in the Graduate School of the Ohio State University

By

Michael Xavier Collins, B.S.
Graduate Program in Linguistics

Ohio State University

2012

Thesis Committee:

William Schuler, Advisor

Peter Culicover
Copyright by

Michael Xavier Collins

2012
Abstract

Certain English constructions permit two semantically equivalent word orders. Corpus studies of these constructions reveal that one variant usually occurs much more frequently than the other. What cognitive process causes speakers to gravitate towards some orders while neglecting others? This thesis reviews three prominent psycholinguistic models of the relationship between word order and processing: Maximum Per Word Surprisal (MPWS) [Hale 2001], Uniform Information Density (UID) [Levy & Jaeger 2007, inter alia], and Dependency Length Minimization (DLM) [Gildea & Temperley 2009]. It then compares the predictions of each of these theories to actual human word order preferences. Quantifying the strength of each psycholinguistic factor yields many interesting insights into human cognition.
Dedication

For Alison
Vita

June 2006......................................................... Raleigh Latin High School

December 2009................................................. B.S. Mathematics, North Carolina State University

2011 to present................................................. Graduate Teaching Associate,
Department of Linguistics, The Ohio State University

Fields of Study

Major Field: Linguistics

Specialities: syntax, psycholinguistics
Table of Contents

Abstract.............................................................................................................................................ii

Dedication.........................................................................................................................................iii

Vita......................................................................................................................................................iv

List of Tables....................................................................................................................................vi

List of Figures....................................................................................................................................vii

Chapter 1: Introduction....................................................................................................1

Chapter 2: Maximum Per Word Surprisal..........................................................6

Chapter 3: Uniform Information Density...............................................................................13

Chapter 4: Dependency Length Minimization.................................................................21

Chapter 5: Experimental Design..................................................................................27

Chapter 6: Statistical Model.........................................................................................33

Chapter 7: Results & Discussion..............................................................................37

Chapter 8: Summary...........................................................................................................51

Chapter 9: Conclusion.........................................................................................................53

References.................................................................................................................................54

Appendix: Mechanical Turk Survey Questions.................................................................58

v
List of Tables

Table 1. Some word order alternations listed in Wasow 2002.................................21
Table 2. Subjects’ word order preferences.................................................................38
Table 3. Base regression model................................................................................39
Table 4. Regression on probability............................................................................40
Table 5. Regression on DLM....................................................................................40
Table 6. Regression on UIDEv..................................................................................41
Table 7. Regression on variance................................................................................42
Table 8. Regression on MPWS..................................................................................42
Table 9. A two factor model of UIDEv and probability.............................................43
Table 10. Correlation of fixed effects.........................................................................44
Table 11. A two factor model of UIDEv and variance..............................................45
Table 12. Correlation of fixed effects.........................................................................45
Table 13. A two factor model of UIDEv and MPWS.................................................47
Table 14. Correlation of fixed effects.........................................................................47
Table 15. A two factor model of DLM and UIDEv...................................................49
Table 16. Correlation of fixed effects.........................................................................49
List of Figures

Figure 1. The noisy channel model of communication...........................................6

Figure 2. Sample PCFG derivation..............................................................................9

Figure 3. A line graph of S1 and S2’s information profiles.......................................11

Figure 4. A line graph of S3 and S4’s information profiles.......................................16

Figure 5. A line graph of S5 and S6’s information profiles.......................................20

Figure 6. The recursive dependency structure of a sentence...................................23

Figure 7. Using DLM to model heaviness effects....................................................24

Figure 8. Experimental design..................................................................................30

Figure 9. The predictor values are balanced across items (a) and (b) when the UIDev metric is used to measure an information profile’s uniformity........................................35

Figure 10. The predictor values are also balanced across items (a) and (b) when variance is used to measure an information profile’s uniformity.........................................36
Chapter 1: Introduction

Natural languages often have many ways of expressing a single idea. These ways are called *alternations*. Some alternations, such as the speaker’s decision to use the active or passive voice, change the syntactic structure of the sentence and assign constituents different semantic roles.

(1) a. David loves Amy.
   b. Amy is loved by David.

Others insert or delete semantically vacuous function words such as “that.”

(2) a. It seems you don’t understand.
   b. It seems [that] you don’t understand.

Still others, in addition to inserting semantically vacuous elements such as “dummy-it,” rearrange constituents so that they appear in a different linear order.

(3) a. [That Patrick is a starfish] [is obvious].
   b. It [is obvious] [that Patrick is a starfish].

Finally, the number of words in the sentence may remain constant, with only their relative
arrangement changing.

(4) a. I talked [about Crispin] [to Jimmy].
   b. I talked [to Jimmy] [about Crispin].

Interestingly, although many options are available, corpus searches show that language users tend to gravitate towards one variant and neglect the others. What motivates language users to choose one alternation over another?

The context in which a sentence is uttered certainly influences speakers’ decisions about which alternation to use. For example, when conversing with a young child, one would probably not use an alternation such as, “that you keep whining and screaming annoys me.” Such a construction is ungainly and difficult even for adults to understand. Presumably, “it annoys me that you keep whining and screaming” would be preferred. However, a different situation may incline one to make use of the complex subject. For example, a lawyer or a politician may favor the complex subject precisely because it is more difficult to understand. The use of this construction may carry with it certain social implications of high intelligence or political dexterity. In this case, using the complex subject may actually be preferred.

Alternations also serve as a means to modify and control discourse and topic structure. Consider the following utterance.
“I saw a cat wandering the streets. It was so cute. This little kitten made friends with everyone it met. Eventually, it made its way to my neighbor’s door...”

The speaker now has a choice about how to phrase the next sentence: does he/she use the active or the passive alternation?

(a) My neighbor took it inside.
(b) It was taken inside by my neighbor.

Each variant carries with it certain consequences for discourse and topic structure. The active alternation (a) shifts the listener’s attention to the new subject “my neighbor.” If the speaker decides to use this form, the listener would expect the focus of the discourse to center now on the neighbor. In contrast, the passive alternation (b) keeps the listener’s focus on the cat. Choosing to use form (b) is a signal to the listener that the cat occupies a place of primacy in this discourse; it relegates the neighbor to a subordinate position. Which alternation is appropriate depends very much on which discourse referent the speaker wishes to emphasize.

This thesis focuses on one particular family of alternations: variable word order. Variable word order is defined as two semantically-equivalent positions for syntactic elements. For example:
(5) a. I looked up the number.
   b. I looked the number up.

(6) a. He is often at the office.
   b. He often is at the office.

Gries 2003 provides a summary of some linguistic factors known to affect word order decisions. He notes contributions from almost every level of linguistic description: phonetic (stress patterns), morphosyntactic (definiteness of NP’s), semantic (focus), and discourse functionality (saliency). The majority of the literature on word order concerns itself with identifying and modeling the structural and pragmatic factors that affect word order preferences, with the end goal of forming a model which predicts which word order will be used in any given situation. This not the focus here. Instead, the goal of this thesis is to control for these factors as far as possible and study how word order relates to general cognitive processing. The central assumption underlying this investigation is that, when the above and other similar factors are controlled for, speakers will favor the word order that minimizes processing difficulty. Studying speakers’ “default” word order preferences reveals how the mind prefers verbal information to be presented. Such data is of interest to both linguists and cognitive scientists.

This thesis reviews three prominent psycholinguistic models of the relationship between word order and processing: Maximum Per Word Surprisal (MPWS) [Hale 2001], Uniform Information Density (UID) [Levy & Jaeger 2007, inter alia], and Dependency Length
Minimization (DLM) [Gildea & Temperley 2009]. It then compares the predictions of each of these theories to actual human word order preferences. Quantifying the strength of each psycholinguistic factor yields many interesting insights into human cognition.
Chapter 2: Maximum Per Word Surprisal

The concepts of Information Theory (Shannon 1948) offer precise mathematical tools with which to investigate linguistic problems. Information Theory models the communicative process in terms of three basic components: the channel encoder, the channel, and the channel decoder. The channel encoder is the source of the message (m). It sends the coded signal (c) through the channel, or communication medium. The channel, in turn, may be noisy (introducing distortion) or clean (without distortion). Finally, the channel decoder attempts to decipher the original message (m) sent by the channel encoder.

Figure 1. The noisy channel model of communication\(^1\)

For example, suppose John and Mary are at a cocktail party and John is trying to ask Mary out for a date. John is the channel encoder. His message is sent through the channel (the air), where it encounters noise (the other people at the party who are talking). Mary, the channel decoder, may not be able to decipher John’s original message, due to the distorting effects of the noise. The channel capacity is defined as the highest upper bound on the

\(^1\) from http://errorcorrectingcodes.wordpress.com/2010/01/25/notes-3-stochastic-channels-and-noisy-coding-theorem/)
amount of information that can be reliably transmitted over the channel. This concept captures the fact that some conversational media, like a quiet college classroom, are better for transmitting messages than others, such as a boisterous bar.

The information or surprisal of an item (the two terms are synonymous) is defined as the negative logarithm of its probability. This definition reflects our intuition that improbable or unpredictable items should carry a large amount of information while likely or predictable items should not.

Information(x) = -log(pr(x)) = log(1/pr(x))

Information is a function of probability. Therefore, the information value a model assigns a particular item depends entirely on how the probability of this item is calculated. Different models make different assumptions; this, in turn, changes the probability value assigned to a particular item. In this thesis, the probabilities of all linguistic items are calculated using the parsing model described in detail in Schuler 2009 and Van Schijndel et al. 2012.

The parsing model is based on a probabilistic context free grammar (PCFG) trained on the Wall Street Journal corpus sections 02 - 21. A PCFG is simply a context free grammar (CFG) that assigns each expansion rule a probability of occurrence. For example, consider the toy CFG below.

S → NP VP
NP → Det N
NP → Det Adj N
VP → V NP
Det → {some, a, the}
N → {cat, dog, gerbil}
V → {eats, attacks}
Adj → {fat}

We can transform this CFG into a PCFG by associating a probability value with each rewrite rule. These probabilities are calculated using corpus counts in the training set. For example, suppose our training corpus contains 100 NP’s. Out of these 100, 90 are composed of Det N, and the other ten are composed of Det Adj N. A PCFG uses this data to assign a probabilistic weight of expansion to NP.

[.9] NP → Det N
[.1] NP → Det Adj N

Continuing with our toy example, we assign hypothetical probabilities to our CFG to transform it into a PCFG.

[1] S → NP VP
[.9] NP → Det N
[.1] NP → Det Adj N
[1] VP → V NP
[.3] Det → {some}
[.4] Det → {a}
[.3] Det → {the}
[.3] N → {cat}
[.5] N → {dog}
[.2] N → {gerbil}
[.5] V → {eats}
[.5] V → {attacks}
[1] Adj → {fat}

The total probability of a sentence $S$ is defined as the product of the probabilities of the expansion rules used to generate it. For example, the probability of this toy PCFG model generating the sentence, “the fat gerbil eats the dog” is calculated as:

$$1 \times .1 \times .3 \times .1 \times .2 \times .1 \times .5 \times .9 \times .3 \times .5 = 0.000405.$$
PCFGs are powerful cognitive models of a language’s lexicon and syntactic structure (Jurafsky 1996). However, there is the intuitive feeling that the probability of a sentence depends not only on the syntactic rules used to generate it, but also on the semantics of the resulting construction. It seems that sentences with extremely odd semantics ought to be assigned a lower probability than those with more conventional meanings. Consider the following pair of sentences.

(7) a. The bird flew into the *bush.*  
   b. The bird flew into the *disposal.*

Both of these sentences have the same syntactic structure. Both “bush” and “disposal” are nouns; the oddity arises from choosing to use the noun “disposal” in this particular context. Intuitively, option (b) should be dispreferred because it is semantically odd. Fortunately, these legitimate concerns about semantic complications have little bearing on the present line of inquiry. The subject of this study is variable word order, which is defined as two or more *semantically equivalent* positions for syntactic elements. Since both orders have the same meaning, there is no danger of one variant being preferred simply because it expresses a meaning that it used more often. The nature of the research topic, which is concerned with studying variations in how a single common idea is expressed, effectively factors out the role of semantics.

Hale 2001 observes a positive correlation between an word’s surprisal value and its comprehension difficulty. If we assume that language users generally wish to minimize the
comprehension difficulty of their utterances, we may formulate the following hypothesis.

**Hypothesis 1:** Language users choose the word order that minimizes the sentence’s maximum per word surprisal (MPWS).

In other words, language users tend to avoid “information bottlenecks” that dramatically slow down comprehension. The word order that produces the smallest obstacle to comprehension should be preferred. Consider the following two hypothetical information profiles.

\[
\begin{align*}
S_1 &= .1 \quad .2 \quad .1 \quad .5 \quad .3 \quad .1 \\
S_2 &= .1 \quad .2 \quad .1 \quad .7 \quad .3 \quad .1
\end{align*}
\]

![Figure 3. A line graph of S1 and S2’s information profiles](image)
According to Hypothesis 1, S1 should be preferred since its MPWS (.5) is lower than S2’s (.7). Language users strive to avoid abrupt peaks in information and order their words accordingly.
Chapter 3: Uniform Information Density

3.1. Classical Formulation

Levy & Jaeger 2007 and Frank & Jaeger 2008 assert that language users prefer structures that approach Uniform Information Density (UID). UID claims that, ideally, each component of the utterance should carry roughly the same amount of information. UID holds a great deal of cognitive appeal by striking a balance between two competing communicative demands: efficiency and reliability. Speakers want to transmit their message as quickly as possible (efficiency), but, at the same time, they want to make the speech signal robust enough to withstand possible interference from noise (reliability).

By transmitting the signal at a constant rate at or near the channel capacity, speakers ensure that their message is being communicated as rapidly as possible. At the same time, since every component of the utterance carries roughly the same amount of information, the signal is robust enough to withstand degradation from noise. If one component of the utterance is lost, it is still possible to piece the signal back together from context. Such a repair process would be much more difficult to achieve if UID did not hold; if the component carrying most of the information were lost, it might be nearly impossible to restore the intended message. UID offers a compromise between these two opposing tendencies and is thus an ideal solution to the problem of rapid, error-free communication (cf. Shannon 1948).
Previous research has demonstrated that speakers strive for UID in many levels of linguistic structure. Aylett & Turk 2004 gives a functionalist account of phonetic/prosodic patterns in English speech. They find that speakers are more likely to delete phonetic segments with low information content. Intuitively, this is because these segments are not contributing much to the structure of the message; their contents can be reasonably inferred from its surrounding environment. Information Theory, which assumes that language users strive for efficiency in communication, gives a satisfying cognitive justification of this behavior.

Levy & Jaeger 2007 and Frank & Jaeger 2008 extend the ideas of Aylett & Turk 2004 to language’s morphosyntactic level. Levy & Jaeger 2007 shows that speakers are more likely to include the function word “that” in object relative clauses (e.g. the boy I hit vs. the boy that I hit) when the relative clause (e.g. I hit) has a high information content. The inclusion of this extra word helps spread out the information more evenly across all the words in the sentence. The decision about whether to use or omit the semantically vacuous function word “that” is determined in part by the following phrase’s information content. Frank & Jaeger 2008 finds similar results; speakers are more likely to use a non-contracted verbal form (e.g. “you are,” as opposed to “you’re”) when the following items are high in information. Again, the same principle applies; the inclusion of extra function words acts as a way to modulate the information density across the utterance.

Finally, Maurits et al. 2010 uses UID as a functionalist explanation for global word order typologies. They explain the extreme prevalence of SOV and SVO word orders and the
relative absence of the other four possible typologies in terms of the relative distribution of information across each of the three primitive components S, V, and O. The dominant SOV and SVO orders are the ones that spread the total information of the sentence most evenly over its parts. They conclude that UID had some hand in shaping the global word order typologies we observe today.

With the exception of Maurits et al. 2010, the above works have applied the concept of UID to situations in which the speaker has a choice about how to construct his or her utterance. UID’s ability to predict language users’ decisions based on constructions’ information profiles suggests that this theory might be helpful in studying other instances of linguistic choice, such as variable word order. With this in mind, we formulate the following hypothesis.

**Hypothesis 2:** Language users choose the word order that results in the most uniform distribution of information across all the words in the sentence.

This statement raises a methodological concern: how does one quantitatively measure uniformity? In Aylett & Turk 2004, Levy & Jaeger 2007, and Frank & Jaeger 2008, this question had a straightforward answer; speakers either deleted or inserted elements in the speech stream. This can be described as a structural change from I/n (reduced form) to I/(n+1) [full form], where I = total amount of information, and n = number of utterance units. It is easy to see that the I/(n+1) form results in a decrease of average information density as compared to I/n.
Unfortunately, this type of logic cannot be applied straightforwardly to variable word order. Here, the number of utterance units containing information remains constant; it is only their relative arrangement that changes. In other words, we are comparing sentences of the form I/n to other sentences of the form I/n. How is it clear which one spreads out information more evenly across utterance units?

Suppose two sentence S3 and S4 have the following information profiles.

\[
S3 = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.5, 0.4, 0.3, 0.2, 0.1\}
\]
\[
S4 = \{0.6, 0.1, 0.3, 0.5, 0.2, 0.4, 0.1, 0.4, 0.5, 0.2, 0.3\}
\]

![Graph of S3 and S4's information profiles](image)

**Figure 4.** A line graph of S3 and S4’s information profiles

UID states that each component of the utterance should carry roughly the same amount of
information. Graphically, a uniform information profile is represented as a flat line near the channel capacity value. As such, it is unclear which of the above sentences has the more uniform information profile. The S3 line is characterized by a gradual ascent and descent to and from the channel capacity (.6). S4, in contrast, contains many sharp peaks and valleys. Neither resembles a flat line. Which of these two sentences comes closer to the definition of uniformity?

The theory of uniform information density claims that, ideally, language users should strive to transmit at a constant rate near the channel capacity. Spending large amounts of transmission time significantly below the channel capacity is undesirable. This concept can be captured quantitatively by taking the variance of a sentence’s information profile.

Variance is a statistic designed to measure the spread of a population’s individual data points (in this case, the information content of each individual word) from the population’s mean. Accordingly, information profiles with high variance (data points far away from the mean) are not “UID like,” while those with low variance (data points closely clustered around the mean) are. The variance of S3 is equal to the variance of S4; this models the fact that both spend an equal amount of time below the channel capacity.

3.2. An Alternative Approach

The classic formulation of UID speaks only to the ideal case. It does not offer any guidelines for how language users should structure their information profiles in the event that a perfectly flat line near the channel capacity cannot be produced. Accordingly, this
thesis proposes a new addition to the theory of uniform information density: if
deviations from the ideal transmission rate must occur, then they should be introduced in a
gradual fashion. Language users should not abruptly switch from transmitting very high
amounts of information to very low amounts; dense information areas and sparse
information areas should be “bridged” (cf. Levy & Jaeger 2007 and Frank & Jaeger 2008).
In other words, the difference between each successive point in the information profile is
crucially important. This claim is intuitively appealing. If we recall that information
content directly correlates with comprehension difficulty (Hale 2001), it is easy to believe
that listeners would desire comprehension difficulty to increase or decrease in a slow and
predictable way. Rapidly shifting from extremely taxing sections to those requiring hardly
any effort at all would be disorienting and result in a decrease in transmission accuracy.

To model this idea quantitatively, we propose the following novel metric.

Let w1, w2, …, w(n-1), w(n) be the words in some sentence S of length n. Let i1, i2, …,
i(n-1), i(n) be the corresponding information content of words w1, w2, …, w(n-1), w(n),
respectively. Then S’s UID Deviation is defined as follows (cf. Maurits et al. 2010’s UID
deviation score):

\[ \text{UIDev}(S) = \frac{|i_2 - i_1| + |i_3 - i_2| + |i_4 - i_3| + \ldots + |i(n) - i(n-1)|}{n} \]

Smaller scores are more “UID like,” while bigger scores are less UID like. Intuitively, this
is because, in a uniform profile, each component carries roughly the same amount of
information. That is, $i_1 \equiv i_2 \equiv i_3 \equiv i_4 \equiv \ldots \equiv i(n)$. Therefore, the sum of the difference between successive points in the profile will be close to zero. In a profile which is not uniform, the difference between successive points is large. Applying the above formula to S3 and S4, we obtain $\text{UIDev(S3)} = (.1 + .1 + .1 + .1 + .1 + .1 + .1 + .1 + .1 + .1)/11 = .10/11 = .00909$ and $\text{UIDev(S4)} = (.5 + .2 + .2 + .3 + .2 + .3 + .3 + .1 + .3 + .1)/11 = .25/11 = .0227$. Therefore, according to this novel conception of uniformity, S3 best approximates the ideal of uniform information density.

UIDev and variance model different aspects of the theory of UID. Comparing the two is very interesting from a theoretical standpoint: which measure of uniformity is more important to language users?

3.3. A Comparison of UID and MPWS

Although UID and MPWS stem from the same intellectual tradition and are similar in some respects, it is important to recognize that they make different empirical predictions. MPWS states that language users should arrange their words to avoid high information bottlenecks; sharp peaks in the information profile are cognitively undesirable. In other words, MPWS sets an upper bound on the possible information content of each word in the utterance. In contrast, UID asserts that each word of the utterance should carry roughly the same amount of information. It follows MPWS in stating that sharp peaks of information are undesirable. However, it differs by claiming that information valleys should be avoided as well. It imposes both an upper and a lower bound on the information value of each individual point.
in the profile; MPWS imposes only an upper bound. Consider the following hypothetical information profiles.

\[ S5 = .3 \quad .2 \quad .1 \quad .1 \quad .2 \quad .3 \]
\[ S6 = .3 \quad .3 \quad .3 \quad .3 \quad .3 \quad .3 \]

![Graph of S5 and S6's information profiles](image)

**Figure 5.** A line graph of S5 and S6’s information profiles

UID claims that S6 should be preferred since each component of the utterance carries the same amount of information. MPWS, however, makes no distinction between S5 and S6 since the maximum per word surprisal of both sentences is the same (.3). Both should be equally preferred. By explicitly comparing the predictions of UID and MPWS, we investigate whether this lower bound of information is of importance to language users. The answer to this question is also of great theoretical interest.
Chapter 4: Dependency Length Minimization

Wasow 2002 lists several examples of variable word order in English.

<table>
<thead>
<tr>
<th>Alternation</th>
<th>Order 1</th>
<th>Order 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heavy NP Shift</td>
<td>I gave the book which had previously been owned by Severus Snape to Ron.</td>
<td>I gave to Ron the book which had previously been owned by Severus Snape.</td>
</tr>
<tr>
<td>Extraposition from NP</td>
<td>The gun which I had cleaned went off.</td>
<td>The gun went off which I had cleaned.</td>
</tr>
<tr>
<td>It-extraposition</td>
<td>That you are an idiot bothers me.</td>
<td>It bothers me that you are an idiot.</td>
</tr>
<tr>
<td>PP ordering</td>
<td>I talked to many people about the movie.</td>
<td>I talked about the movie to many people.</td>
</tr>
<tr>
<td>Verb Particle</td>
<td>I looked up the number.</td>
<td>I looked the number up.</td>
</tr>
<tr>
<td>Dative alternation</td>
<td>I gave Greg the book.</td>
<td>I gave the book to Greg.</td>
</tr>
</tbody>
</table>

Table 1. Some word order alternations listed in Wasow 2002

Wasow claims that, since the semantic difference between each variant is negligible, our
decision to use one order or the other is motivated primarily by processing considerations. He claims that the Principle of End Weight (PEW) can offer a unified explanation for these alternations. The Principle of End Weight (PEW) simply states that, when given a choice, speakers prefer the “heaviest” constituent (defined in terms of both syntactic and phonological complexity) to occur as close to the end of the sentence as possible. Constituents are usually ordered by increasing heaviness. Wasow 2002 (pg. 1) uses a quote from Steven Pinker’s *The Language Instinct* to illustrate how this principle is manifested in everyday language use. It is reproduced below (formatting mine). Note how the constituents after the verb “document” are arranged in order of increasing syntactic and phonological weight.

“In my laboratory we use it as an easily studied instance of mental grammar, allowing us to document [in great detail] [the psychology of linguistic rules] [from infancy to old age] [in both normal and mentally impaired people,] [in much the same way that biologists focus on the fruit fly *Drosophila* to study the machinery of the genes.]”

There have been various attempts to situate PEW within a general theory of linguistic processing. A particularly prominent model is Gildea & Temperley 2009’s Dependency Length Minimization (DLM), which builds on the work of Gibson 1998. A *dependency* is defined as an asymmetrical syntactic relationship between two words, the head and the dependent. This idea corresponds roughly with the idea of syntactic subcategorization or argument selection. The authors make the assumption that each constituent type is headed by a word of the corresponding type: VP’s by V’s, NP’s by N’s, etc. For example, in the
NP “the boy,” “boy” is the head (being an N) and the determiner “the” is the dependent.

Heads, in turn, depend on other heads, forming a recursive structure, as shown below. The auxiliary verb, if the sentence has one, is taken to be the head of the sentence as a whole. Here, and in all subsequent diagrams, arrows point from heads to dependents.

![Diagram of recursive dependency structure]

Figure 6. The recursive dependency structure of a sentence (from Gildea & Temperley 2009, pg. 287)

*Length* is defined as the number of words (graphemes) separating the head from the dependent. Adjacent words are assumed to have a dependency length of one. For example, in the phrase “a nonexecutive director,” the dependency length between the head “director” and the dependent “a” is two.

In essence, DLM adopts the commonsense notion that related words in a sentence ought to appear close together in order to facilitate processing. The theory appeals to working memory constraints as a measure of processing difficulty. Long distance dependencies are considered difficult because they strain working memory. DLM thus provides cognitive underpinnings for PEW’s empirical generalizations.

The following diagram illustrates how DLM can give a convincing cognitive model of
heaviness effects. Obeying PEW results in a dramatic reduction of dependency length, and hence, a decreased strain on working memory. The order that observes PEW (top arrangement) results in a total dependency length of eight; the order that violates it (bottom arrangement) yields a total dependency length of 16. The drastic increase in dependency length mirrors the drastic increase in processing difficulty.

Figure 7. Using DLM to model heaviness effects (from Gildea & Temperley 2009, pg. 288)

With this in mind, we formulate the following hypothesis.

**Hypothesis 3:** Language users choose the word order that minimizes the total dependency lengths of all components of the sentence.

This hypothesis holds a great deal of intuitive appeal. For example, consider the following
alternation.

(8) a. I study in the library.
    b. In the library, I study.

Native English speakers display a strong preference for order (a). Order (b), while acceptable, seems strange and awkward. At issue is the distance between the head “study” and the dependent “in.” In order (a) these two words have a total dependency length of one, while in order (b) their length increases to four. DLM thus gives our intuitions a mathematical justification.

However, DLM has limitations. It is clear that heaviness has an effect on verb-particle ordering preferences in examples such as “I looked up [the number that was in the very back of the yellow pages]” (cf. ?I looked [the number that was in the very back of the yellow pages] up). However, things become murkier when considering the ordering of “lighter” constituents. For example, “I looked up [the number]” and “I looked [the number] up” are both felicitous. Searches in the Corpus of Contemporary American English (COCA) [Davies 2008] confirm that both options are used frequently; however, the [V Particle Det N] variant occurs almost twice as often as [V Det N Particle]. The two variants are almost identical in terms of dependency length (one has length four, the other five). DLM predicts that they should incur approximately the same processing cost; it is unable to explain the clear preference for [V Particle Det N].

25
Adverb placement offers another challenge for DLM. Consider the following alternation:

(9) a. I usually am in the house.
   b. I am usually in the house.

Both orders have exactly the same dependency length of three. However, COCA searches reveal that order (b) is used ~27x more often than order (a). DLM cannot explain this fact. A similar pattern arises in the placement of adverbs within complex verb phrases:

(10) a. Goblins often have visited the ruins.
   b. Goblins have often visited the ruins.

COCA searches reveal that order (b) occurs ~2x more often than order (a). Both orders have dependency length of three; DLM is thus powerless to explain why one should be so clearly preferred.

The cases above highlight the fact that DLM is a discrete metric; because it measures processing difficulty in terms of integer counts of words separating dependents from heads, it can provide no insight into cases in which the dependency lengths of the competing word order variants are equal. Furthermore, DLM is essentially a generalization of PEW; as such, it is designed to model heaviness effects. Although this has been the traditional focus of research on word order, heaviness forms only a part, not the whole, of the empirical facts. DLM’s ability to model only a fraction of word order data is some cause for concern.
Chapter 5: Experimental Design

MPWS, UID, and DLM are all cognitively motivated models of the relationships between word order and processing. From a purely theoretical perspective, it is not immediately clear which provides the most convincing account of speakers’ word order preferences. This thesis seeks to answer this question by comparing each theory’s prediction to actual human preferences in a wide array of variable word order phenomena. Each theory’s respective ability to duplicate human choices constitutes evidence of its cognitive reality.

Amazon Mechanical Turk is an online service that describes itself as “a marketplace for work that requires human intelligence.” Mechanical Turk assumes that there are still tasks that humans can do more quickly and accurately than a computer. Examples include identifying objects in a photo or transcribing audio recordings. Mechanical Turk is of interest to syntacticians because it allows them to move away from purely introspective judgments (which may or may not reflect actual linguistic behavior) towards usage-based data from naive informants. It provides a cheap, easy, and efficient way for syntacticians to test their intuitions about constructions, including those involving variable word order.

Since Mechanical Turk is available to anyone in the world with an internet connection, it is natural to ask if the data obtained from Turkers (Mechanical Turk workers) is as reliable and accurate as that obtained in a traditional laboratory setting. In a series of tests comparing the results from Mechanical Turk to results from experiments using an
undergraduate subject pool, Munro et al. 2010, Schnoebelen & Kuperman 2010, Snow et al. 2008, and Sprouse 2011 conclude that Mechanical Turk provides data that is of comparable quality to standard laboratory experiments, provided that the experimenter is careful to scrub the data set properly. On the whole, Turkers are honest and competent. There will always be outliers and troublemakers, as with any experiment, but methods have been developed for ensuring that they do not contaminate the larger data pool. In this thesis, we follow the data scrubbing procedures outlined in Schnoebelen & Kuperman 2010.

(i) Amazon allows the experimenter to restrict Turkers to a certain geographical location. Since this thesis is concerned only with judgments about English, we restricted all Turkers to United States locations only.

(ii) We asked all Turkers what language they grew up speaking. If it was anything other than English, we did not use that data point.

(iii) We included a short demographics survey at the beginning of the task. If the Turkers did not fill out this section completely, we did not use that data point (indicates lack of interest in the task, akin to subjects in laboratory experiments who fall asleep).

(iv) Amazon records how long it takes a Turker to complete a task. We did not use any data points that fell more than two standard deviations away from the mean time (again, indicates a lack of interest or competence).
We used Mechanical Turk to gauge human preferences on eight semantically equivalent variable word order “templates.” They are given below.

Template 1: (a) NP Adverb Be PP (I usually am in the house)
   (b) NP Be Adverb PP (I am usually in the house)

Template 2: (a) NP Adverb Have V-en NP (Goblins often have visited the ruins)
   (b) NP Have Adverb V-en NP (Goblins have often visited the ruins)

Template 3: (a) NP V Particle Det N (She looked up the number)
   (b) NP V Det N Particle (She looked the number up)

Template 4: (a) NP V Adverb PP (I sleep often on a bench)
   (b) NP V PP Adverb (I sleep on a bench often)

Template 5: (a) Heavy NP In Situ (The young boy gave the beautiful green ring that had
   been in the jewelry store for weeks to the girl)
   (b) Heavy NP Shift (The young boy gave to the girl the beautiful green
   ring that had been in the jewelry store window for weeks)

Template 6: (a) Complex Subject (That the administrator lost the medical reports
   bothered the intern)
(b) It-extraposition (It bothered the intern that the administrator lost the medical reports)

Template 7: (a) PP In Situ (I work in the house)
(b) PP Extraction (In the house, I work)

Template 8 (a) NP V Particle NP [heavy] (She looked up the number that was in the very back of the yellow pages)
(b) NP V NP [heavy] Particle (She looked the number that was in the very back of the yellow pages up)

We created ten token pairs of test sentences for each of these eight templates. Each token pair was judged by ten different Turkers, yielding an even 100 judgments per word order template. Thus, a grand total of 800 workers participated in this task, as shown in the following diagram.

![Experimental design diagram]

Figure 8. Experimental design
A sample question is presented below.

Which sounds more natural to say?

(a) I looked up the number.
(b) I looked the number up.

The token pairs’ order of presentation was balanced; in half of the questions, variant 1 is listed as option (a), while in the other half it is given as option (b). The data scrubbing procedures necessitated the removal of 15 individual points (~1.9% of the total responses) from the data pool. A complete list of the test sentences and Turkers’ responses is given in the appendix.

The underlying assumption of this thesis is that, all else being equal, speakers will choose the word order that minimizes processing difficulty. The experimental methodology, in which sentences are presented in pairs completely devoid of context\(^2\), is designed to approach this “all else being equal” condition. Presenting each token pair in isolation minimizes the effects of the pragmatic and discourse structural variables discussed in Chapter 1. Additionally, the phonetic effects of stress and intonation are controlled for by

\(^{2}\) One could argue that it is impossible for language to exist without some sort of utterance context. In this case, the context is the fact that the sentences are presented as part of a survey. As such, subjects may experience a certain amount of social pressure to choose what they think the correct answer “should be” according to the rules of prescriptive grammar. This factor did not seem to have a significant effect on the experimental results, but it is worth pointing out as a possible subsidiary influence.
presenting the stimuli in written form\(^3\). Limiting the effects of these confounding factors encourages participants to choose the word order that corresponds most closely to the mind’s default or “unmarked” configuration, the true object of this investigation.

We now compare the predictions of each psycholinguistic theory to actual human preferences. Which hypothesis correctly models the empirical facts?

\(^3\) Again, it is possible to argue that prosodic structure is inextricably bound to the grammatical structure of a sentence; even when a language user reads silently, he/she implicitly hears prosody “in the mind’s ear.” As such, the effects of prosody on word order cannot be controlled for entirely; however, the written presentation of stimuli appears to be the best way to minimize its contribution.
Chapter 6: Statistical Model

We used the parser of Van Schijndel et al. 2012 to calculate the MPWS, UIDv, and variance values for each sentence in each token pair. Dependency length was calculated by counting the number of words separating each relevant head from each relevant dependent (see Chapter 4). In addition, we used the parser to calculate the total probability value of each sentence. Such information is important to include in our analysis for the following reasons.

(i) Certain syntactic configurations are less common than others (e.g. complex subjects). Subjects might be inclined to disprefer them simply because they are infrequent or are used only in a very special register, not necessarily because they are difficult to process. Reporting the probability value of each sentence enables our analysis to account for frequency effects.

(ii) Do MPWS, UID, and DLM made significantly different predictions than a naive baseline algorithm for word order decisions: always choose the most likely sequence?

Mixed effects regression models were employed to analyse the influence of MPWS, UID and DLM on the subjects’ word order preferences. These statistical tools provide a powerful means to separate out the different contributions of many independent random variables on a data set (Quene & van der Bergh 2008). All computations were performed
using the lme4 package of the open source statistical analysis program R (R Development Core Team 2008). The general form of the regression model is given below.

\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + E \]

The dependent variable \( Y \) is expressed as the sum of the contributions from the intercept constant \( \alpha \), one or more independent variables \( X_1, X_2, \ldots, X_n \), and error constant \( E \). Each independent variable \( X_i \) is assigned a weight \( \beta_i \) proportional to its contribution.

In our model, the dependent variable \( Y \) is each individual subject’s decision to choose item (b) in the 80 Mechanical Turk survey questions. The independent variables (\( X_1, X_2, \) and \( X_3 \)) represent the effects of MPWS, UID, and DLM on this decision. Recall that the order of presentation of answers in the survey questions is balanced; for each word order template, variation 1 appears as item (a) half of the time and as item (b) the other half. A token-by-token comparison of each of the three independent variables MPWS, information profile uniformity (measured using both UIDev and variance), and DLM in the 80 survey questions reveals that they are distributed very evenly across all items (a) and (b).

The null hypothesis is that none of these three variables has an effect on the decision to choose item (b) in each of the surveys. If there is indeed a preference for item (a) or (b), it is merely the subjects’ preference to choose either the first or second option in a survey. This statistical design thus provides a fair means to assess the relative contributions of MPWS, UID, and DLM on subjects’ word order preferences.
Figure 9. The predictor values are balanced across items (a) and (b) when the UIDev metric is employed to measure an information profile’s uniformity.
Figure 10. The predictor values are also balanced across items (a) and (b) when variance is used to measure an information profile’s uniformity
Chapter 7: Results & Discussion

The results from the Mechanical Turk surveys are displayed in the following table.
<table>
<thead>
<tr>
<th>Template</th>
<th>Order 1</th>
<th>% Chose 1</th>
<th>Order 2</th>
<th>% Chose 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Template 1</td>
<td>NP Adverb Be PP</td>
<td>13</td>
<td>NP Be Adverb PP</td>
<td>87</td>
</tr>
<tr>
<td>Template 2</td>
<td>NP Adverb Have V-en NP</td>
<td>16</td>
<td>NP Have Adverb V-en NP</td>
<td>84</td>
</tr>
<tr>
<td>Template 3</td>
<td>NP V Particle Det N</td>
<td>63</td>
<td>NP V Det N Particle</td>
<td>37</td>
</tr>
<tr>
<td>Template 4</td>
<td>NP V Adv PP</td>
<td>29</td>
<td>NP V PP Adv</td>
<td>71</td>
</tr>
<tr>
<td>Template 5</td>
<td>Heavy NP In Situ</td>
<td>70</td>
<td>Heavy NP Shift</td>
<td>30</td>
</tr>
<tr>
<td>Template 6</td>
<td>Complex Subject</td>
<td>6</td>
<td>It-Extrapolation</td>
<td>94</td>
</tr>
<tr>
<td>Template 7</td>
<td>PP In Situ</td>
<td>100</td>
<td>PP Extraction</td>
<td>0</td>
</tr>
<tr>
<td>Template 8</td>
<td>NP V Particle NP (heavy)</td>
<td>95</td>
<td>NP V NP (heavy) Particle</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 2. Subjects’ word order preferences
We first ran a base regression model to assess the overall fit of the model to the data. The results are displayed below.

|        | Estimate | Error | z value | Pr(|z|) |
|--------|----------|-------|---------|--------|
| Intercept | -0.5431 | 0.3528 | -1.539  | 0.124  |

Table 3. Base regression model

The intercept is interpreted as the predicted value of the dependent variable (whether or not the subjects chose order (b) in each survey question) when all of the independent variables (MPWS, UID, and DLM) are set to zero. The base regression results give further confirmation that the experimental materials are balanced. Subjects chose option (b) only slightly less than half the time, as indicated by the slightly negative value of the intercept.

We now investigate the contribution of each individual psycholinguistic factor on subjects’ preferences. We first analyzed the effects of total sentence probability. The results are displayed below.
|                     | Estimate | Error  | z value | Pr(>|z|) |
|---------------------|----------|--------|---------|----------|
| Intercept           | -.56560  | .29201 | -1.937  | .0528    |
| Probability Coefficient | .24732    | .04034 | 6.130   | 8.77e-10 (p=0) |

Table 4. Regression on probability

The model assigns probability a positive weight. This indicates a positive relationship between the probability of item (b) and the preference for item (b); the higher the probability value, the more likely the subjects were to choose it. This result is highly statistically significant.

We then ran a regression model to calculate the effects of DLM on the subjects’ preferences. The results are displayed below.

|                     | Estimate | Error  | z value | Pr(>|z|) |
|---------------------|----------|--------|---------|----------|
| Intercept           | -.50774  | .32828 | -1.547  | .122     |
| DLM Coefficient     | -.24137  | .05726 | -4.215  | 2.49e-05 (p=0) |

Table 5. Regression on DLM
The model assigns the DLM coefficient a negative value. This indicates an inverse relationship between the dependency length of order (b) and the subjects’ preference for it; the smaller the total dependency length of order (b), the more likely they were to choose it. The result is also highly statistically significant.

We repeated this procedure for UIDev.

|                | Estimate | Error  | z value | Pr(>|z|) |
|----------------|----------|--------|---------|----------|
| Intercept      | -.5398   | .2769  | -1.949  | .0513    |
| UIDev Coefficient | -7.3396  | 1.2237 | -5.998  | 2.00e-9 (p=0) |

Table 6. Regression on UIDev

The model assigns UIDev a negative weight. This indicates an inverse relationship between the UIDev value of item (b) and the preference for item (b); the lower the UIDev score (the more the sentence’s information profile approaches the ideal of uniformity information density) the more likely subjects were to choose it. Once again, the result was highly statistically significant.

We also ran a regression on variance. The results are displayed below.
|          | Estimate | Error | z value | Pr(>|z|) |
|----------|----------|-------|---------|---------|
| Intercept| -0.5184  | 0.3024| -1.714  | 0.0865  |
| Variance Coefficient | -7.0769  | 1.5169| -5.081  | 3.76e-7 (p=0) |

Table 7. Regression on variance

The model assigns variance a negative weight. This indicates an inverse relationship between variance and subject preference for item (b); the higher the variance of the sentence (the less uniform the information profile), the less likely were subjects were to choose it. This result was highly statistically significant.

Finally, we ran a regression on MPWS.

|          | Estimate | Error | z value | Pr(>|z|) |
|----------|----------|-------|---------|---------|
| Intercept| -0.5839  | 0.3373| -1.731  | 0.08346 |
| MPWS Coefficient | -4.8878  | 1.5514| -3.151  | 0.00163 (p=.001) |

Table 8. Regression on MPWS
The model assigns MPWS a negative weight. Once again, this indicates an inverse relationship between MPWS and the preference for item (b); the lower the MPWS, the more likely subjects were to choose it. And yet again, the result was highly statistically significant.

To summarize thus far, regression on each individual factor suggests that each is a highly significant predictor of word order preferences. However, it would be interesting to discover whether any one of these individual models’ predictions could be reduced to any of the others. We first investigate whether UIDev adds any predictive power to probability and vice versa. The results from the two factor regression analysis are displayed below.

|                | Estimate | Error  | z value | Pr(>|z|)    |
|----------------|----------|--------|---------|------------|
| Intercept      | -.54693  | .27498 | -1.989  | .04670 (p=.01) |
| Probability Coefficient | .02933  | .07209 | .407    | .68417 (p=1) |
| UIDev Coefficient | -6.58210| 2.20135| -2.990  | .00279 (p=.001) |

Table 9. A two factor model of UIDev and probability
<table>
<thead>
<tr>
<th></th>
<th>Intraclass correlation</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>-.025</td>
<td></td>
</tr>
<tr>
<td>UIDEv</td>
<td>.007</td>
<td>.837</td>
</tr>
</tbody>
</table>

Table 10. Correlation of fixed effects

The two factors are strongly correlated. This result is not surprising, given the similarities between the two models. A PCFG defines the total probability of a sentence as the product of the phrase structure rules used to generate it. UIDEv calculates the sum of the differential between each successive point in the information profile and divides the result by the total number of words in the sentence. Both metrics depend on probability values. The crucial difference between the two is that, in UIDEv, the linear order of these values matters. The maximum probability hypothesis says nothing about the order in which the probability values should optimally occur; it simply states that if order (a) has a higher overall probability than order (b), then order (a) should be preferred.

The regression results show that probability does not add any significant predictive power to UIDEv. In other words, the effects of probability in the one factor model above came about simply because of probability’s strong relationship with UIDEv.

We then ask if the two different ways of quantifying an information profile’s uniformity, UIDEv and variance, can be reduced to one another. The regression results are displayed
below.

|                | Estimate | Error | z value | Pr(>|z|) |
|----------------|----------|-------|---------|---------|
| Intercept      | -.5253   | .2733 | -1.922  | .0546   |
| Variance       | -1.5158  | 1.7528| -.865   | .3872 (p=1) |
| Coefficient    |          |       |         |         |
| UDev           | -6.6073  | 1.4379| -4.590  | 4.44e-6 (p=0) |
| Coefficient    |          |       |         |         |

Table 11. A two factor model of UDev and variance

<table>
<thead>
<tr>
<th></th>
<th>Intraclass Coefficient</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance</td>
<td>-.059</td>
<td></td>
</tr>
<tr>
<td>UDev</td>
<td>.065</td>
<td>-.550</td>
</tr>
<tr>
<td>Coefficient</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12. Correlation of fixed effects

It is interesting to note that UDev and variance are negatively correlated. This seems strange, since UDev and variance measure different aspects of an information profile’s uniformity (see Chapter 3). A manual examination of the raw data, however, confirms this
result. Since the UIDev and variance metrics output very different statistical information, there are many instances in which they make conflicting predictions (e.g. assigning some sentences high variance and low UIDev or vice versa). They are powerful single predictors, but they do not often agree.

We do observe, however, that variance contributes no additional predictive value to UIDev. Thus, we conclude that neither total sentence probability nor an information profile’s variance has any effect on subjects’ word order preferences. It appears that language users are more sensitive to the difference between the successive points in the information profile than to the total time spent transmitting below channel capacity. This result makes good sense in light of the theoretical arguments for the cognitive plausibility of this conception of uniformity (see Chapter 3.2).

We now run a two factor model comparing UIDev and MPWS. The results are displayed below.
|                      | Estimate | Error | z value | Pr(>|z|) |
|----------------------|----------|-------|---------|----------|
| Intercept            | -.5375   | .2748 | -1.956  | .0505    |
| MPWS Coefficient     | .1387    | 1.5211| .091    | .9273 (p=1) |
| UIDev Coefficient    | -7.3498  | 1.2779| -5.751  | 8.85e-9 (p=0) |

Table 13. A two factor model of MPWS and UIDev

<table>
<thead>
<tr>
<th></th>
<th>Intraclass Coefficient</th>
<th>MPWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPWS</td>
<td>.004</td>
<td></td>
</tr>
<tr>
<td>UIDev</td>
<td>.047</td>
<td>-.313</td>
</tr>
</tbody>
</table>

Table 14. Correlation of fixed effects

Here too we observe a negative correlation between the two metrics; this indicates a lack of agreement between the two predictors. Again, this fact can be explained by examining the raw data. In many token pairs, the MPWS value of both variants is exactly equal. In these cases, MPWS can make no prediction about which option should be preferred. In contrast, UIDev, a metric with more descriptive power, can make more fine-grained distinctions and
differentiate between cases in which MPWS fails.

This observation is confirmed by the regression results. The two factor regression model suggests that MPWS adds nothing to UIDev’s predictions; MPWS is, in fact, subsumed by UIDev. This result is of great theoretical interest. Recall that both UID and MPWS claim that sharp spikes in information should be avoided. UID, however, goes one step further and states that “valleys” of information should be avoided as well. The empirical results presented here substantiate this claim. It is not enough to impose a “ceiling” on maximum per word surprisal; one must also create an information “floor” as well. This result is of great theoretical interest.

Recall that Information Theory models the communicative process as a balance between two opposing tendencies: efficiency and reliability. In essence, MPWS sacrifices the first for the sake of the second. It correctly claims that language users should strive for reliability in communication, and minimizing MPWS certainly aids in this goal. However, as it stands, MPWS does not require communicative efficiency. Taken to its logical extreme, MPWS claims that language users have complete freedom to be as uninformative and inefficient as they want, just as long as the maximum per word surprisal stays below some threshold. This idea is not cognitively plausible and does not reflect actual language behavior. As it stands, MPWS is an incomplete theory of communication. UID, which builds upon and extends the fundamental insight of Hale 2001, is both more cognitively plausible and empirically powerful.
Finally, we ran a two factor model comparing UIDev and DLM. The results are displayed below.

|               | Estimate | Error | z value | Pr(|z|) |
|---------------|----------|-------|---------|--------|
| Intercept     | -.52396  | .27264| -1.922  | .0546  |
| DLM Coefficient | -.09801  | .04886| -2.006  | .0448  |
| UIDev Coefficient | -6.54039 | 1.24903| -5.236  | 1.65e-07 |

Table 15. A two factor model of DLM and UIDev

<table>
<thead>
<tr>
<th></th>
<th>Intraclass correlation</th>
<th>DLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLM</td>
<td>-.004</td>
<td></td>
</tr>
<tr>
<td>UIDev</td>
<td>.052</td>
<td>-.303</td>
</tr>
</tbody>
</table>

Table 16. Correlation of fixed effects

These two variables are less strongly correlated with one another. However, the fact that the regression model assigns both factors a statistically significant value suggests that both
UIDev and DLM make substantial *independent* contributions to the subjects’ decision to choose item (b). Neither theory can be subsumed by the other. However, UID has a much more significant effect than DLM. This result is also of great theoretical interest.

A particularly striking feature of Table 2 is that subjects’ preferences are not always what Wasow 2002’s Principle of End Weight (PEW) and DLM would predict. For instance, PEW/DLM predicts that heavy NP’s should always undergo heavy NP shift. However, subjects clearly preferred heavy NP’s in situ. Additionally, in Template 7, the PP (always of the form P Det N) contains more syntactic and phonological weight than the adverb. Therefore, according to PEW/DLM, PP should come after the adverb. However, this order was not what the subjects tended to prefer. In these templates, subjects chose the exact opposite of what DLM predicted. At the same time, other templates show that PEW is not completely invalid. The subjects’ response to Templates 6 and 8 receive a straightforward PEW/DLM explanation. Dependency length seems to be a salient feature in some word orders, while in others it is relegated to a subordinate position. DLM clearly has an effect, but it cannot account for all of the data. UID is a much better overall predictor of word order preferences.
Chapter 8: Summary

The mixed effects regression model yielded two major results:

(i) MPWS does not seem to have a significant effect on word order preferences.

(ii) DLM does have a significant effect on word order preferences, but UID’s effect is much greater.

These results suggest that processing difficulty is most heavily affected by how the information content of a sentence is linearly arranged. Certain arrangements facilitate processing and others hinder it. Language users tend to arrange their words in a way that both minimizes dependency length and results in a uniform overall information contour. Crucially, language users are sensitive both to abrupt spikes and troughs in an information profile (contra MPWS).

Chapter 4 delineates the weaknesses of DLM. It is a metric that is largely designed to model heaviness effects. As such, it encounters difficulty in modeling word order preferences involving “light” constituents. In addition, there are certain word order variations characterized by equal dependency lengths between variants. In cases like these, DLM is able to make no prediction about which should be the preferred order. The empirical results presented here support this theoretical critique. DLM was shown to have
a statistically significant effect on word order preferences, but its effects were relatively small compared to those of UID. UID, which does not suffer from the above limitations, gives a more complete account of word order phenomena.

Our results extend the work of Aylett & Turk 2004, Levy & Jaeger 2007, and Frank & Jaeger 2008 by demonstrating the effects of UID at the sentence level. UID was far and away the most powerful influence on variable word order preferences. Variable word order thus serves as a way to modulate and control the information profile in a sentence in the spirit of Aylett & Turk 2004, Levy & Jaeger 2007, and Frank & Jaeger 2008.
Chapter 9: Conclusion

This thesis has made several important contributions to the literature on word order and processing. First, it extends the empirical domain by considering word order variations of “light” elements such as adverbs and NP’s such as Det N. The ordering preferences of these constituents have often been neglected by previous psycholinguistic models of word order.

Second, it suggests that word order preferences are not determined by a single psycholinguistic metric, but rather a combination of different contributing factors. Attempts at reduction, such as PEW, often cannot account for important empirical facts. The joint effects of both DLM and UID on word order preferences are in line with psycholinguistic models that stress the interactive nature of language processing (Marslen-Wilson 1975, Tanenhaus et al. 1995, inter alia).

Finally, this thesis shows that the ideas of Aylett & Turk 2004, Levy & Jaeger 2007, and Frank & Jaeger 2008 can be fruitfully applied to model variable word order. Language appears to strive for uniform information density at almost every level of description. Such a fact provides support for using the tools of Information Theory (Shannon 1948) to investigate complex problems involving the relationship between language and cognition.
References


Quene, H. & H. van den Bergh. 2008. Examples of mixed effects modeling with crossed
random effects and with binomial data. *Journal of Memory and Language* 59(4): 413-425. (Special Issue: Emerging Data Analysis).


Appendix: Mechanical Turk Survey Questions
Template 1: NP Adverb Be PP vs. NP Be Adverb PP

1.  
   (a) I often am in the house.  
   (b) I am often in the house.  

2.  
   (a) People are sometimes in the park.  
   (b) People sometimes are in the park.  

3.  
   (a) Sharks usually are in the ocean.  
   (b) Sharks are usually in the ocean.  

4.  
   (a) Deer are frequently in the road.  
   (b) Deer frequently are in the road.  

5.  
   (a) Horses seldom are in the kitchen.  
   (b) Horses are seldom in the kitchen.  

6.  
   (a) Gold is always in the treasury.  
   (b) Gold always is in the treasury.  

7.  
   (a) Professors rarely are in the cafe.  
   (b) Professors are rarely in the cafe.  

8.  
   (a) She is occasionally in the gym.  
   (b) She occasionally is in the gym.  

9.  
   (a) The dishes never are on the table.  
   (b) The dishes are never on the table.  

10.
(a) I was just in the same store!
(b) I just was in the same store!

Template 2: NP Have Adverb V-en NP vs. NP Adverb Have V-en NP

1.  
(a) Goblins have often visited the ruins.
(b) Goblins often have visited the ruins.

2.  
(a) I sometimes have told lies.
(b) I have sometimes told lies.

3.  
(a) Students have usually passed the exam.
(b) Students usually have passed the exam.

4.  
(a) This worker frequently has disappointed his superiors.
(b) This worker has frequently disappointed his superiors.

5.  
(a) The athlete has seldom forgotten his gear.
(b) The athlete seldom has forgotten his gear.

6.  
(a) Farmers always have eaten bread.
(b) Farmers have always eaten bread.

7.  
(a) The president has rarely deceived the voters.
(b) The president rarely has deceived the voters.

8.  
(a) The government occasionally has admitted its mistakes.
(b) The government has occasionally admitted its mistakes.

9.  
(a) The dunce has never done his homework.
(b) The dunce never has done his homework.
10.
   (a) My uncle just has won the lottery.
   (b) My uncle has just won the lottery.

**Template 3: NP V Particle Det N vs. NP V Det N Particle**

1.
   (a) She looked up the number.
   (b) She looked the number up.

2.
   (a) You ran the battery down.
   (b) You ran down the battery.

3.
   (a) Sandy put up the pictures.
   (b) Sandy put the pictures up.

4.
   (a) The man switched the light off.
   (b) The man switched off the light.

5.
   (a) The researcher figured out the problem.
   (b) The researcher figured the problem out.

6.
   (a) The bartender filled my glass up.
   (b) The bartender filled up my glass.

7.
   (a) The lifeguard took off his shirt.
   (b) The lifeguard took his shirt off.

8.
   (a) I turned the promotion down.
   (b) I turned down the promotion.
9.
(a) The girl tried on the dress.
(b) The girl tried the dress on.

10.
(a) The drug dealer used his supplies up.
(b) The drug dealer used up his supplies.

Template 4: NP V Adverb PP vs. NP V PP Adverb

1.
(a) I sleep often on a bench.
(b) I sleep on a bench often.

2.
(a) Children play in the field sometimes.
(b) Children play sometimes in the field.

3.
(a) I ran reluctantly across the desert.
(b) I ran across the desert reluctantly.

4.
(a) I eat at this restaurant frequently.
(b) I eat frequently at this restaurant.

5.
(a) The dog swims happily in the lake.
(b) The dog swims in the lake happily.

6.
(a) Most campers stay on the path willingly.
(a) Most campers stay willingly on the path.

7.
(a) The mail comes rarely at this time.
(b) The mail comes at this time rarely.
8. (a) The lawyer acts with some decency occasionally.  
(b) The lawyer acts occasionally with some decency.

9. (a) The businessman arrived yesterday without his briefcase.  
(b) The businessman arrived without his briefcase yesterday

10. (a) The king thought about his son today.  
(b) The king thought today about his son.

**Template 5: Heavy NP In Situ vs. Heavy NP Shift**

1. (a) The young boy gave the beautiful green ring that had been in the jewelry store window for weeks to the girl. 
(b) The young boy gave to the girl the beautiful green ring that had been in the jewelry store window for weeks.

2. (a) The landlord talked about the problem to the visitor who had just arrived and was staying for the summer. 
(b) The landlord talked to the visitor who had just arrived and was staying for the summer about the problem.

3. (a) I received the books which my uncle left me as part of my inheritance on Sunday.  
(b) I received on Sunday the books which my uncle left me as part of my inheritance.

4. (a) John dedicated to Cynthia the song that he had been working on intermittently for weeks. 
(b) John dedicated the song that he had been working on intermittently for weeks to Cynthia.
5. (a) I postponed the ambitious project that would require hours in front of the computer and several days out in the field to next week.  
(b) I postponed to next week the ambitious project that would require hours in front of the computer and several days out in the field.

6. (a) She accepted from Paul the richly embroidered quilt encrusted with seven different varieties of precious gemstones.  
(b) She accepted the richly embroidered quilt encrusted with seven different varieties of precious gemstones from Paul.

7. (a) The executive lent all the profits of his multibillion dollar dot com company to his nephew.  
(b) The executive lent to his nephew all the profits of his multibillion dollar dot com company.

8. (a) The magician showed to the audience a large black top hat that looked entirely ordinary.  
(b) The magician showed a large black top hat that looked entirely ordinary to the audience.

9. (a) The driver supplied the eight ton tractor trailer that was carrying all the gold in Fort Knox with fuel.  
(b) The driver supplied with fuel the eight ton tractor trailer that was carrying all the gold in Fort Knox.

10. (a) The manager sold to his friend the baseball team that had won four world series and numerous other smaller titles.  
(b) The manager sold the baseball team that had won four world series and numerous other smaller titles to his friend.
Template 6: Complex Subject vs. It-extrapolation

1.  
   (a) That the administrator lost the medical reports bothered the intern.  
   (b) It bothered the intern that the administrator lost the medical reports.

2.  
   (a) It annoys me that the computer program rarely works correctly even when it should.  
   (b) That the computer program rarely works correctly even when it should annoys me.

3.  
   (a) That the government is full of corrupt and greedy politicians is disgraceful.  
   (b) It is disgraceful that the government is full of corrupt and greedy politicians.

4.  
   (a) It is rather sad that mathematics is both feared and shunned by the general populous.  
   (b) That mathematics is both feared and shunned by the general populous is rather sad.

5.  
   (a) That Mozart was a child prodigy possessed of unparalleled talent is obvious.  
   (b) It is obvious that Mozart was a child prodigy possessed of unparalleled talent.

6.  
   (b) It is doubtful that Shakespeare plagiarized his plays and poems.  
   (a) That Shakespeare plagiarized his plays and poems is doubtful.

7.  
   (a) That people should not be punished for a crime they did not commit is clear.  
   (b) It is clear that people should not be punished for a crime they did not commit.

8.  
   (a) It is idealistic to assume that people always seek to better themselves.  
   (b) That people always seek to better themselves is idealistic to assume.
9.  
(a) That water bottles are not to be used as floatation devices in the event of a crash landing is apparent to anyone with a brain.  
(b) It is apparent to anyone with a brain that water bottles are not to be used as floatation devices in the event of a crash landing.

10.  
(a) It surprises me that you would think Chris capable of such treachery.  
(b) That you would think Chris capable of such treachery surprises me.

Template 7: PP In Situ vs. PP Extraction

1.  
(a) I work in the house.  
(b) In the house, I work.

2.  
(a) In the forest, I walk.  
(b) I walk in the forest.

3.  
(a) I live in the city.  
(b) In the city, I live.

4.  
(b) In the country, I hike.  
(a) I hike in the country.

5.  
(a) I study in the library.  
(b) In the library, I study.

6.  
(a) In the office, I sit.  
(b) I sit in the office.

7.  
(a) The cat climbed up the tree.

66
(b) Up the tree, the cat climbed.

8.
(a) On my shirt, the chili fell.
(b) The chili fell on my shirt.

9.
(a) He jumped into the puddle.
(b) Into the puddle, he jumped.

10.
(a) On the table, the cat slept.
(b) The cat slept on the table.

Template 8: NP V Particle NP (heavy) vs. NP V NP (heavy) Particle

1.
(a) She looked up the number that was in the very back of the yellow pages.
(b) She looked the number that was in the very back of the yellow pages up.

2.
(a) You ran the battery that was powering both the fridge and the microwave down.
(b) You ran down the battery that was powering both the fridge and the microwave.

3.
(a) Sandy put up the pictures of her vacation in the Virgin Islands.
(b) Sandy put the pictures of her vacation in the Virgin Islands up.

4.
(a) The man switched the light that was in the kitchen off.
(b) The man switched off the light that was in the kitchen.

5.
(a) The researcher figured out the problem that had been troubling him for weeks.
(b) The researcher figured the problem that had been troubling him for weeks out.

6.
(a) The bartender filled my glass which was on the counter next to my wallet up.
(b) The bartender filled up my glass which was on the counter next to my wallet.
7.  
(a) The lifeguard took off his shirt that had been sprayed with shaving cream.  
(b) The lifeguard took his shirt that had been sprayed with shaving cream off.

8.  
(a) I turned the promotion that would increase my salary by twenty percent down.  
(b) I turned down the promotion that would increase my salary by twenty percent.

9.  
(a) The girl tried on the dress that was purple with pink polka dots.  
(b) The girl tried the dress that was purple with pink polka dots on.

10.  
(a) The drug dealer used his supplies that he had purchased on the corner up.  
(b) The drug dealer used up his supplies that he had purchased on the corner.