It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.

– Albert Einstein, 1933.

University of Alberta

The Processing of Lexical Sequences

by

Cyrus Shaoul

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Psychology

Spring 2012
Edmonton, Alberta

Kerstin, Darius and Ilona
For inspiring me to go on.

Abstract

Psycholinguistics has traditionally been defined as the study of how we process units of language such as letters, words and sentences. But what about other units? This dissertation concerns itself with short lexical sequences called *n*-grams, longer than words but shorter than most sentences. *N*-grams can be phrases (such as the 3-gram *the great divide*) or just fragments (such as the 4-gram *means nothing to a*). Words are often thought to be the universal, atomic building block of longer lexical sequences, but *n*-grams are equally capable of carrying meaning and being combined to create any sentence. Are *n*-grams more than just the sum of their parts (the sum of their words)? How do language users process *n*-grams when they are asked to read them or produce them? Using evidence that I have gathered, I will address these and other questions with the goal of better understanding *n*-gram processing.

Acknowledgements

The research in this thesis would not have been possible without the support of Dr. Chris Westbury and the members of the Westbury Lab. I collaborated with Georgie Columbus on the eye-tracking research in Chapter 3 (She was responsible for the experimental software programming and part of the data collection). Dr. Chris Westbury, Dr. Harald Baayen, Dr. Christina Gagné and Dr. Norman Brown provided advice and inspiration throughout the development of this dissertation.

Chapter 1, in part, is a revised version of the material as it appears in Shaoul and Westbury (2011). The dissertation author was the primary investigator and author of this material.

Chapters 2, 3 and 4, in full, are currently being prepared for submission for publication. The dissertation author was the primary investigator and author of this material.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

There is a new and growing interest in psycholinguistics in the mental representation of (not necessarily phrasal) lexical sequences and in how knowledge of these sequences relates to word, phrase, and sentence knowledge. In this chapter I summarize the evidence for the existence of distinct mental representations for these types of sequences. Studies of sentence processing, contextual ambiguity resolution, speech production and compound word processing provide indirect evidence for frequency effects for lexical sequences. Recent studies of adult reading behaviour have looked more directly at the effects of holistic frequency on reading performance. I end by considering the relevance of lexical sequences to existing cognitive models of language and speculating on how they may impact future models.

The urge to analyze language in a reductionist way has been evident from the earliest psycholinguistic experiments. Language has long been conceived of as a stream of discrete words that can be broken down into their components (morphemes, syllables, phonemes, and letters) or combined together to produce phrases or sentences. Relatively little consideration has been given to linguistic units intermediate between these two levels of analysis: lexical sequences that are not necessarily phrasal. In this chapter I address two main questions: Are there any units of language that have a mental representation that is larger than a word but smaller than a phrase, and do they matter to models of human language use? I will review relevant research from the fields

of psycholinguistics and linguistics that bear on these questions.

Though words and sentences have been studied extensively by psycholinguists, the concept of lexical sequences as a behaviorally relevant unit is a relatively new one (Bybee & Scheibman, 1999) [1]. The concept has been given many names by different researchers. Some terms that have been used (not quite identically) include: lexical bundles (Biber, Conrad, & Cortes, 2004), formulae, collocations, adjuncts, idioms, multi-word expressions, multi-word sequences, chunks, holophrases, prefabricated routines, lexical patterns, word combinations, multi-word combinations, N-grams, and formulaic sequences (Wray, 1998). Some definitions of these sequences allow for flexibility in intervening words with open slots, some allow for arbitrarily long sequences, some imply figurative meaning, and some imply semantic compositionality or non-compositionality. The scope of these terms is potentially very large. In this dissertation I will use the term *n*-grams, and limit it strictly to the subset of non-lexicalized word sequences that are between two and five words long.

Despite the lack of consensus on the definition of formulaic sequences, linguists agree that they are extremely prevalent in both spontaneous speech and writing. Biber (1999) proposed that for a series of words to be considered to be a formulaic sequence, it must occur at least ten times per million in a corpus for sequences between two and four words long, and at least five times per million for longer sequences. This is obviously an arbitrary threshold, but its use has become a common way for investigators to identify formulaic sequences. According to Erman and Warren (2000), over 50% of spoken and written language is made up of such formulaic sequences. The Google Web1T Database (Brants & Franz, 2006), which consists of approximately one trillion word tokens of text found in publicly accessible Web pages, lists approximately 78 million formulaic sequences of two words, 244 million sequences of three words, 328 million of four words, and 294 million of five words. These numbers are so large that it may seem beyond the scope of possibility that the human

[1] I will not directly address the extensive literature on compound word processing in this dissertation, since it remains unclear if, or to what extent, lexicalized multi-morphemic words are processed as *n*-grams.

brain could represent any information about formulaic sequences. However, a growing body of evidence suggests that they can. I begin by presenting a brief history of the study of formulaic sequences and of the evidence suggesting that they may be represented.

1.1 Foundations: Information & Learning Theory

The conceptual roots for conceiving language in terms of *n*-grams comes from information theory. The behavioral manifestations of an information-theoretic conceptualization of the mind have been studied in learning theory. In this section I briefly review these two foundational theories.

1.1.1 Information Theory

One of the first people interested in the probabilistic nature of language was Claude Shannon (Shannon, 1948). He demonstrated that strings of letters generated at random with the constraint that they follow real bigram distributions produced words that were word-like and pronounceable, and that sentences created by stringing together words in the same way produced sentences that were sometimes readable. The computational resources he had available for this work were of course tiny compared to those that exist today. After discussing how sentences could be approximated using two-word transition probabilities, Shannon remarked that "It would be interesting if further approximations could be constructed, but the labor involved becomes enormous at the next stage." (Shannon, 1948, p.8). Sixty years later, advances in computational technology have made this enormous labor tractable. The arguments made by Shannon (1948) form a foundational assumption of information theory relevant to *n*-grams: that analyzing the probability of discrete patterns in a signal can lead to insights about how to simplify the processing of that signal. During the decade after Shannon's paper was published, psychologists delved into information theory. They felt that this new direction of probabilistic inquiry would have powerful implications for the future of psy-

cholinguistics (Osgood et al., 1954). However, during the Chomskian era that followed, with its emphasis on rule-based generative models, almost no further work was done on probabilistic models of psycholinguistics (Newmeyer, 1996). The beginning of the large-data revolution in the 1990s marked the return of probabilistic psycholinguistics (Jurafsky, 2003).

1.1.2 Learning Theory

Much of what we know about how organisms deal with statistically patterned information comes from traditional learning theory. Learning theory and theories of language learning and use have largely been de-coupled for decades, in part due to the belief in the "poverty of the stimulus" — the claim originally due to Chomsky (1980) that there is insufficient information in the language stream for a language user to master language, especially with little or no negative evidence. The corollaries (or, perhaps more accurately, the pre-assumptions) of this claim were that language was special, that it required special resources for learning and processing, and that therefore language learning could never be explained by the same learning theories being used to explain other forms of learning. I believe that the claim of the poverty of the stimulus is an example of what the philosopher Daniel Dennett (1991, p. 401) has called "Philosopher's syndrome": mistaking a failure of imagination into an insight of necessity. One of the exciting aspects of work on sequences is that it schools our imaginations, making it possible for us to imagine how traditional learning theory might be able to account for language learning and use. It does this by giving us a principled way of understanding how much information can actually be extracted from the language stream. When we conceive of language as a set of nested sequences (of letters or phonemes and words) whose (first-order and higher-order[2]) co-occurrence probabilities may be simultane-

[2]First order co-occurrence refers to things that co-occur in close proximity — the normal meaning of *co-occurrence* and the main focus of this dissertation. Second order co-occurrence refers to things that do not occur in close proximity to each other but do occur in close proximity to the same things — in other words, to things that share context. Second-order co-occurrence is of central importance in co-occurrence models of semantics (see Landauer & Dumais, 1997). Higher-order co-occurrences are well-defined and may be relevant, but they are so far unstudied in language.

ously computed, the language stream appears richer in information than it has traditionally been imagined to be. Early learning theorists studied how the frequency of exposure to a stimulus improved fluency in processing that stimulus. A well-known outcome of increased exposure is increased processing speed, as captured in the power law of learning (Newell, 1990; Speelman, 2005) which describes the relationship between practice and performance in the acquisition of a wide range of cognitive skills. Just as with other skills, exposure to words has the largest impact in the early stages of language acquisition, but after enough repeated exposure, the impact shrinks. The shape of the tail of the learning distribution curve is a general property of practice. As such, any theory of formulaic language will need to address how frequency of exposure shapes learning. Diessel (2007), building on ideas from learning theorists (Anderson, 1982; Newell, 1990), proposed three psychological mechanisms that underpin frequency effects specifically in learning word sequences:

1. Increased frequency causes the strengthening of linguistic representations. Increased exposure reinforces the representation in memory. This, in turn, influences the activation and interpretation of these representations during language use.

2. Increased frequency causes the strengthening of expectation. Words are arranged in recurrent orders, and people develop expectations as to which word or words may occur after a particular word or set of words.

3. Increased frequency causes the automatization of chunks. Words that are frequently combined together may develop into discrete processing units, where the boundaries between words become unclear and the whole chunk becomes reduced or compressed, as, for example, in the reduction of the phrase *going to* to *gonna* or the phrase *don't know* to *dunno*.

These three simple mechanisms together provide a set of tools for constructing a rich model of language. I will return to the relationship between learning theory, information theory, and *n*-grams in the final section of this chapter.

1.2 Empirical Evidence of *N*-gram Processing

Learning theory and information theory provide the theoretical foundations for building a model of how n-grams may be sufficient for building a model of language. In this section I review some of the empirical evidence that suggests that people are indeed sensitive to the probabilities of sequences, across multiple dependent measures and subdomains of psychological research.

1.2.1 Evidence from memory research

Memory research has addressed word-association and how theories of episodic memory may apply to the learning of associations. Recall accuracy are used as a measure of well an association has been learned. By manipulating the a priori co-occurrence between two words, one deduce whether or not those a priori probabilities are impinging upon the strength of association of those words, and therefore whether or not they were already associated in a person's mind. By manipulating the context of recall, one can deduce whether or not that association is context-dependent.

The work of Prior and Bentin (2003) is a good example of this kind of work, and of particular relevance. Prior and Bentin were interested in the possibility of incidental word associations being stronger when the words were seen in a sentence rather than when they were seen without any context. They used a three-stage experiment. First, pairs of nouns were read by the subjects in one of two styles: in a sentence or as separate words. Subjects were asked to make an incidental semantic category judgment (*flower or jewelry*?). These pairs were shown five times to each subject, during which time learning or implicit memory encoding took place. Subjects were unaware that they would later be asked to remember paired associations. In the second stage, subjects performed an explicit learning task with the same words as in the first stage, with the addition of a set of words that were not shown in the first stage. They were asked to memorize word pairings for later testing. In the final testing stage, subjects were asked to perform a cued-recall task (given one member of a pair, recall the other one) and a new-old single word recognition

test (with an equal number of old and new words). Critically, performance on the cued-recall task was reliably more accurate for the incidental sentence exposure stimuli than for the non-sentential stimuli. There was no difference between the non-incidental stimuli and the non-sentential stimuli. To rule out the confounding influence of sentence memorization, Prior and Bentin did the experiment again, using five difference sentence contexts instead of repeating the same sentence five times. This manipulation did not change the results. Their conclusion, providing a basis for the psychological reality of *n*-grams, was that the mere act of reading words in the context of a sentence triggers an associative process that links co-occurring words together.

In a subsequent study, Prior and Bentin (2008) demonstrated that the mechanism behind this phenomenon was that words are incidentally associated as part of semantic integration during sentence comprehension. They used a similar experimental method, but instead of using only meaningful sentences, they added syntactically acceptable but semantically anomalous sentences such as *The brown shoe pleased the tired fly.* Consistent with their previous (2003) study, they found that episodic memory of a shared context created an association that transferred over to later paired memory tasks, but only for the coherent sentences. Anomalous sentences created significantly weaker associations. They also tested for sentence recall, finding that coherent sentences were recalled much more accurately than anomalous sentences.

These results point to the quality of the memory traces as the source of the association boost. Anomalous sentences that could not gain any boost from semantic integration created weaker memory traces. To determine if the explicit paired association task was required to obtain this effect, Prior and Bentin (2008) conducted a final experiment using a set of anomalous and coherent sentences. This time there was no explicit memorization during test. Rather, implicit association was tested using a sequential new/old recognition task. Subjects again had a strong associative priming effect for word pairs that were seen in coherent sentences, but not for those that were seen in anomalous sentences.

Together these findings have clear implications for a theory of *n*-gram pro-

cessing. Words that appear during natural language use in the same context build associations with each other over time, and this association process is automatic and efficient when the context is semantically coherent.

1.2.2 Evidence from linguistic ambiguity research

Studies of lexical ambiguity in language have also provided empirical evidence suggesting that we are sensitive to word co-occurrence frequencies. Evidence that word co-occurrence statistics are involved in the interpretation of an ambiguous sentence would imply that n-gram probabilities are both accessible and used in normal language processing.

One of the first studies to approach semantic ambiguity effects from this perspective was a study by Macdonald (1993). She looked at reading times just after ambiguous words embedded in sentences, such as the word *fires* in the sentence *The union told reporters that the warehouse fires many workers each spring without giving them proper notice.* She found that there was a strong tendency towards incorrect interpretations for ambiguous sentences when there was a supportive bias towards the incorrect interpretation. Reading times were reliably slower for sentences in that category. Both the probability of the first word being the head or modifier and, most importantly, the probability of the two words co-occurring accounted for variance in the ambiguity resolution tasks.

1.2.3 Evidence from acceptability judgment research

Acceptability judgments are closely related to ambiguity judgments, and the same logic applies for the use of acceptability judgment in studies of formulaic sequences: evidence of word co-occurrence impinging on acceptability judgments is taken as evidence that those co-occurrences are accessible and normally used in language processing. However, acceptability judgments are also of particular interest because they address an important theoretical point of contention, the poverty of the stimulus argument. One of motivations for the poverty of the stimulus argument and related arguments for "special functionality" for language processing is that it just seems intuitively obvious that

language users should not be able to understand sequences they have never been exposed to, but it is also obvious that we can understand them. When we think of sentences as being composed of multiple levels of co-occurrence rather than as a single flat structure, it becomes possible to start doubting that intuitively obvious belief: that is, it becomes possible to see the poverty of the stimulus argument is a failure of imagination.

The key to overcoming the idea that unencountered sentences contain no information is the idea of statistical smoothing (closely related to second order co-occurrence, which was mentioned in Footnote 2). Statistical smoothing is a set of techniques invented by natural language processing engineers that allows them to assign probabilities greater than zero to word strings that have never been encountered by their language models, and which therefore have empirical probabilities of zero from the model's point of view. One technique, called distance-weighted averaging, starts from the assumption that the probability of a particular string is not only dependent on the transition probabilities of words in that string, but also on the transition probabilities of words in strings that are similar. For example, if we have never seen word *aardvaark* followed by the word *dig*, we can compute the probability that word *dig* will follow words that are related to *aardvaark*, and use the average probability of those related strings to estimate the probability of the unencountered string. If the language stream tells us that many other mammal names can be followed by the word *dig*, then we have some reasonable grounds to assume that the unencountered phrase *aardvaarks dig* is probably acceptable. Roberts and Chater (2008) used statistical smoothing to see if novel sentences would be more acceptable if they had a higher estimated probability. They extracted adjectives and nouns from a corpus of English and constructed new sentences with them that included novel (zero frequency) co-occurrence sequences. One example of a pair of these zero frequency sequences is: *The intolerable mediocrity will be discussed* and *The unequivocal tsarist will be discussed*. Adjectives that co-occurred with similar nouns were the ones chosen by their statistical smoothing technique. In the example given, the smoothed frequency of the first sentence is higher than the second. Sentences with higher smoothed frequencies were

rated significantly higher on a scale of acceptability by their participants. Acceptability for these types of sentences was accurately predicted by smoothed co-occurrence probabilities. This is strong evidence for a probabilistic model of lexical preference in these types of tasks.

1.2.4 Evidence from phonology research

Related evidence of sensitivity to subtle informational probabilities in the lexical stream comes from studies of phonological reduction. The efficiency of producing spoken language is predicted to increase with practice. Words that co-occur together more frequently should receive more practice than words that occur together less frequently and should, therefore, be more predictable for the listener who will be able to understand the words even in a reduced form. The information contained in word co-occurrence data should therefore predict phonologically reduction. One measure of word co-occurrence is Mutual Information (Church & Hanks, 1990), which is defined the following way:

$$MI(X;Y) = log\left(\frac{(Frequency(XY))}{(Frequency(X)) \times (Frequency(Y))}\right) \tag{1.1}$$

The MI between two linguistic tokens is the degree to which the first token predicts the occurrence of the second. Several researchers have shown that this informational measure predicts phonological reduction.

For example, Pluymaekers, Ernestus, and Baayen (2005a) studied face-to-face conversations in Dutch and found that articulatory planning is continuous and sensitive to informational redundancy. They focused on words using the adjectival suffix *-lijk* and measured the time to say the words in different contexts. One of the best predictors of phonological reduction was Mutual Information for the preceding and following words. Similar findings have been reported by Church and Hanks (1990), Bybee and Scheibman (1999) and Bybee (2002).

1.2.5 Evidence from eye movement studies

Many psycholinguists have begun using gaze tracking to gather data about the location and time of visual fixation of subjects as they read natural language. Eye-tracking measures take advantage of the fact that contextual predictability has a large impact on the ease of processing a text. Ease of processing manifests itself operationally in measures such as the probability of making a regression (re-reading text that has already been read), how long a person fixates on each word, and second-pass reading time (how much time is spent in re-reading). One of the benefits of using this methodology in formulaic sequence research is that the stimuli used can be more ecologically valid than in many other paradigms, allowing for the study of regular text instead of just single words. Furthermore, the fixation data are the outcome of automatic processes that are independent of participant control and decision-making, allowing us to find evidence that sequence probabilities affect the automatic functional structuring of language processing. These factors make eye tracking a potentially powerful tool to study responses to the probabilistic nature of language.

One of the first studies done on the influence of transitional probabilities on eye movements was carried out by McDonald and Shillcock (2003). They used corpus data from the British National Corpus (BNC) to calculate the probability of the previous and next word when reading, looking for a relationship between this transitional probability and eye movements. They found that both forward and backward transitional probabilities predicted an early processing measure (first fixation duration, how long a person fixates on a word for the first time) and a later processing measure (gaze duration) independent of other factors such as length, launch distance, and word frequency. As would be expected, first fixation and gaze duration were both shorter for bigrams with higher transitional probabilities. They note that these results were contingent on availability of parafoveal information about target words, implying that frequency effects for fixations across word pairs depend on fast feedback from the visual system using parafoveal preview of the next word. McDonald and Shillcock (2003) propose that linguistic experience fosters the

creation of a representation of contingency statistics. This ability to predict upcoming words using lexical statistical information is a computationally inexpensive mechanism that may contribute to proficient reading.

1.2.6 Evidence from reading speed and repetition studies

A more direct way to measure the effect of information measures on language than looking at eye movements or phonological reduction is to look for effects of sequence probabilities on reading speed. The assumption is that well-learned sequences should be easier to read or repeat (as measured by reading or repetition time) than less well-learned sequences. Despite the apparent ease by which this kind of experiment can be carried out, this type of investigation has not been attempted until very recently. Several experimental issues make this type of research more difficult than it may seem. The first is that stimuli must be carefully chosen because if care is not taken, any effects could be related to the frequency of the individual words in the sequence, or perhaps to substrings in the sequence (a two or three word sequence that is contained within a four word sequence, for example). The second is that it is not trivial to obtain large corpora and to calculate all the frequencies for all sequences of all lengths in those corpora. Selecting matched stimuli from these lists is also a challenge. For these reasons, some of the recent attempts to look at reading of formulaic sequences have been weakened by a lack of experimental control.

For example Conklin and Schmitt (2008) found evidence that formulaic (idiomatic) sequences, such as *a breath of fresh air*, were read faster than near-identical sequences that had small changes in their word order rendering them non-formulaic (e.g. *fresh breath of some air*). These were embedded in short stories, and subjects read the stories in a self-paced reading task. Conklin and Schmitt reported facilitation (for both native and non-native English speakers) for formulaic versus non-formulaic sequences embedded in short stories. Unfortunately they did not attempt to control for embedded sequence frequency or substring frequency, weakening the relevance of their evidence.

In a similar experiment, N. C. Ellis and Simpson-Vlach (2009) looked at

reading times for compositional formulas and used Mutual Information (MI) and whole-sequence frequency to try to predict reading time performance (Experiment 1). They chose to add MI to the list of predictors because they felt that the MI score captured the coherence of the sequence whereas the frequency of the whole sequence only captured familiarity. They found that MI was a reliable predictor of RT, whereas sequence frequency was not. Interpretation of the study is limited by the fact that they did not control for substring frequency or substring MI and had a sample size of only 11 participants.

The first experiment to use strict substring frequency control was a study of phrase repetition in children by Bannard and Matthews (2008). Taking the usage-based language acquisition stance that was proposed by Tomasello (2003), they wanted to investigate how children process frequent four word sequences, pointing out that in child-directed speech, many sequences of this length occur more frequently than single words. They proposed that children simultaneously use several complementary representations of language at different levels of granularity (morpheme, word, and multi-word). The fact that some sequences are very frequent may explain why children exploit them to build representations. To deal with the issue of substring frequencies in these four-word sequences, they used a technique that was first proposed by Taft (1979) to investigate the contribution of morpheme frequency in regularly inflected words. Taft reasoned that if the frequency of whole forms affected processing independently of the frequency of their components, then people must be storing information about the whole forms. Bannard and Matthews (2008) extended this reasoning to word sequences by hypothesizing that, when they controlled for all component frequencies, any increase in speed or accuracy for higher frequency word sequences would indicate whole sequence storage. They found that the surface frequency of four word sequences predicted the accuracy and speed with which two and three year-olds repeated these sequences and therefore concluded that young children have a stored representation of those sequences. They also pointed out that this is evidence against purely algorithmic processing of word combinations, as proposed by Ullman (2001). The same authors recently used a similar experimental paradigm to look at the effects

of predictability, semantic density, and entropy on performance (Matthews & Bannard, 2010). As in their earlier study, they used stimuli that were matched on all sub-sequence frequencies except the final one, two, or three words. For each of the initial three words they calculated the slot entropy for the final word's slot, using the following entropy equation:

$$H(X) = -\sum_{x \in X} p(x) log_2 p(x) \tag{1.2}$$

where X is a slot, and $p(x)$ is the probability of seeing word x in that position (p. 467). Word sequences with a high slot entropy value have increased uncertainty for what word goes in that slot. Such sequences might be easier to generalize and therefore be easier to process for children than sequences with lower slot entropy. They also calculated HAL-like co-occurrence vectors (Burgess, 1998) for all the words that filled these slots and measured the similarity of all these vectors using the cosine function. They then calculated a measure of distributional homogeneity they called semantic density, which is the mean pair-wise distance of all the possible completions in that slot. They hypothesized that sequences with many semantically similar completions (e.g. *back in the* _____) might be processed differently from sequences with semantically diverse completions (e.g. *a piece of* _____). Since the ten sub-string frequency measures for each four word sequence were highly multi-collinear, they extracted four orthogonal dimensions using principal component analysis (PCA), and entered these four principal component factors as predictors. Their two- and three-year old subjects were more likely to correctly repeat unfamiliar sequences with high slot entropy. The linear mixed effects model they used included the PCA-based frequency factors, age, and semantic density, but semantic density was not a strong predictor for unfamiliar sequences. For familiar sequences, the reverse was true: semantic density was a strong predictor of accuracy but slot entropy was not. The authors were not able to provide any explanation of why familiar sequences should show this effect, and warn that it could have been due to peculiarities in their very small stimuli set (only nine of the 27 stimuli were of this category). Their final conclusion was that *n*-grams do have some psychological primacy across the life span, and

play a role in language behaviour at the ages of two and three years.

The first study to look at the role of sequence frequency in adults was recently reported by Arnon and Snider (2010). The goals and the rationale they used were identical to those of Bannard and Matthews (2008) in that they were looking to find speedier processing of sequences when they controlled for component frequency. Another goal was to compare the threshold theory of *n*-gram processing (advantages only for sequences of a frequency above a particular frequency) versus a continuous effect of frequency (incremental advantages for all sequences spanning the full frequency spectrum). The stimuli set was made up of matched pairs of four word sequences from three different frequency bands (High, Medium and Low) measured in a twenty million word corpus of transcribed spontaneous speech from telephone conversations. Twenty-six subjects performed a timed plausibility judgment task on the sequences, with some of the items being nonsense (e.g. *I saw man the and jump during the pool*) and the rest being matched pairs occurring with high frequency (e.g. *Don't have to worry*, 15.3 times/million) or low frequency (e.g. *Don't have to wait*, 1.5 times/million). To avoid repetition effects, each member of a matched pair was presented in a different block, with a 5-minute filler task between blocks. Despite these efforts, they did find a strong block effect in the reaction times. They used linear mixed effects models to analyze their data, included the frequency (coded as high or low) as a categorical predictor. Models that included this dichotomous frequency predictor fit the data better than models that didn't include it. In a separate analysis that included data from all three frequency categories, they used the sequence frequency as a continuous predictor. Models that used dichotomous frequency predictors fit the data poorly when compared to the models with continuous frequency predictors, leading them to conclude that sequence frequency influences processing across the frequency spectrum.

Finally, the most recent work in this area was a study of reading times of lexical bundles by Tremblay, Derwing, Libben, and Westbury (2011). As in the first analysis done by Arnon and Snider (2010), they created a dichotomous category for their stimuli. Sequences that occurred more than five times

per million (e.g. *in the middle of the*) were hypothesized to have a processing advantage over less frequent sequences (e.g. *in the front of the*). Their choice of stimuli extends the type of sequence beyond those used byBannard and Matthews (2008), Arnon and Snider (2010), and Matthews and Bannard (2010), who all limited their stimuli to sequences that were either constituents (verb phrases, noun phrases, prepositional phrases) or intonational phrases, who all compared sequences that differed only in their final word, and who all displayed the sequences out of sentential context. Tremblay et al. hoped to show that frequency effects could be found for stimuli that were not restricted in that way. They used only non-constituent sequences, matched them on words in the non-final position, and embedded the sequences in valid sentences such as *I sat in the middle of the bullet train.* To investigate sequence reading performance, Tremblay et al. (2011) performed three self-paced reading experiments: word-by-word reading, portion-by-portion reading and whole sentence reading. In all of these self-paced reading experiments, HF/LF was a strong predictor of reading speedup.

One question that arises is: Does the ease of processing that they found for HF sequences come from practice effects during reading without any benefit of a stored representation, or does it come from a stored representation that exists because of their frequency? To tease these sources of facilitation apart, Tremblay et al. created a memory task to look for recall advantages for *n*-grams in working memory. Their rationale was that if HF sequences have a holistic stored representation, the memory load for these sequences should be smaller than for LF sequences, which should not be expected to have such a representation. Two memory experiments were done: one in the auditory modality and another in the visual modality. Each sequence was followed by a list of six random words. Participants heard these read by a speech synthesizer, and were asked to immediately type in all that they had heard. The two dependent measures were accuracy of sentence recall and accuracy of word recall. In the auditory modality the HF/LF predictor explained the difference in sentence recall accuracy, but it did not explain the number of words recalled in each of the trials. In the visual modality, the HF/LF predictor

had a reliable effect for both sentence recall and number of words. These results are evidence for stored representations of high frequency sequences. However, there are some issues with the design of the experiment that need to be addressed (as acknowledged by the authors): the factorial design creates problems for continuous measures such as frequency, and lack of control for substring frequency could be a confound.

1.3 Methodological issues in studying *n*-grams

The research on *n*-grams that I have discussed has involved many different paradigms and methodologies with differing levels of stimulus selectivity. Future research should aim to avoid the pitfalls of previous experiments and take into account the experimental variables that are known to affect performance. To aid in this process, I present a list of methodological issues that should be considered during experimental design and analysis.

Frequency and collinearity

Tremblay et al. (2011) used a single categorical frequency variable to describe their n-grams (one they called lexical bundles, defined as having a whole n-gram frequency greater than or less than 10 occurrences per million words for 4-grams). The key difficulty with this approach is that power is lost when a continuous variable is arbitrarily dichotomized (Baayen, 2010b). Since frequency is a covariate and not a treatment variable, correlational designs are much more appropriate for any studies involving stimuli made up of lexical sequences. Some of the most common stimulus covariates that are taken into consideration in single-word lexical processing experiments are orthographic frequency, orthographic neighborhood size, mean bigram frequency, imageability/concreteness, length, and age of acquisition. Of these covariates, orthographic frequency often absorbs more variability in statistical models than the others. For this reason, many experiments attempt to constrain their stimulus sets so that these covariates do not reliably differ, other than for a single manipulated factor. The danger in relying on this type of stimulus selection is

that the number of items that satisfy these constraints may be so small that the results can no longer be claimed to be applicable to any other words, as the stimulus set is a nearly exhaustive list of the items in the language that meet the criteria. Baayen and Hendrix (2011) point out that Arnon and Snider (2010) should not have considered their stimuli to be generalizable beyond a very small set of n-grams for this very reason. The potential list of stimuli that fit these requirements is extremely small compared to the full set of n-grams. Baayen and Hendrix (2011) argued that case items must be treated as a fixed effect rather than a random effect. The reason that Arnon and Snider (2010)employed their constraints (matched n-grams must not vary in frequency except for a two frequency variables, the final word and the whole 4-gram frequency) was that they wanted to avoid undue influence of frequency variability in the stimuli on the reading time they measured. The covariation of frequency is the most important issue to be dealt with, and I will now discuss it in more detail. When moving from words to lexical sequences, the issues of stimulus covariates are magnified: the powerful effect of frequency now has many components. Each n-gram can be described by a large set of frequencies: the whole n-gram frequency, the sub-component n-gram frequencies, and the word frequencies. For example, the 3-gram *the blue lagoon* has seven frequencies associated with it:

- One whole n-gram (*the blue lagoon*)
- Two 2-grams (*the blue* and *blue lagoon*)
- One non-contiguous 2-gram (*the lagoon*)
- Three single word frequencies (*the*, *blue*, and *lagoon*)

With longer n-grams, the number of frequency measures grows larger. As of yet, the individual predictive properties of these frequencies are unclear and their interactions have yet to be explored. What is known is that these frequencies are often collinear. This makes it problematic to enter them into statistical models that assume uncorrelated predictors. This problem is highlighted by a reanalysis of Arnon and Snider's (2010) stimuli by Baayen and

Hendrix (2011). They found that by applying a very powerful statistical model, a generalized additive model, and added non-linear relationships between the frequency variables and the reading time, a complex interaction of 4-gram frequency and 4th word frequency emerged. Baayen and Hendrix (2011) found a facilitatory effect of n-gram frequency across the full range of fourth word frequencies. This effect of fourth word frequency, for a fixed n-gram frequency, was non-linear and inverse U-shaped – the greatest response latencies were predicted for intermediate fourth word frequencies. This is the danger of attempting to eliminate the effect of frequency covariates by matching using only a subset of these covariates.

The methods used by Matthews and Bannard (2010) to mitigate the influence of n-gram frequencies offer one solution to this issue. They used principal component analysis (PCA) to transform the set of multi-collinear frequency covariates into uncorrelated set of principle components (PCs). They also tested these PCs for collinearity before entering them into their regression model to make sure that they were not confounded with each other. (Matthews & Bannard, 2010, Appendix S2) (Matthews and Bannar, 2010, Appendix S2). This method of analysis makes sense, and is not difficult to do. If these PCs are entered into statistical models, their contribution can be easily understood and taken into account.

Tremblay and Baayen (2010) chose to include all the component n-gram probabilities into their statistical models without de-correlating them because they found that the collinearity did not effect the outcome of their analysis. They then added many other covariates, in particular the log conditional probabilities (the log-transformed probability of obtaining a 4-gram given one of the contained 3-grams for example.) They eliminated all frequencies that did not improve the fit of their multiple regression model with mixed effects. This demonstrates another approach to the problem of including frequency information into statistical models: show how the complete set of covariates influences the outcome, and demonstrate the null effect of multi-collinearity.

A more general statistical issue exists for experimental designs that involve n-grams: the statistical models that are chosen must be suited to the data

collected. In most experiments discussed in this chapter the designs include many fixed and random effects, both nested and crossed. Instead of merely dealing with group averages, trial-level predictions make much more sense for these types of experiments. For this reason, many analyses have used linear mixed effects models that contain crossed random effects for items and subjects (Baayen, 2008). Applying the by-item and by-subject ANOVA or multiple regression without dealing with the crossed random effects structure of the model leads to loss of power and other statistical problems (Baayen, 2008).

There is a final methodological issue that needs discussing: the volatile nature of this area of investigation. With each new model presented, multiple derivative statistical measures are being proposed. As we have seen, conditional probability (Tremblay & Baayen, 2010), entropy (Matthews & Bannard, 2010) , and mutual information (N. C. Ellis & Simpson-Vlach, 2009) have all been used to explain variations in response to *n*-gram stimuli. Whenever possible, researchers should attempt to calculate and enter these competing measures into their models and see if there is any evidence that they are reliable predictors. It is difficult to keep track of these many new statistical measures, but without a consistent focus from all investigators in the field it will be very difficult to discern the informative measures from the less informative ones.

1.3.1 Models of probability, expectation and predictability in language

In this final section before concluding, I consider how the evidence for the existence of *n*-gram effects fits with extant language models. Most psycholinguistic models attempt to predict behaviour in word reading or sentence reading; relatively few address word sequences. However, there are some models that indirectly implicate sequence processing because they have the capacity to simultaneously take into account frequency at multiple granularities. I will focus on these models.

Some of the first models that considered word sequences were built by people investigating sentence processing. Mitchell, Cuetos, Corley, and Brysbaert (1995) created an exposure-based model that looked at fine-grained versus

coarse-grained analyses of sentences. In this kind of parsing research, the goal is to model behaviour for sentences with ambiguous parses. They found evidence against exclusively fine-grained (lexical) record keeping, as well as abundant evidence suggesting that the statistical information used to resolve ambiguities is based on counts using categories that are higher than the lexical level. Mitchell et al. (1995) did not specifically mention word sequences, but their work provides a pathway towards a sequence-processing model.

Jurafsky (1996) created a probabilistic model of lexical access and disambiguation that gets closer to sequence modeling. He proposed a single probabilistic algorithm that modeled both the access and disambiguation of linguistic knowledge. The algorithm was based on a parallel parser that ranked constructions by their conditional probability using a Bayesian or evidential access algorithm that accounted for data from access and disambiguation experiments by ranking constructions according to their posterior probability given the evidence. This type of model is interesting because it unifies lexical and supra-lexical information in a single probabilistic model. In a similar fashion, Levy (2008) built an expectation-based model of sentence comprehension that also included probabilistic information. Levy looked at a construct called the surprisal of words in sentences. Surprisal, also called self-information, is a measure of the information content associated with the outcome of a random variable (Shannon, 1948). Levy created a parallel model using a constraint-based, resource-allocation paradigm of ambiguity resolution. Like the work of Mitchell et al. (1995) considered in the last paragraph, the ideas in these models were not directly applied to word sequences, but may help us understand how to model them.

Bell, Brenier, Gregory, Girand, and Jurafsky (2009) created a model of speech production that integrated sequence frequency. In their regression study of conversational speech, they found that frequency, contextual predictability, and repetition each made separate contributions to the length of time it took to produce words. They also found that content-and function-word durations were affected differently by their frequency and predictability. Content word utterances had shorter durations when more frequent, while

function words had no change in their duration. Both content and function words were influenced by predictability from the word following them. Sensitivity to predictability from the preceding word was largely limited to very frequent function words. This evidence supports the use of probabilistic information in planning articulation for speech but does not use any formal mathematical models to predict word duration.

Automated speech recognition (ASR) is one domain of language studies in which statistical models of language have become the dominant tool. Statistical methods, especially stochastic processing with hidden Markov models (HMMs), were introduced to the field in the early 1970s. HMMs are the only way ever found for computers to perform accurate ASR (Baker et al., 2009). HMMs are trained from corpus data, and they learn to recognize conditional probabilities in the input speech signal despite the high variability of speech. HMMs can link the acoustic signal to candidate sequences, and ASR systems include complex sequence models that increase the performance of the systems.

Google used probabilities from the world's largest set of *n*-grams to beat twenty other translation systems in translation accuracy at a competition conducted by the Speech Group of the US National Institute of Standards and Technology's Information Access Division (Geer, 2005). This type of purely statistical machine translation depends on simple models and large datasets, whereas most other systems use complex sets of grammatical rules to translate between languages. Although these two examples of word sequences being used to help machines to perform linguistic tasks do not bear directly on the question of human representations, they do provide suggestive evidence of the utility of the information available in word sequence probabilities.

Another type of probabilistic model that deals with the temporal dependencies of words is the connectionist Simple Recurrent Network model (SRN; Elman, 1990). This model, extended by Loewenstein, Tabor, and Tanenhaus (1999), uses a hidden layer of nodes that take the current state of the network as input (hence the recurrency). SRNs have the ability to encode probabilities of events across time, making them good candidates for understanding *n*-gram effects. They have been trained to predict the next letter in a newspaper arti-

cle corpus after being trained on similar text (Rodriguez, 2003). These models assume that linguistic units are emergent consequences of a learning process operating over the latent structure in the language stream. The lack of other assumptions in this type of model allows *n*-gram units to emerge from the statistical information in the language stream, making it one of the few models that does. Words emerge from a stream of letters, and sentences emerge from a stream of words.

Finally, a recently proposed model by Baayen, Milin, Djurdjevic, Hendrix, and Marelli (2011) is perhaps the best exemplar of the power of modeling language using statistical learning algorithms. Their model uses the standard learning equations of the Rescorla-Wagner model at equilibrium (Danks, 2003) to model morphological processing effects. They go on to extend this model beyond single words and compounds and show how it is also able to explain the phrase frequency effects found by Arnon and Snider (2010). Remarkably, Baayen et al's model contains no representations that correspond to whole words or whole phrases, only letter unigrams and bigrams. Nevertheless, the statistical regularities observed by the model allow it to learn any regularities from a language stream. This is the reason they call their model a naive discriminative learning framework: the model begins with no information about the language, and builds a precise representation of the language by learning how form maps to meaning without using any explicit rules for parsing. The model applies the basic principles of discriminative learning to the problem of mapping form to meaning. They define the activation of a linguistic meaning as the model's estimate of the posterior probability of that meaning given its unigrams and bigrams and the co-occurrence probabilities of these unigrams, bigrams, and meanings. The simplicity of this model and its success at modeling a wide variety of phenomena powerfully demonstrate that probabilistic information processing can help explain data from experiments on n-gram processing.

1.4 Beyond rule-based descriptions of language

We began this chapter by arguing that there are two theoretical foundations to studying *n*-grams: information theory and learning theory. I have reviewed evidence showing that information-theoretic measures predict human behavior on a wide range of language-related tasks, and reviewed a number of models that allow for the learning of this information without assuming the existence of any language-specific machinery. The study of *n*-grams can seem like a curiosity, since it largely ignores the sentence structure and the special status of grammatical phrases that many believe are what makes human language special. However, a thorough investigation of *n*-gram effects may result in a radical re-conceptualization of human language in the near future – indeed, I believe that this re-conceptualization is already underway. I conceive of language as a probability calculation across multiple levels of granularity (including second-order co-occurrence, which is shared context), from phonemes or letters to sentences. This conception allows several thorny problems – most notably, the problem of the [alleged] poverty of the stimulus and the closely related problem of how we can understand sentences we have never encountered (Plato's Problem, see Landauer & Dumais, 1997)- to simply evaporate. The success of Google's *n*-gram based machine translation software and Baayen et al. (2011) NDR model of language learning demonstrate empirically that relatively simple and algorithmically well-defined processes of probability maximization could underlay much or all of language.

More generally, putting language on such a basis also allows us to see a way to overthrow the philosophical dominance that rule-based models of language processing have held for so long. I believe that rules are useless as explanatory devices because (as David Bloor has written in a related but different context): "Verbalised principles, rules and values are the phenomena to be explained. They are dependent, not independent variables." (Bloor, 1983, p. 137). Linguistic rules are high-level post-hoc descriptions of language. The role of language scientists must not be merely to describe, but to explain, and to do so with as few "special mechanisms" as possible. Statistical models of

language that build on widely-accepted general learning principles best satisfy the demands of Occam's Razor, that the number of theoretical entities not be multiplied beyond necessity.

One potential impediment to understanding word sequences is the predominant concept of the mental lexicon. Lexical sequence representations do not fit well within the standard models of a mental lexicon because these symbolic stores cannot easily accommodate the information contained in non-lexical entities. Sequences do, however, fit very well into Elman's framework of lexical knowledge without a lexicon (Baayen et al., 2011; Elman, 2009, 2011). As Elman notes: "In this scheme of things there is no data structure that corresponds to a lexicon. There are no lexical entries. Rather, there is a grammar on which words operate. Crucially, the system has the capacity to reflect generalizations that occur at multiple levels of granularity." (Elman, 2009, p. 566) The multiple interactions of lexical context within a word sequence can be seen as an emergent property of the system. I believe that this type of process model shows the most promise for explaining *n*-gram effects in language.

This position is a controversial one: if there is no lexicon and no syntactic rules to be implemented in the brain, what about all the language behaviour that so obviously grammatical? There must be a mechanism that allows us to produce and comprehend longer, more complex sequences, and it is possible that our prodigious memory for *n*-grams and our ability to generalize patterns using probabilistic smoothing may not fit the bill, but pushing these simple models to their logical conclusion is undoubtedly worth the effort.

1.5 Summary

I have reviewed from various experimental paradigms that help justify more empirical work as well as more theoretical development of the role of *n*-grams in psycholinguistics. Even if the definition of an *n*-gram is limited to those sequences that occur frequently, the number of word sequences is very large. It may seem that retaining information about so many combinations of words would be highly inefficient and a waste of finite mental resources. In addition, a

mental representation of sequences implies that linguistic competence is built from experience and nothing else. This implication was seen as counter to psycholinguistic reality until very recently. With the shift currently taking place in science to massive-data-driven inference, these ideas for alternate ways of representing language have started to take root. By analyzing enormous corpora and computing very large numbers of word transition probabilities, for the first time we have a chance of approaching the issue of *n*-gram processing.

All this does not, of course, mean that human language is not special. Clearly, human beings are able to pull more information out of the linguistic stream than other animals, since human children are the only animals who have ever become fluent language users (or even just language comprehenders) from exposure to that stream. The language faculty is no more and no less than a product of a quantitative improvement in the basic cognitive capabilities that are seen in other organisms, rather than a qualitatively different process that is seen in no other communicative organisms. This viewpoint is, I believe, most consistent with the development of an evolutionarily grounded, biologically-plausible scientific approach to the study of language. In the terminology of semiotics, words are a type of sign. All *n*-grams are signs, but they are signs built from simpler signs.

1.6 Three experimental studies

In the following three chapters I will report three lines of research that I have conducted on *n*-gram processing. These studies use various stimuli, methodologies and experimental designs, but are unified by their goal of better understanding how we process *n*-grams.

The first line of research, presented in Chapter 2, investigates the perception of the subjective frequency of *n*-grams. Next, in Chapter 3, I investigate the process of *n*-gram comprehension by looking at people's eye movements while reading *n*-grams. Finally in Chapter 4, the focus will be on production rather than comprehension as I look at the words participants choose to complete *n*-grams in a cloze task. These three chapters may be read in any order –

the content of each chapter is independent from the others. In the final chapter I will synthesize the results from these three disparate research projects and build a coherent picture of a new understanding of n-gram processing.

Chapter 2

The subjective frequency of *n*-grams.

When asked to assign think about the subjective frequency of an *n*-gram, what properties of the *n*-gram influence the respondent? *N*-grams that were more frequently found in a corpus of English were read faster than less frequent *n*-grams, an effect that is analogous to the frequency effects in word reading and lexical decision. The subjective frequency of words has also been extensively studied and linked to performance on linguistic tasks. I investigated the capacity of people to gauge the absolute and relative frequencies of *n*-grams. Subjective frequency ratings were collected for 352 *n*-grams. Their subjective frequency ratings showed a strong correlation with corpus frequency, in particular for *n*-grams with the highest subjective frequency. These *n*-grams were then paired up and used in a relative frequency decision task (e.g. Is *green hills* more frequent than *weekend trips*?). Accuracy on this task was reliably above chance, and the trial-level accuracy was best predicted by a model that included the ratio of corpus frequencies of the whole *n*-grams or the ratio of the frequencies of the component *n*-grams. These results support models of reading that posit traces in long-term memory for *n*-grams as well as words, models that take advantage of the probabilistic information in each *n*-gram.

2.1 Introduction

What are the grain sizes of language that are represented by our minds? The word-sized unit has been the dominant size for most psycholinguistic research, with the next largest unit being the sentence, which is made up of words. There has been as of yet little work on groups of words called *n*-grams (Shaoul & Westbury, 2011). *N*-grams are any combination of two or more words, and are not restricted to complete, compositional phrases (both *the red hat* and *the*

hat that are *n*-grams). Any stream of language can be broken down into its component *n*-grams in the same way that a word can be segmented into morphemes or phonemes. *N*-grams have similar properties to other units: each *n*-gram will have a probability of occurring at any point in time, and that probability will depend on the context. The probability of any *n*-gram occurring can be estimated from its frequency of occurrence in a corpus, and the larger the corpus, the more accurate the estimate (Kilgarriff & Grefenstette, 2011). *N*-gram probabilities throughout this chapter will derived from frequency information found in a one trillion-word corpus of English web documents created by Google (Brants & Franz, 2006). These probabilities have the potential to explain aspects of language behaviour that are beyond the reach of non-probabilistic psychological models of language.

Some theories of language predict that there should be no effects for the transitional probabilities of words in sentences or *n*-grams (Harris, 1951; Chomsky, 2005). In a generative framework of language, the long-term memory system is not necessary for lexical sequence processing. This could be dismissed as being theoretically unimportant for a system of rule representations, but there is no point in differentiating *competence* from *performance* in empirical psycholinguistic research. The power of the grammar/rule systems are able to do all the heavy lifting without the need to gather probabilistic information about words and word combinations (*n*-grams). Ullman (2001), for example, describes language as a mental lexicon of memorized words that are arranged by the rules contained in a mental grammar. The procedural operations in this model, and others like it, assemble larger structures from hierarchical compositions of smaller structures (morphemes into words, words into sentences). When these compositions are fully productive (e.g. *walk* – *walked* or *ideas* – *green ideas*), they are posited to be purely rule driven. Any effects of *n*-gram probability or frequency are inconsistent with these models because the unfolding of abstract rule processing operations should not be affected by the amount of experience with a stimulus. In some recent work on sex differences in language processing, Ullman, Miranda, and Travers (2008) noted that "women depend more on lexical/declarative memory for the processing of complex lin-

guistic forms, while men tend to rely more on the rule-governed combination of these forms in the grammatical procedural system" (p. 301). This would imply that humans universally use two distinct systems to process language, a procedural system and a declarative system, and that women depend more on one system than the other.

Compositional semantics is another area that has seen attempts at rule-based theories of representation and processing. Jackendoff (2007) has offered models that build semantic combinations from a set of lexical items and relationships, but the empirical validations of this model is not forthcoming. The assumption of this and other semantic models is that the grain-size of language is the word, and larger structures are based on operations on words, similar to the syntactic dualism of words and rules (Pinker & Ullman, 2002).

N-grams and words share many properties, somehow represented as entries in a lexicon, and that there is a search process across this lexicon as proposed by Forster and Hector (2002)

The inherent unwieldiness of dualist models has spurred demand for more parsimonious model that can explain our linguistic capabilities. These *emergentist* theories of language propose that experience is used to build representations of linguistic patterns without any need for systems of grammatical rules (Baetes & Elman, 1993; Elman, 1990; Tomasello, 2003; Goldberg, 2006; Bod, 2009; Dilkina, McClelland, & Plaut, 2010a; Baayen et al., 2011; Frank & Bod, 2011). Why use the word *emergent* to describe language? A spirit of *reductionism* has long been at the core of many theories of language (e.g. a word is just the sum of its spelling, sound and meanings). Rather than understanding the whole by studying the parts, these new theories attempt to generate the properties of the whole by understanding the parts. My definition of the emergentist school of thought is broad and inclusive, but the trait that links these models is consistent: these models all include linguistic context and linguistic content and allow context and content to interact.

The following summary of current research on *n*-gram processing provide evidence for broad, probabilistic effects of linguistic experience on language processing task, in turn providing support for this emergentist school of

thought.

In the last few years there has been an increase in the number of studies reporting such probabilistic effects, in particular *n*-gram frequency effects. Bannard and Matthews (2008) studied children's production of *n*-grams, and found that *n*-gram frequencies influence their accuracy when children repeat back short phrases that differ only by one word. Arnon and Snider (2010) replicated this effect using similar stimuli, a reading task and undergraduate student participants. They found that participants read the more frequent *n*-grams faster than the less frequent *n*-grams. In both studies the effect was not due to the frequency of the individual words or substrings and it was observed across the entire frequency range (for low, mid- and high frequency *n*-grams).

Matthews and Bannard (2010) found more accurate production of *n*-grams when the experimenters asked 2 and 3-year olds to repeat them, even after controlling for multicollinearity in the frequency measures. In the studies mentioned so far, the authors limited all of their stimuli to *n*-grams that were *constituents* (verb phrases, noun phrases or prepositional phrases) or *intonational phrases*, meaning that they did not cross over traditional phrase boundaries. The first study to look at reading times for *n*-grams that were sampled without imposing any restrictions on phrasehood was done by Tremblay et al. (2011). They used only non-constituent *n*-grams in a self-paced reading experiment and found that there was a whole *n*-gram frequency advantage. Tremblay and Baayen (2010) followed up with an ERP study for an immediate free recall task for sets of three non-constituent 4-grams. They found that whole *n*-gram probability as well as internal word and 3-gram frequency predicted recall as well as P1 and N1 amplitudes. These results suggest that *n*-gram frequency is contributing something to the language system, and that *n*-grams representations may operate alongside word representations.

Eye tracking experiments have also been used to look at *n*-gram frequency effects. Siyanova-Chanturia, Conklin, and Heuven (2011) presented subjects with two types of 3-grams: binomial phrases (*bride and groom*) and those same phrases reversed (*groom and bride*). These two types of *n*-grams are naturally very closely matched on many lexical variables, and they proposed that any

differences in processing must arise from effects of *n*-gram frequency. The binomial 3-grams had an average frequency in the BNC that was 10 times that of the reversed 3-grams (2.473 per million versus 0.274 per million). Thirty 3-grams of each type were embedded in sentences and read by participants in the eye tracker. They found that binomial phrases were read faster than reversed phrases. They also found that phrasal frequency facilitated reading even after taking into account the effect of phrase type, more evidence that increased exposure to an *n*-gram contributes to its entrenchment.

If language is being represented as a stream of words and *n*-grams of different lengths, it follows that we should be able to see implicit learning of word sequences. Interestingly, Remillard (2010) recently reported that subjects were able to implicitly learn 5^{th}-order and 6^{th}-order sequential probabilities of certain non-linguistic stimuli. In their experiment they taught their participants to push one of six buttons corresponding to the location of a box on the screen. After two sessions of training spread over two days, subjects showed improved speed and accuracy in their responses. After 16 sessions of training were completed, participants were able to reliably predict the 5th element of a sequence based on the conditional probability of the previous four elements. This result provides some support to the idea that implicit learning of *n*-gram transitional probabilities for 2, 3 , 4 and 5-grams is feasible.

In a related line of research, implicit sequence learning ability has been shown to be linked to performance on language processing tasks by Conway, Bauernschmidt, Huang, and Pisoni (2010). They looked for individual differences in their participants' perceive degraded speech, a task that is highly dependent on the ability to predict upcoming words based on context. They found that a reader's sensitivity to sequential structure during implicit learning was the best predictor of these individual differences, even after taking into account their performance on tasks measuring short-term and working memory, attention and inhibition, and vocabulary.

Moving beyond orthographic frequency, other probabilistic measures are now being studied. Tremblay and Tucker (2011) investigated the influence of two additional measures, conditional logarithmic (log) probability, and Point-

wise Mutual Information (PMI), on the recognition and production of 4-grams. Conditional probability is a measure of likelihood of seeing a word given a specific context, or predictability. PMI is an index of how strongly words are associated with each other and is calculated by dividing the probability of the whole *n*-gram by the product of the individual word probabilities. They asked participants to read 432 4-grams as quickly as possible after viewing them and they recorded the onset time (the time taken to read the 4-gram and prepare for the articulation) and duration of the utterance (the time to articulate). *N*-gram frequency was found to explain much more of the deviance in production durations than conditional probability or PMI, leading the authors to conclude that *n*-gram frequency relates to the fluency of production due to entrenchment from exposure. Recognition time, as measured by the onset latency, had more deviance explained by conditional probability and PMI, with a smaller contribution from frequency. This implies that the degree of competition between *n*-gram family members is the main process underlying recognition, which dovetails nicely with recent work on competition-based models of recognition of compound words (Juhasz & Berkowitz, 2011; Kuperman, Schreuder, Bertram, & Baayen, 2009). In terms of which length *n*-gram contributed most to explaining deviance in onset latencies, probabilistic measures for the 3-grams were strongest, followed by unigram probabilities. For production duration, unigram probabilities were the dominant measure in reducing deviance. Tremblay and Tucker propose that the 3-gram is a key unit of language that is long enough to contain complex meaning, but short enough to be processed efficiently. This pattern of results points to a complex, dynamic system, with information from internal *n*-grams influencing the processing of the wholes.

These studies all provide evidence for general *n*-gram frequency sensitivity, using different types of stimuli and different experimental paradigms. Is frequency purely a measure of the familiarity of an *n*-gram? Frequency effects can be also be thought of as complex phenomena that arise from more than just pure exposure. The key realization is that repetition implies contextual diversity, and so repetition itself may not be what gives high frequency *n*-

grams their advantage (McDonald & Shillcock, 2001). Frequency is inevitably correlated with many other measures. McDonald and Shillcock (2001) identified *contextual distinctiveness* (CD) as a measure that can explain effects of orthographic frequency. CD was expressed as the relative entropy between a word's context and the context for all words in the language.

In a similar vein Baayen (2010a) calculated the contribution of 17 lexical variables from many categories : frequency, genre distribution, CD, syntactic entropy, morphological entropy, and orthographic features in predicting LDRT. Once the other predictors were used to predict RT, orthographic frequency did not contribute to the final model. This idea could be called the *frequency-effect-as-epiphenomenon position*, another case of frequency effects emerging from models that do not use lexical frequency counts. In these experiments a key covariate used is *n*-gram frequency, but it is critically important to state that frequency itself is at the heart of the process, but rather other probabilistic measures of *n*-grams that we do not yet have access to, such as those mentioned above, are involved. The models used here can tell us much despite the fact that they are simpler and do not include the covariates mentioned above.

Over time *n*-grams do become more familiar. This feeling of familiarity with a word sequence (its *subjective frequency*) must come into play when reading *n*-grams. This study aims to delve deeper into the question of *n*-gram subjective frequency and to better understand what is driving these varying degrees of sequence familiarity.

The first question to be addressed in this work is: How does the probability of an *n*-gram in a large corpus of text relate to the subjective frequency of the *n*-gram? In the first part of the chapter I will attempt to detect any contribution of *n*-gram frequency to subjective frequency ratings. This evidence will provide a basis for *n*-gram probability in the formation of *n*-gram familiarity. The second question addressed is: How sensitive is the language system to the relative probabilistic information contained in language? Subjective frequency judgements are by definition absolute (from *VERY FREQUENT* to *VERY RARE*), but relative frequency judgements change depending on what *n*-grams are being compared. Comparing two very common *n*-grams may be

different from comparing two very uncommon *n*-grams. Yet relative frequency judgements should tap into the same implicit familiarity knowledge that is used to generate subjective frequency ratings. In the second part of the chapter the impact of *n*-gram probability on subjective relative frequency judgements is investigated. Will there be an impact of the frequency of the internal *n*-grams, the whole *n*-gram or both? My goal is to better understand how the probabilistic information contained in *n*-grams influences their processing.

2.1.1 Words and *N*-grams

One theme in this research is the similarities between *n*-grams and words. Evidence for this conjecture has come from many sources. Kuperman, Bertram, and Baayen (2008) studied compound words, and found that compound word frequencies, constituent lexeme frequencies, and conditional probabilities for all the morphemes in the compound word had a role to play in their model of compound word reading. Compound words are in many ways similar to 2-grams, leading me to speculate that models of *n*-grams may need to take similar information into account. Since *n*-grams have been shown by Arnon and Snider (2010) and Tremblay et al. (2011) and others to have a word-like frequency advantage, it is possible that words and *n*-grams have even more in common. I will first look at subjective frequency, a well studied aspect of word knowledge.

2.1.2 Subjective and objective frequency of words and *n*-grams

The subjective frequency of words has been investigated by psycholinguists since the 1960s (see Gernsbacher, 1984 for a review). Connine, Mullennix, Shernoff, and Yelen (1990) found subjective frequency to be predictive of word naming times when the stimuli were presented auditorily, but found no effect for orthographic frequency in this modality. This led Connine et al. to conclude that objective and subjective frequency effects for words were task and modality dependent. Furthermore, subjective frequency was concluded to be a post-lexical component that was related to ease of production. Balota,

Pilotti, and Cortese (2001) talked about what influences subjective frequency: objective frequency and meaningfulness, as defined by Toglia (2009). They found that meaningfulness was a better predictor of subjective frequency for low frequency words and orthographic frequency was a better predictor of subjective frequency for high frequency words. More recently, Colombo, Pasini, and Balota (2006) used Italian words and found that subjective frequency and meaningfulness explained variance in lexical decision response times, but not in naming response times. Orthographic frequency explained variance for both tasks. Thompson and Desrochers (2009) found lower correlations between the orthographic frequency of low frequency words and their subjective frequencies, replicating the results (Balota et al., 2001), but with French words.

Baayen, Feldman, and Schreuder (2006) attempted to explain the variability in subjective frequency ratings using various objective predictors. They built a statistical model that absorbed more than two thirds of the variance in subjective word frequency ratings using predictors such as orthographic frequency, written-spoken ratio, word category (noun or verb), noun-verb ratio, orthographic neighbourhood density, derivational entropy and inflectional entropy. These predictors are also important inputs into most models of visual lexical decision response time and word naming response time. The parallels between the two sets of predictive variables supports the notion that subjective frequency is an "off-line inverse of visual lexical decision" (Baayen et al., 2006, p. 305).

What is subjective frequency? Subjective frequency is nothing more or less than a self-reported measure that expresses a person's introspective understanding of their amount of exposure to a stimulus. Lexical subjective frequency data is collected by asking people to rate how frequently they have encountered a word. The instructions in these experiments define encounters as hearing the word, saying the word or reading the word. The variance in these subjective frequency norms for words have been used to explain variance in lexical decision tasks, word naming tasks and others. Taking an emergentist stance, I posit that the subjective frequency rating for a word arises from the same emergent process that is in play when we use words – from the in-

teractions of various processes that operate according to very basic principles of non-symbolic processing and representation (Elman, 2011). If *n*-grams are word-like, an *n*-gram's subjective frequency should be available to people during a rating task, just as a word's subjective frequency is available. In my first experiment I collected subjective frequency norms for a set of *n*-grams and then analyzed these ratings to see how strong their relationship to objective frequency was. My hypothesis is that if *n*-grams have a word-like subjective frequency, corpus frequency should be strongly correlated with subjective frequency when the effect of constituent word and *n*-gram frequencies are taken into account. Furthermore the direction of the correlation should be positive (higher ratings for more frequent *n*-grams), and the correlation should be strongest for the most frequent *n*-grams, replicating the results of Balota et al. (2001).

2.2 Experiment 1

There are many data sets available that provide subjective frequency ratings for words (Balota et al., 2001), but there are no previous reports of the collection of subjective frequency norms for *n*-grams. To see if *n*-grams would have a stable, subjective frequency in the same way that words do, I collected ratings and looked for similarities between *n*-gram ratings and word ratings.

2.2.1 Participants

One thousand five hundred and forty eight students at the University of Alberta participated in this experiment in exchange for partial course credit. The mean age was 19.2 years old, 64% were females and 74% of the students were native English speakers.

2.2.2 Methods and Materials

179 pairs of *n*-grams were chosen from the Google Web1T data set (Brants & Franz, 2006): 60 pairs of 2-grams, 43 pairs of 3-grams, 36 pairs of 4-grams and 38 pairs of 5-grams. The *n*-grams were chosen to cover a broad range

of frequencies and relative frequencies. They were also grouped into pairs and matched on the geometric mean of their constituent word frequencies. This was done so that there would be no bias caused by the relative lexical frequency of the items when they were later used in a relative frequency judgement task. Arnon and Snider (2010) chose to only use *n*-grams that were intonational phrases, that is, *n*-grams that sound complete when uttered on their own. The stimuli were not restricted to clausal or intonational units so as to demonstrate that *n*-gram effects are not limited to those types of constructions. The *n*-grams had frequencies ranging from the very frequent (1139 per million, *to the*) to the very infrequent (0.00006 per million, *to know and keep the*). Subjects were given a web-based survey with a seven point scale next to each *n*-gram. The *n*-grams were presented in the same pseudo-random order to all participants. The instructions stated: "Please rate how frequently the phrases below are used. A rating of *almost never* means that the phrases are used very rarely. A rating of *very often* means that the phrases are used very frequently." The two extremes of the scale were labeled, but the intermediate ratings were not labeled. Each person was asked to rate 31 *n*-grams, providing me with approximately 130 ratings per *n*-gram. Each subject also rated the frequency of three nonsense *n*-grams (e.g. *sanity toast blanket*) to confirm that they understood the instructions.

2.2.3 Results

Our participants understood the task I asked of them. The mean rating for the nonsense *n*-grams was $\mu = 1.35$, $\sigma = 0.2$, a very low rating on a scale of 1 to 7. The mean rating for all the sensible *n*-grams was $\mu = 3.83$, $\sigma = 1.07$. The nonsense *n*-gram ratings were removed from the rest of the analyses. I measured rater reliability using the intra-class correlation coefficient (ICC, Shrout & Fleiss, 1979). For all of the sub-groups of subjects who rated the same set of 32 items, all the ICCs were greater than 0.37, and all of the 95% confidence intervals around the ICCs did not include 0, showing consistency in item ratings across participants.

To understand the relationship between the ratings I gathered and the cor-

pus frequency of the n-grams, I had to see if the internal n-gram frequencies were participating in driving the subjective ratings. This task is complicated by the fact that all of these frequencies are highly inter-correlated, and entering all the predictors simultaneously into a regression model could lead me to mis-evaluate the importance of the predictors. Principle Component Analysis (PCA), chosen by Matthews and Bannard (2010) to reduce the multicollinearity of the component frequencies of their 4-grams, was considered as a potential way to reduce multi-collinerarity in this experiment. A disadvantage of PCA is that the orthogonal components that it produces can be extremely difficult to interpret in terms of the original variables. To sidestep the problem of multicollinearity while properly assessing which predictors are most relevant, I made use of *random forests*. Random forests are a type of recursive partitioning algorithm for performing nonparametric regression with a large numbers of predictors (Breiman, 2001). They are a powerful type of Classification and Regression Tree (CART) method, and since they make no assumptions about the types of relationships between variables they have been found to be superior to multiple regression in predicting performance on various tasks (Finch et al., 2011). To understand which of my predictors was important, I measured the conditional importance of each variable in a random forest model and then only used the most important predictors in my regression models. A method for performing this type of conditional importance analysis has been described by Strobl, Malley, and Tutz (2009) in this way: variable importance is assessed by permuting the data in each predictor variable and then testing the model with the permuted variable and the remaining non-permuted variables until all the variables have been permuted. The prediction accuracy of each inference tree in the forest decreases substantially if the permuted variable was involved in predicting the response. The difference in prediction accuracy before and after permuting a variable, averaged over all trees, is one measure of variable importance, the *marginal permutation importance*. An improvement on this unconditional permutation importance measure is the *conditional permutation importance* (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008) in which the permutation importance is conditioned on each of the partitions that arise from

the recursive partitioning in the random forest as a conditioning grid. This conditional variable importance is less susceptible to preferring correlated predictor variables and takes into account both main effects and interactions. In all of my analyses I used the **R** package called **party** (Strobl et al., 2009), and I tested my random forests with several different starting values to make sure that the ranking of variable importance did not change depending on the starting value. I report the results below after confirming that there was no change in the ranking of conditional importance being caused by the initial conditions of the pseudo-random number generator. The results of my analysis are shown in Figure 2.2.3, and can be summarized as follows:

- For 2-grams, the whole *n*-gram and second word frequencies were important.
- For 3-grams, the whole *n*-gram frequency was important, with a smaller contribution from the third 2-gram's frequency. Interestingly, the third 2-gram frequency, **bf3**, is the frequency with which the first and third words appear together, which I call a *split-gram*.
- For 4-grams, the whole *n*-gram, the first 2-gram and the second 3-gram frequencies were important.
- For 5-grams, the first 4-gram and the whole *n*-gram frequencies were important.

Was *n*-gram frequency helpful in predicting my outcome variable? Using the variables identified by the random forest analysis, I created linear models for each size of *n*-gram with and without the whole *n*-gram frequency in each model and then performed a model comparison. I compared the Akaike Information Criterion (AIC, Akaike, 1974) of all the models to determine which one had the best fit. The AIC is a measure of the quality of a model that incorporates both the goodness of fit and the number of free parameters in the model. Nested models with fewer parameters that have a better fit with the

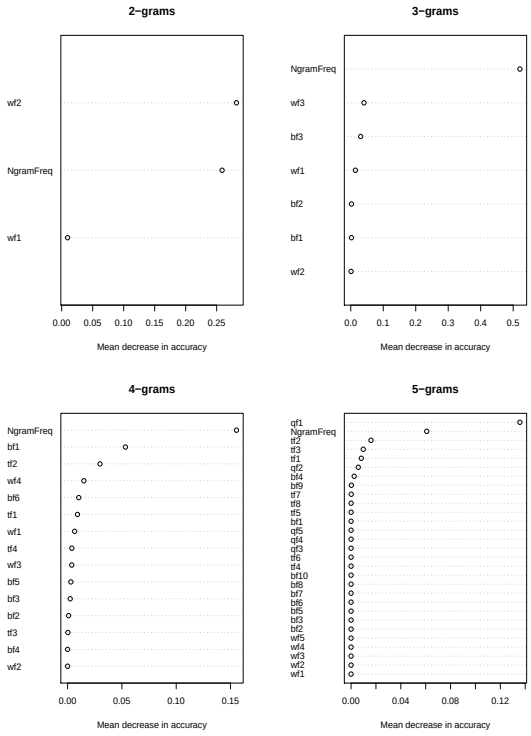

Figure 2.1: Importance for predictors in a random forest model of mean item rating in Experiment 1. After creating random forest models, I calculated the relative importance of all of the log transformed *n*-gram frequency variables in predicting mean subjective frequency ratings adjusted for correlations between predictor variables (both for the main effects and the interactions). The names of the frequencies are abbreviated in the following manner: 2, 3 and 4-grams are assigned the letters **b**,**t**, and **q**. The abbreviation **tf2** stand for **Second Trigram Frequency**. A full description of all these abbreviations is given in Section 2.7.

Table 2.1: Regression Model Comparisons for Experiment 1. Two models for predicting the mean subjective frequency ratings of *n*-grams are given for each size of *n*-gram, with the first model nested within the second. Models in bold type were the best models for each type of *n*-gram. Δdf denotes the change in the number of free parameters between the two models being compared.

	AIC	Δdf	χ^2	p
2-grams: *n*-gram freq only	330			
2-grams: *n*-gram freq and w2f	317	1	15.98	0.00001
3-grams: *n*-gram freq only	212			
3-grams: *n*-gram freq and bf3	206	1	8.07	0.00566
4-grams: *n*-gram freq only	154			
4-grams: *n*-gram freq, bf1 & tf2	145	2	6.81	0.00200
5-grams: qf1 only	172			
5-grams: qf1 and *n*-gram freq	174	1	0.29	0.59329

data are given a lower AIC. This means that the absolute value of the AIC is not important, but rather the difference between two AIC values shows which is better, and how much better. The results of these comparisons of nested models are shown in Table 2.1.

The picture for the relationship between objective and subjective frequency for *n*-grams is more complicated than the one for words described by Balota et al. (2001); it is much more than a linear relationship between the meaningfulness of words or their simple whole form corpus frequency. There were effects of the internal *n*-gram frequencies that came into play. In this section, I will report regression effect sizes using Cohen's f^2, a measure of effect sizes appropriate for regression models. Cohen (1988) suggested that effect sizes of 0.02, 0.15, and 0.35 should be considered as being *small*, *medium*, and *large*. Each model was re-fit 1000 times with bootstrapped replicants giving a distribution of f^2 values. I then calculated the 95% confidence interval of the effect size from this distribution, reported below. For 2-grams, the subjective frequency ratings were predicted by both the 2-gram's frequency ($f^2 = 0.45$, 95% CI 0.32-0.56) and the second word's frequency ($f^2 = 0.07$, 95% CI 0.02-0.13). This result could imply a recency effect: the frequency of the last word read had more impact on the rating than the first word. For the 3-grams, the whole

n-grams's frequency had the largest effect size ($f^2 = 0.45$, 95% CI 0.27-0.59) and there was a weak effect of the split-gram ($f^2 = 0.05$, 95% CI 0.01-0.14).

For the 4-grams, a more complicated model was the best fitting. The whole *n*-gram frequency had the largest effect ($f^2 = 0.34$, 95% CI 0.17-0.51), followed by a weak effect of the first bigram ($f^2 = 0.08$, 95% CI 0.01-0.19) and an unreliable effect of the second trigram ($f^2 = 0.03$, 95% CI 0-0.11).

For the 5-grams, the addition of the whole *n*-gram frequency did not improve the model, so the simpler model prevailed. This simpler model had a strong effect of 4-gram frequency, with the effect size being ($f^2 = 0.27$, 95% CI 0.14-0.43).

In all the analyses above the amount of mulit-collinearity between the predictors was reasonable (in all models, $\kappa < 8$).

Finally, I noted that Balota et al. (2001) had found that the group of words with the highest subjective frequency ratings had a strong relationship between objective and subjective frequency, and that the opposite was true for the words with the lowest subjective frequency ratings. I replicated this result: I performed a median split on all of the items bases on their average subjective frequency rating, and calculated a bootstrapped Pearson correlation with corpus frequency for each of the two groups. The magnitude of the correlations with frequency were larger for the set of items with the higher subjective frequency ratings: for the upper half, $r(177) = 0.22$, 95% CI 0.19-0.48, and for the lower half, $r(176) = 0.11$, 95% CI 0.01-0.19.

2.2.4 Discussion

In this exploratory look at the frequency measures that influence the subjective ratings for *n*-grams I found a complex pattern of evidence for *n*-gram frequency effects. Each *n*-gram size had a different pattern of frequency effects, with no clear, over-arching pattern. The first interesting result was the size limitation seen in the 5-gram data, which could be related to limitations of the short term memory system. Participants were sensitive to the frequency of the first four words of the 5-gram, and nothing else. This implies that the subjective frequency estimation process either cannot use — or does not need to use —

information about the probability of 5 words occurring together to accomplish this task and perhaps other tasks like it. I interpret the results from the 2-grams as showing more of a recency bias since the final word's frequency in the 2-grams had a large impact on the ratings, but the first word's frequency did not. For the 3- and 4-grams, the whole *n*-gram frequencies had very large effects whereas the internal frequencies had relatively weak effects. The key finding in this experiment was that, excluding the 5-grams, there was strong evidence in the 2-,3- and 4-gram data for a dominant effect of whole *n*-gram frequency (for all these effects, Cohen's $f^2 > 0.34$) and a subordinate effect of the sub-frequencies (for all these effects, Cohen's $f^2 < 0.08$). This supports my hypotheses about the sources of implicit frequency judgements, and provides the justification for the next experiments on relative subjective frequency estimation reported below.

2.3 Experiment 2

2.3.1 Relative frequency of words

With evidence from my subjective frequency rating task pointing towards *n*-gram frequency effects for subjective frequency ratings, the next place I looked for effects was in a more complicated task: relative frequency judgements. All rating tasks are limited by the use of absolute Likert scales, which are not immune to artifacts (for example, I have treated ordinal-scale data in the ratings as if they were interval-scale, using the mean ratings instead of the mode, etc.). To avoid these issues, I chose to develop a relative frequency task that does not suffer from the same issues. In this type of task the participants are shown two items at once and are asked to judge which one of them, in their experience, is more frequent. I manipulated the relative corpus frequency of the items, both in absolute terms (low frequency vs. low frequency, high frequency vs. high frequency) and in relative terms (a very small difference in frequency relative to each other or a very big difference). The power-law distribution of *n*-grams provides ample examples of items that fall into all of these categories, and the stimuli were chosen to cover a broad swath of the frequency spectrum

to make sure that my results were generalizable to a majority of *n*-grams.

Before attempting this task with *n*-grams, I sought to confirm that a relative subjective frequency judgement task was reasonable and feasible with simpler stimuli. I created a single word task that I could later extend to *n*-grams, and looked for evidence that my paradigm was valid for investigating relative frequency judgments.

2.3.2 Participants

Thirty-three students from the University of Alberta participated in this experiment in exchange for partial course credit. The mean age was 19.4 years and 57% of the participants were females. All were right-handed native English speakers. None had any visual or neurological disabilities that would interfere with their participation.

2.3.3 Methods and Materials

120 pairs of words were chosen to meet specific experimental criteria. To avoid any effects of a relative difference in orthographic neighbourhood size, each pair of words had minimal difference between their Orthographic Levenshtein Distance (OLD, Yap & Balota, 2009). The mean of the differences between the OLD in all of the word pairs in my stimuli was 0.007 with a standard deviation of 0.2, meaning that each word was matched with a word with an orthographic neighborhood of almost identical size. I also used words of different lengths. There were 51 pairs of four letter words, 37 pairs of five letter words and 32 pairs of six letter words. Each word pair was selected to provide the broadest possible coverage of the frequency ratio space. The distribution of the item frequencies is shown in Figure 2.2.

I used the ACTUATE experimental design package (Westbury, 2007) to collect RT and accuracy data in my task. Each trial began with the display of a fixation cross for a random period of time between 500ms and 1000ms. At that point the fixation cross was removed and each pair of words, displayed directly above and below the location of the cross. The words were displayed

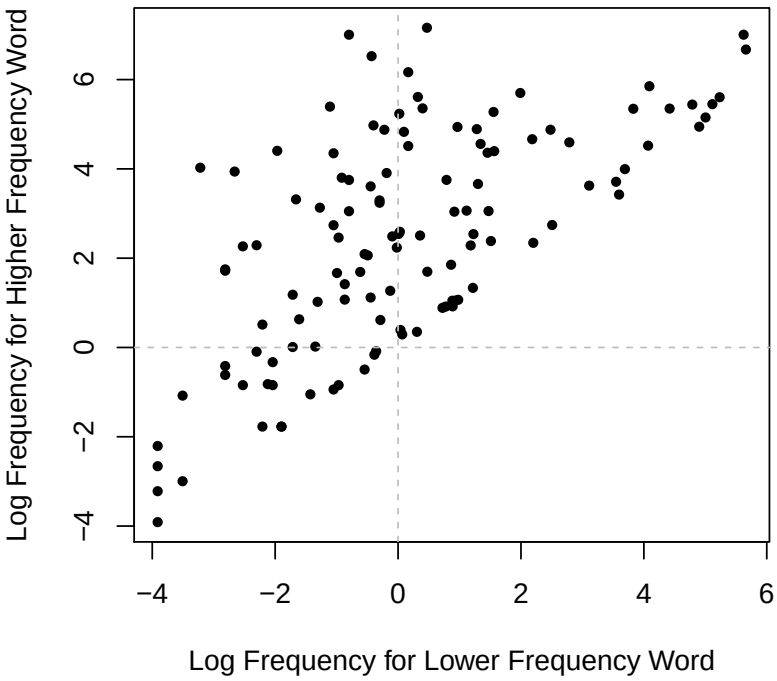

Figure 2.2: Distribution of relative frequencies of stimuli for all word pairs presented in Experiment 2.

in 18 point times roman font on a white background. Each subject had 10 practice trials and then all the word pairs were presented in pseudo-random order. Participants were instructed to press the k key if the word on top was more used more frequently or the m key if the word on the bottom was more used more frequently. The more frequent n-gram appeared above the less frequent n-gram 50% of the time. After completing ten practice trials with feedback, all the experimental trial were completed without any feedback.

2.3.4 Results

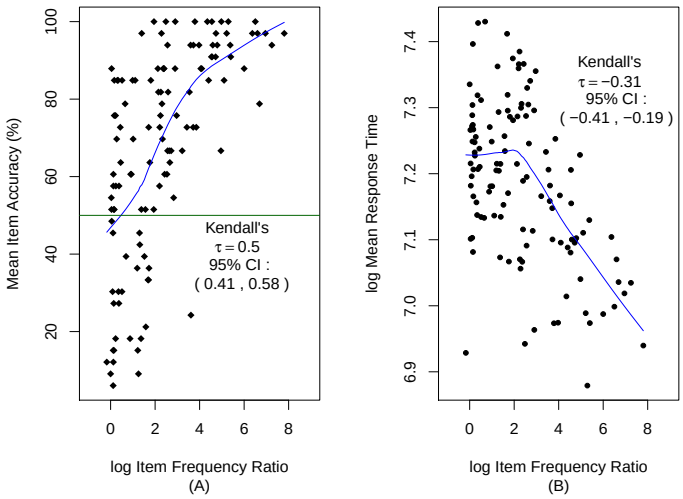

Figure 2.3: A) Relationship between item accuracy and log frequency ratio for all the word pairs in Experiment 2. The green line is at the 50% accuracy level. (B) Relationship between frequency ratio and response time for all the word pairs in Experiment 2. In both of these graphs, Kendall's τ is reported rather than Pearson's r due to the heteroskedasticity of the distribution, and I have included bootstrapped 95% confidence intervals. The blue lines show the LOWESS (locally weighted scatterplot smoothing) smooths.

I first used a graphical analysis to understand the relationship between my two dependent variables and my predictors of interest. In Figure 2.3 (A), I saw that the mean item accuracy increased with the ratio of the orthographic frequencies (Kendall's $\tau = 0.5$, bootstrapped 95% CI 0.41,0.59). In Figure 2.3

(B), I saw a negative relationship between the corpus frequency ratio and RT (Kendall's $\tau = -0.31$, bootstrapped 95% CI -0.41,-0.18). To quantify these effects, I created statistical models to fit the data. I used generalized linear mixed effects models to understand the relationship between the independent variables (such as stimulus properties) and the accuracy of the participants' judgements (Baayen, Davidson, & Bates, 2008). As with the subjective frequency data models above, I compared AIC values to find the best fitting model, and the results of those comparisons are shown in Table 2.2. All models include two fully crossed random factors, Subject and Item.

Table 2.2: Accuracy Model Comparisons for Experiment 2. Pos is the position of the higher frequency word, either above or below the fixation cross. FreqRatio is the log transformed ratio of the word frequencies. The base model included crossed random effects of subject and item, and random slopes were fitted for each subject based on their sensitivity to the item's frequency ratio.

	AIC	Δdf	χ^2	p
1) Base Model	3997			
2) FreqRatio Only	3928	1	70.51	0.0000
3) Position + FreqRatio	3921	1	9.06	0.0026
4) Position + FreqRatio with Random Slopes	3908	0	15.09	0.0001
5) Position $\times$ FreqRatio with Random Slopes	3910	1	0.00	0.9509

Table 2.3: Coefficients for the fixed effects in the generalized linear mixed effects model fitted to the observed accuracy for word pairs in Experiment 1. FreqRatio is the log transformed ratio of the word frequencies and Position is the location of the higher frequency word on the screen. This model also included random intercepts of subject and item as well as random slopes for each subject based on their sensitivity to the item's frequency ratio.

	Coef β	SE(β)	Wald's z	p
Intercept	-0.65	0.20	-3.20	0.0014
FreqRatio	0.61	0.07	9.32	0.0000
Position	0.69	0.22	3.09	0.0020

The best fitting, simplest model was an additive model that included Position (the part of the screen that the higher frequency word was placed in) as

well as the ratio of the frequencies, with the greater frequency ratio producing greater accuracy. The importance of the Position variable suggests that the participants were more accurate when the more frequent word was placed above the less frequent word. The accuracy of each subject on each item was also a function of the frequency ratio, with the larger frequency ratios showing greater accuracy, but the interactive model was not a better fit than the simpler additive model, and contained more free parameters, forcing me to reject it. Adding random slopes for the effect of the frequency on each subject improved the model fit ($\chi^2(1) = 15.09, p = 0.0001$), implying that some subjects were more sensitive to the frequency ratio information than others. The inclusion of random slopes for each subject's sensitivity to the frequency ratio did not eliminate the effect of the frequency ratio. The slope for frequency ratio remained significantly different from zero, as shown in Table 2.3. I did not include the number of letters in the word in this analysis, since when I compared the final model with a model that contained word length as an additive effect, I found no improvement in model fitness ($\chi^2(1) = 2.45, p = 0.12$).

Table 2.4: RT Model Comparisons for Experiment 2. FreqRatio is the log transformed ratio of the word frequencies. Length is the number of letters in the word. The Random Slopes in these models were fitted for each subject based on their sensitivity to the frequency ratio. All models except the first one contain the fixed effect of previous trial RT.

	AIC	Δdf	χ^2	p
1) No Fixed Effects	2667			
2) Base Model (includes PrevTrialRT)	2577	1	92.40	0.00000
3) FreqRatio Only	2537	1	41.03	0.00000
4) Length + FreqRatio	2523	1	16.93	0.00004
5) Length + FreqRatio with Random Slopes	2515	0	9.72	0.00183
6) Length $\times$ FreqRatio with Random Slopes	2517	1	0.15	0.69629

I also performed a linear mixed effects model comparison for the log transformed response times obtained in this experiment to look at the processing load involved in making this type of judgement. Before beginning the analysis, I removed 88 outlier observations from the data set (2% of the data, RTs that

were two and a half standard deviations above or below the grand mean RT). Again, all of my models contained crossed random effects for Subject and Item, but in this analysis, all models also included the log transformed RT from the previous trial (the first trial for each subject was assigned that subject's mean RT). This predictor was inserted to account for inter-trial temporal dependencies, which were pronounced in this experiment (Baayen & Milin, 2010). The other predictors were the ratio of the word frequencies and the length of the word in letters. In Table 2.4 I present the results of this model comparison.

From the model comparison I can infer that word length and the frequency ratio are important predictors, but the interactive model was no better than the simpler additive model. The best model included a random slope for the effect of the frequency ratio on each subject. The fact that this model was superior to all the others suggests that there was some variation in each subject's sensitivity to the frequency of the least frequent word. To confirm that the position of the words did not influence RT, I compared the final model with a model that contained word position as an additive effect, and found that it did not improve the model ($\chi^2(1) = 0.14, p = 0.7$). There was also no benefit in adding the trial number into the model ($\chi^2(1) = 0.06, p = 0.8$). The direction of the relationships in the best model, shown in Table 2.5, provide evidence for a negative relationship between frequency ratio and RT, meaning that there was facilitation when the frequency ratio was larger. The opposite direction was found for word length, due to the fact that longer words take more time to read. The effect of Previous Trial RT was also positive, suggesting that participants exhibited a spillover effect of RT across trials.

2.3.5 Discussion

I created a novel relative frequency judgement task for pairs of words and found that the ratio of the words' frequencies was a powerful predictor of the participants' accuracy in detecting the more frequent word as well as their response time in the task. By matching word pairs on orthographic neighbourhood size, I avoided potential confounds caused by orthographic encoding

Table 2.5: MCMC-based estimates for the coefficients for the fixed effects in the linear mixed effects model fitted to the observed RT in Experiment 2. FreqRatio is the log-transformed ratio of the word frequencies, Length is the length of the word in letters and PrevTrialRT is the log-transformed RT for the preceding trial.

	Estimated β	β_{MCMC}	HPD lower	HPD upper	p_{MCMC}
Intercept	6.0303	5.9852	5.7421	6.2320	0.0001
FreqRatio	-0.0344	-0.0344	-0.0430	-0.0254	0.0001
Length	0.0471	0.0470	0.0255	0.0676	0.0001
PrevTrialRT	0.1415	0.1478	0.1189	0.1756	0.0001

differences. Word pairs that were very close in frequency were much more difficult to judge accurately. Word pairs that were very close in frequency also took longer to process, suggesting that there is a greater cognitive load in distinguishing the relative frequency of items that are very similar in their subjective frequency.

The pattern of reaction time results in this experiment are the inverse of a well known reaction time effect, the symbolic distance effect (SDE, Banks, 1977; Moyer & Dumais, 1978). This effect has been found when the symbolic magnitudes of stimuli are compared in a binary decision task. It takes longer to make a decision about stimuli that have a greater difference in their symbolic magnitude. An example of a typical stimulus would be *Which is larger, Jamaica or Canada?*. In experiments on the SDE, this decision takes longer than for the stimulus *Which is larger, Jamaica or Cuba?* The fact that an effect in the opposite direction was found implies that the cognitive distance between the frequency of two words is not a symbolic distance. It is more likely correlate of familiarity, and the cognitive load is least when the words are very distinct in their subjective frequency. Words that were very close in frequency were harder to judge quickly, not easier.

Why was this so? Balota et al. (2001) and Baayen et al. (2006) found a strong correlation between the subjective and objective frequency of words. This experiment tapped into the same subjective knowledge of word frequency used to rate subjective frequency and these results suggest that this implicit

knowledge for words is being used to judge relative frequency. The accuracy of the participants on this task shows the strong relationship between relative corpus frequency and relative subjective frequency.

My next step was to extend this paradigm to the judgement of the relative frequency of *n*-grams rather than words, making it possible to compare participants' performance on multi-word stimuli to their performance on single word stimuli.

2.4 Experiment 3

2.4.1 Relative frequency of *n*-grams

In this experiment I applied the experimental paradigm that I found to be sensitive to lexical frequency ratios in Experiment 2 to pairs of *n*-grams instead of pairs of words. As I saw in the analysis of Experiment 1, *n*-grams are composed of smaller *n*-grams, and the frequencies of those internal *n*-grams can be predictive of performance on a subjective frequency task. To be certain that the impact of the whole *n*-gram is real, my models needed to simultaneously take both the whole *n*-gram frequency and any relevant internal *n*-gram frequencies into account. The results from Experiment 1 led me to speculate that the same predictors that influenced absolute subjective frequency rating would also influence subjective relative frequency judgements. Any similarities in the patterns of results of Experiment 1 and this experiment would support a common style of processing in both tasks.

2.4.2 Participants

Forty-nine students from the University of Alberta participated in this experiment in exchange for partial course credit. The mean age was 19.3 years old, and 65% were females. All were right-handed native English speakers. None of them reported any visual or neurological issues that would interfere with their ability to participate in the experiment. None had participated in Experiments 1 or 2.

2.4.3 Materials

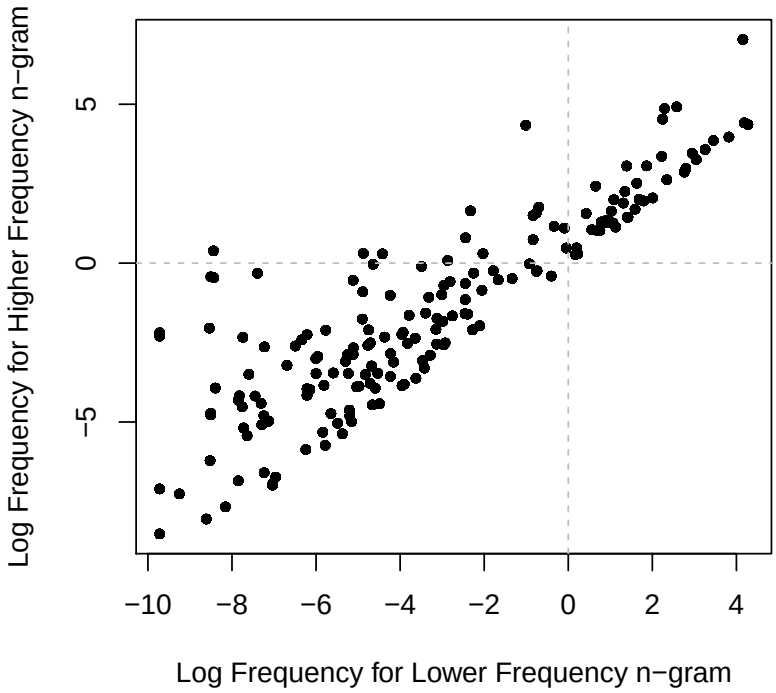

Figure 2.4: Distribution of relative frequencies of stimuli for all *n*-gram pairs presented in Experiment 3.

The same 179 pairs of *n*-grams that were rated by subjects in Experiment 1 were used to create pairs of *n*-grams that covered a wide range of frequency ratios. Figure 2.4 shows the distribution of ratios for all the stimuli in this experiment. The coverage of the majority of the frequency space was excellent, covering most of the space, except for the upper left and upper right quadrants of the space, which are very sparse in the corpus. I wanted to control the influence of the cue of word frequency in the *n*-grams and so I calculated the geometric mean of the word frequencies of the words in each *n*-gram using the unigram frequencies from the Google Web1T corpus (Brants & Franz, 2006). I then matched each *n*-gram with an *n*-gram that had the identical geometric mean. By doing this, I hoped to eliminate any relative frequency cues coming

from individual words in the *n*-grams, cues that I knew to be salient, as I found they influenced performance in the relative frequency judgement task in Experiment 2. With the effect of lexical frequency balanced on each trial, I restricted the source of variation to other types of information. The distributions of the frequency ratios of all of the *n*-grams used in this study are given in Appendix B.

2.4.4 Methods

I used the same method as in Experiment 2. After ten practice trials with feedback, all of the *n*-gram pairs were presented in pseudo-random order for each participant, with no feedback. The more frequent *n*-gram appeared above the less frequent *n*-gram 50% of the time. The presentation format and instructions were identical to those used in Experiment 2.

2.4.5 Results: Accuracy

The overall accuracy with which our participants identified the higher frequency *n*-gram was above chance. I used a bootstrapped confidence interval to assess the accuracy, and found that for 2-grams, the mean accuracy for all subjects on all items was 0.6 (95% CI: 0.58-0.62), for the 3-grams it was 0.62 (95% CI: 0.6-0.64), for the 4-grams is was 0.57 (95% CI: 0.55-0.6), and for the 5-grams it was 0.56 (95% CI: 0.54-0.58). Before attempting to model the accuracy data, I investigated the relative conditional importance of all the frequency ratio variables in predicting mean accuracy using the same random forest methodology described in the analysis of Experiment 1. The relative importance of the predictor variables in predicting the mean accuracy is shown in Figure 2.5. Since there are the same large number of multi-collinear predictors here in Experiment 3 as there were in Experiment 1, I wanted to see which frequency components contributed the most. Before presenting a more formal statistical analysis, I begin with this informal summary of the results of this analysis:

- For 2-grams, the whole *n*-gram frequency ratio was important.

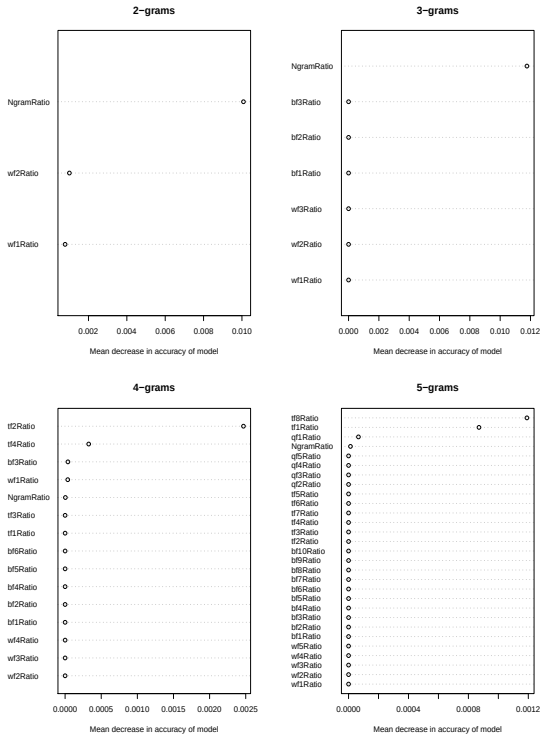

Figure 2.5: Conditional importance of predictors in a random forest model for accuracy in Experiment 3. After creating random forest models, I calculated the relative conditional importance of all of the n-gram frequency variables in predicting mean accuracy, adjusted for correlations between predictor variables, both for the main effects and interactions.

- For 3-grams, the whole *n*-gram frequency ratio was important.
- For 4-grams, the second 3-gram's frequency ratio was the most important. The first trigram, first word and a split 3-gram, **tf4** had weak influence. The whole 4-gram frequency ratio was not a strong predictor.
- For 5-grams, the first 3-gram was a strong predictor along with a split 3-gram, **tf8**, made up of the first, second and fifth words of the *n*-gram. The first 4-gram was a slightly weaker predictor. The whole 5-gram frequency ratio was not a strong predictor.

Table 2.6: Accuracy GLME Model Comparisons for Experiment 3. All models contain crossed random effects for subjects and items. Models in bold type were the best models for each type of *n*-gram. All models include a random intercept for each item and a random slope for the effect of the frequency on each subject.

	AIC	Δdf	χ^2	p
2-grams: Position	3330			
2-grams: Position + *N*-gram Ratio	3327	1	5.42	0.020
3-grams: Position	2387			
3-grams: Position + *N*-gram Ratio	2372	1	16.47	0.000
4-grams: No fixed effects	2286			
4-grams: tf1 Ratio + tf2 Ratio	2282	2	7.92	0.019
5-grams: No fixed effects	2381			
5-grams: tf8 Ratio + tf1 Ratio	2378	2	7.21	0.027

Next I used generalized linear mixed effects models (Baayen et al., 2008) to understand the relationship between the stimuli and the trial-level accuracy of the participants' judgements using the most important variables found in each of the random forest models. Just as in my analysis of the data from the single word experiment, Experiment 2, all of my models included the random effect of item on the intercept crossed with a random slope for the effect of the frequency ratio of each item on each subject. Stimulus position was only included in the 2- and 3-gram models, as it did not enhance the model fitness in the models for the other *n*-gram lengths. The comparison of these models

is shown in Table 2.6. From the model comparison it becomes clear that the ability of the models to predict trial-level accuracy improved when the appropriate frequency ratios were added. I also compared these models shown with other models that included predictors such as trial number and all the individual word frequencies, but these models are not shown in my model comparison table because these models were uniformly lower in fitness than the models shown. Since the individual word frequencies were not found to improve the fit of any of the models for accuracy for any of the *n*-gram types in the experiment this method of matching pairs of *n*-grams was successful in preventing lexical frequency cues from influencing the participants' relative frequency judgements. Finally, stimulus position did not improve the fitness in the models for the 4-grams and 5-grams, and was dropped from those models.

2.4.6 Discussion: Accuracy

After inspecting the coefficients of my models, I noted that in all of my models, except the 4-gram model, the direction of all the relationships was positive – larger ratios led to a higher likelihood of a correct response. As seen in the coefficients for the 4-gram model listed in Table 2.7, the two ratios are pushing in opposite directions. The larger the ratio of the initial 3-gram, the lower the likelihood of a correct response, with the reverse true for the second 3-gram. This points to the 3-grams being strong sources of information for the judgement process. When these two sources were in conflict, accuracy on the task decreased. In the model for the 5-grams shown in Table 2.8, the strength of the effects of the **tf8** and **tf1** ratios were weak despite the results of the model comparison which showed improvement in model fitness when I added those two ratios to the model. Unlike all the other *n*-grams, the evidence from my data for frequency effects influencing accuracy for 5-grams is too weak to be considered likely.

Table 2.7: Coefficients for the fixed effects in the generalized linear mixed effects model fitted to the observed accuracy on 4-gram pairs in Experiment 3.

	Coef β	SE(β)	Wald's z	p
Intercept	0.32	0.17	1.88	0.06
tf1 Ratio	-0.11	0.05	-2.20	0.03
tf2 Ratio	0.16	0.06	2.45	0.01

Table 2.8: Coefficients for the fixed effects in the generalized linear mixed effects model fitted to the observed accuracy on 5-gram pairs in Experiment 3.

	Coef β	SE(β)	Wald's z	p
Intercept	0.27	0.15	1.82	0.07
tf8 Ratio	0.08	0.05	1.71	0.09
tf1 Ratio	0.04	0.03	1.50	0.13

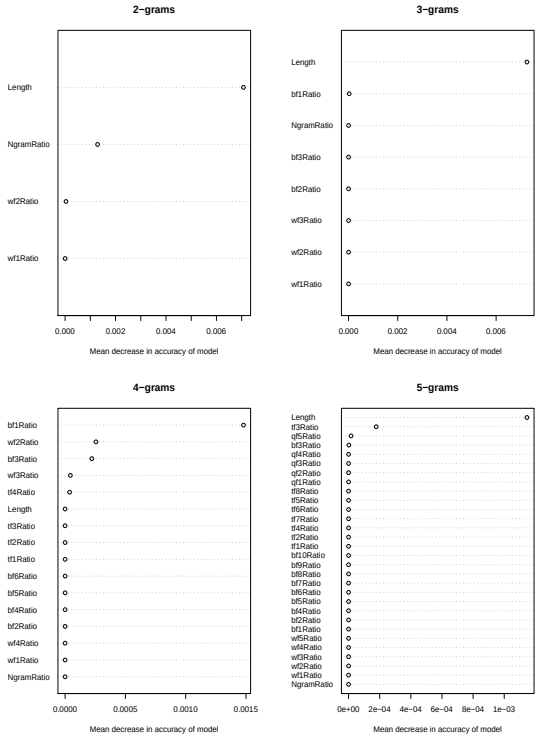

Figure 2.6: Importance for predictors in a random forest model for RT in Experiment 3. After creating random forest models, I calculated the relative importance of all of the *n*-gram frequency variables as well as string Length in predicting mean subjective frequency ratings adjusted for correlations between predictor variables, both for the main effects and interactions.

Table 2.9: Markov-chained Monte Carlo (MCMC) based estimates of the coefficients for the fixed effects in the linear mixed effects model fitted to the observed RTs on 4-gram pairs in Experiment 3.

	Estimated β	$\bar{\beta}_{MCMC}$	HPD lower	HPD upper	p_{MCMC}
Intercept	7.0148	6.7595	6.3329	7.1955	0.0001
Length	-0.0013	-0.0012	-0.0054	0.0031	0.5692
PrevTrialRT	0.1170	0.1490	0.0995	0.1963	0.0001
bf1 Ratio	0.0222	0.0221	0.0081	0.0360	0.0030

2.4.7 Results: Response Time

To begin my analysis of the response time data, I performed a measurement of the conditional importance of the predictors using random forests in the same way as in Experiment 1, predicting mean log-transformed RT from all the frequency ratios and the length (in letters) of the stimuli. The results of this analysis are shown in Figure 2.6. For all *n*-grams except the 4-grams, Length was the only predictor that was important. For 4-grams, the ratio of the first 2-grams was a important predictor of RT, and Length was not an important predictor. To confirm the result of the random forest analysis, I attempted to fit an item-level linear mixed effect model to the data for all the *n*-gram RT data, but even with the addition of crossed random effects for subject and item, only the 4-gram models showed any impact of *n*-gram frequency ratios on model fitness. This correspondence between the random forest models and the linear mixed effect models supported my decision to use random forest models to identify important predictors. After fitting a model that included Length as well as previous trial RT, I performed model criticism by removing data with residuals that were greater or less than 2.5 times the mean residual. This did not change the outcome of the analysis and indicated that my model was not overly influenced by extreme values. The estimated coefficients for all of the fixed effects in this model and their 95% highest posterior density intervals are shown in Table 2.9.

I also performed a model comparison between the base model without the effect of **bf1Ratio** and one with the effect of **bf1Ratio**. The comparison

showed an improvement in model fitness for the second model: $\chi^2(\ 1\ ) = 9.28$, $p = 0.002$). The direction of this relationship was positive, meaning that items with a larger ratio of their initial 2-grams were responded to more slowly. The meaning of this relationship between the first 2-gram frequency and the RT is opaque.

2.4.8 Discussion: Reaction Time

Response times in this task were affected by the number of letters in the *n*-grams, but they were not predicted by any of the frequency ratios, except in the case of the 4-grams, which showed a small effect for the first bigram frequency ratio. This result was quite different from the results in the single word task in Experiment 2, where response time increased as the frequency ratio decreased. There was no symbolic distance effect or inverse distance effect. The reason that the effect went away may be that the impact of the frequency ratios on the cognitive load of the more complicated *n*-gram task is not very large. In the case of the 4-grams, the reason for the increase in response time is, as yet, unclear.

Thus far, I have extended the relative frequency judgement task from pairs of words to pairs of *n*-grams. The results I obtained with *n*-grams for stimuli were similar to those with words – the ratio of certain *n*-gram frequencies predicted the accuracy in detecting the more frequent *n*-grams. This result suggests that the subjective frequency of *n*-grams is something that is accessible to us when it is useful. There appeared to be a preference in the 4-grams and 5-grams trials to use the relative frequency of constituent 3-grams to make a decision, which suggests that the relative frequency of *n*-grams made of 4 or more words was not used by the participants in this task. Potential explanations for the difference in the type of frequency ratio that best predicted accuracy changing for the shorter *n*-grams and the longer *n*-grams will be addressed in the next section along with the implications of the results of the previous experiments.

In summary, I found in Experiment 3 that the probability that the participants could correctly identify the *n*-gram with the higher corpus frequency was

linked with the relative frequency of one or more *n*-grams. For shorter *n*-grams, the whole *n*-gram frequency ratio was important and for longer *n*-grams, the internal *n*-gram frequency ratios were important. *N*-gram probabilities of some type were involved in all of my models.

2.5 Conclusion

In the three experiments presented here I looked at the subjective frequency of words and *n*-grams and how it related to their objective frequency. In Experiment 1, I found that, similar to those for words, the subjective ratings of frequency for *n*-grams were correlated with their corpus frequencies. In Experiment 2 I introduced a relative frequency judgement task and applied it to the relative frequency of words. In Experiment 3 I extended this task to *n*-grams, and I saw that the ratio of the frequencies of *n*-grams can predict the likelihood of correctly choosing the higher frequency *n*-gram in a forced choice task. My efforts to remove lexical frequency cues by matching stimuli by the geometric mean of their component word frequencies were successful, as I saw no predictive input from word frequencies or the ratio of their word frequencies. Relative *n*-gram frequency was the key predictor of accuracy in Experiment 3. These results imply that people have implicit access to the relative frequency of *n*-grams, and that *n*-gram and word probabilities are involved in our processing of *n*-grams, just as letter and word probabilities are involved in our processing of words.

Subjective frequency knowledge is often used implicitly in many tasks, linguistic and non-linguistic, such as word segmentation (Saffran, Aslin, & Newport, 1996), lexical recognition (M. S. Seidenberg & McClelland, 1989), visual object perception (Kirkham, Slemmer, & Johnson, 2002), social learning (Bandura, 1997) and many others. It is not surprising to see subjective frequency effects for groups of words, but the sheer number of *n*-grams that humans are exposed to in our lives makes it difficult to see how it is possible to keep track of our familiarity with each *n*-gram. This concept of a mental lexicon, with various entries for each word or compound word or *n*-gram, has

been recently criticized by Elman (2009, 2011) and Dilkina, McClelland, and Plaut (2010b). Using Elman's ideas, I see *n*-gram representations as dynamic, interactive relations between many types of non-symbolic knowledge. Memory systems could be interacting with comprehension systems and production systems when reading *n*-grams. In this kind of representation, recall of episodic memory traces, ease of articulatory simulation and ease of semantic accessibility all contribute to our ability to judge the absolute and relative frequency of *n*-grams. Frequency of exposure and the depth of the entrenchment of *n*-grams can contribute to the strength of a representation in all of these mental systems, and this could explain why my data show such a consistent influence of *n*-gram frequency on performance in my tasks.

Another way that *n*-gram subjective frequency may emerge is from the sensation of fluency which arises from accessing the meanings of an *n*-gram. Much as lexical access takes longer for words that we do not know the meaning of, *n*-gram access may take longer for *n*-grams that we do not know the meanings of. If *n*-gram subjective frequency emerges from the same processes that produce lexical subjective frequency, and if subjective frequency is related to speed of recognition, then we can look at recent models of word recognition for ideas on how this may happen. Some recent models posit a process of accumulation of evidence when we read and recognize words (Norris & Kinoshita, 2008; Dilkina et al., 2010b; Baayen et al., 2011). One of these models, the Naive Discriminative Reader (NDR) has already been applied to modeling the reading of *n*-grams. It is important to note that the NDR does not assume separate representations for word forms or n-gram forms, but rather shows the emergence of morphological and lexical effects using nothing but sub-lexical probabilistic information. Baayen and Hendrix (2011) used the NDR model to predict reading times for the stimuli used by Arnon and Snider (2010). The NDR model predicted the reading time from the model's knowledge of the statistical properties of pattern of letters and letter bigrams in the input. This model is an example of the kinds of long-term memory traces that are being created from our experience with words and *n*-grams– distributed probabilistic traces. More work will need to be done to link subjective frequency models to

models of reading, but I feel that this may be a promising direction to head in if we are to discover what creates the qualia of word or *n*-gram frequency.

In my results there were some effects that I could not predict when I designed my studies, but that I was able to detect due to the correlational design of my experiments. By choosing stimuli that simultaneously covered a broad span of frequencies and frequency ratios I was able to capture the influence of component *n*-gram frequencies. For example in Experiment 1 I found that both the *n*-gram frequency and the split bigram frequency **bf3** contributed to predicting participants subjective frequency ratings. This result suggests that the first and third words are salient for subjective frequency judgements in 3-grams, but not for other *n*-grams. These split-grams may be related to *discontiguous subtrees* proposed by Bod (2009). They are used in Bod's data-oriented parsing (DOP) model to help explain our ability to parse nonadjacent dependencies such as "BA carried *more* people *than* cargo in 2005" (Bod, 2009, p. 764). This discontiguous subtree, *more XX than* bears a striking resemblance to the split 3-gram, and the influence of the split 3-gram's frequency might provide some behavioural support for parsing models that allow these discontiguous constructions. In contrast, most of the other split-grams that I included in my analyses (see Section 2.6 for the full list) had no detectable influence on the outcomes. The only other time a split-gram entered one of my models was in the relative frequency accuracy model for 5-grams, when the 3-gram, consisting of the first, second and fifth words of the 5-gram rose in importance above the other variables. More evidence will need to collected before any links can be made between probabilistic reading models and syntactic models that presume a representation for many different types of split-grams.

Our results also hint at the existence of differences in the amount of influence of the various grain-sizes. If there was a recurring size of *n*-gram that predicted performance in all of my tasks, it was the 3-gram. In Experiment 3, 3-gram frequency ratios were found to be the most salient *n*-grams for judging the relative frequency of 3-, 4- and 5-grams. One possible explanation for this could be that the probability of seeing groupings of three words provides a particularly strong signal to the language system compared to other size *n*-grams.

This result extends the work of Tremblay and Tucker (2011) by finding that 3-gram ratios were being used in my tasks, just as they found that probabilistic information in 3-grams was being used more than other *n*-gram sizes to recognize 4-grams in their experiment. Furthermore, in both absolute (Exp. 1) and relative (Exp. 3) frequency judgments for 5-grams, the frequency of their component *n*-grams were more predictive of the outcome than the frequency of the 5-gram itself. These results suggest that when we read longer *n*-grams, the subjective frequency of the shorter internal *n*-grams is involved somehow in the process. This type of converging evidence strongly supports continued exploration of the contribution of internal *n*-gram probabilities in future analyses of *n*-gram processing. Such analyses could explore whether specific subset of internal probabilities will come into play in the majority psycholinguistic tasks.

If 3-gram frequency is implicated in the processing of 4-grams and 5-grams, I speculate that 3-gram probability information is being used continuously during reading longer streams of text, and is being done so implicitly. This simultaneous interaction between the probabilities of multiple *n*-gram components in my evidence bears a striking resemblance to recent results from research into processing polymorphemic and compound words, where lexeme/morpheme frequency and meaning are all simultaneously involved in processing, even when they are not required, or even helpful, for the task (Kuperman, Dambacher, Nuthmann, & Kliegl, 2010; Kuperman et al., 2008; Gagné & Spalding, 2009; Juhasz & Berkowitz, 2011). This probabilistic interaction within *n*-gram processing buttresses the argument that *n*-gram processing may be analogous to word processing, with the only difference being the length and probabilistic complexity of the input. The possibility that words and *n*-grams are somehow represented as entries in a lexicon, and that there is a search process across this lexicon as proposed by Forster and Hector (2002), looks increasingly untenable. The sheer number of representations that would be required in a localist model of language that included words *and* *n*-grams in a lexicon would be around 10^9, and even if this search could proceed faster than the fastest known parallel search algorithms, it would still be too slow to be plausible.

My results are compatible with an emergent account of lexical processing that does not depend on unique representations for words or *n*-grams.

The work presented in this chapter supports the notion that *n*-gram probability is a new and important element in psycholinguistics, one that will allow us to explore language processing in new ways. The vast majority of models for word and sentence processing have thus far avoided dealing with the impact of arbitrary *n*-grams on language performance. I have presented experimental evidence that the granularity of language extends beyond words to *n*-grams, and that the probability of *n*-grams influences their subjective frequency. The evidence I have presented here, built upon the work of many others, suggests that subjective *n*-gram probability effects exist at many grain-sizes. Considering the accumulation of evidence presented here, the time has come to bring *n*-gram probability information into language processing models. New models of reading, such as the NDR model (Baayen et al., 2011) that can predict *n*-gram frequency effects and incorporate linguistic knowledge of patterns of varying sizes and levels of abstraction will give us the necessary context to better understand experimental results and to determine what cognitive limitations shape our ability to process *n*-grams. There may be fundamental upper bounds to the complexity of the probabilistic information that we can use when reading *n*-grams and those constraints will require further exploration before they become clearly defined.

2.6 Experimental Stimuli

Table 2.10: 2-grams and 3-grams used in Experiment 3.

2-grams	3-grams
richest man : push ahead	who never returned : committed to tradition
at least : due to	rare vinyl records : juvenile detention facility
wheat flour : snack foods	long curly hair : dubious scientific value
boxer shorts : midterm exam	hip and stylish : never been easier
to work : the future	may end up : cast members of
diet pill : prime minister	residents who are : until my final
human rights : hang out	high school students : step by step
make sure : this album	with the result : told from the
over time : not pay	to be accurate : offered the single
start leaning : veterans studied	discovers that the : but owners of
metric tons : inner workings	just as she : the women sent
music on : and freedom	not good enough : with old structures
aware of : copies of	almost every city : space around him
we finally : wildlife in	the property of : value of the
conveniently located : periodic table	the way of : was not the
is hard : night of	an opportunity to : the end result
rather than : no longer	the fact that : at the end
people enjoy : published since	procedures outlined in : designed to convince
can we : be important	on fossil fuels : pass intercepted by
up until : once we	sooner or later : paves the way
end up : my friends	of their respective : can be used
law enforcement : cell phones	boys and girls : black and silver
cash machines : work visas	that most people : to paying that
watching the : and pull	you will find : feel free to
to the : and in	man and woman : has one too
the pain : limits of	the first time : going to be
string quartet : alarm bells	was added to : of the material
hope that : made on	credit card debts : their offending behaviour
items from : what are	congestive heart failure : deputy prime minister
and pleasure : he only	chopped fresh parsley : gall bladder surgery
to others : may only	be forgiven for : among the passengers
swimming pools : winning streak	just about anyone : manage it well
health care : figure out	appear to be : who is now
travel tips : water shortage	is usually required : return your application
green hills : weekend trips	one or two : not yet been
the following : and most	can you do : to be done
active in : for ideas	must also be : and from yellow
heart disease : motor vehicle	ever so slightly : an accurate understanding
brussel sprouts : crankshaft pulley	work in the : next to the
that are : some of	law enforcement agencies : prescription drug coverage
heavy snoring : paranoia starts	popular tourist attractions : minimize potential impacts
final phase : latest songs	densely populated areas : happy belated birthday
most recently : every game	vast majority of : need to escape
youth in : your foot	
several times : her eyes	
of our : more and	
parking lot : law firms	
affordable toy : scores suffer	
all the : been the	
be able : not know	
human body : no jurisdiction	
recent novel : little wings	
umbilical cord : drought tolerant	
with you : much of	
people with : the streets	
minimum wage : crude oil	
whitewater rafting : unleavened bread	
the time : has and	
as do : any by	
snowboard rental : lighter burden	

Table 2.11: 4-grams used in Experiment 3.

4-grams
be based in the : the business is on
of matter and energy : by means of local
physical and mental disabilities : feel really sad about
protection of human rights : realized that he had
affair of the heart : incredible success in the
are not the cause : and the first paper
and the country was : is of a whole
it is found to : is made at a
this appears to be : first introduced to the
is part of the : and is home to
has been approved by : who you truly are
is the time to : is the year the
part of the first : that the author is
information for the current : home of the latest
brag about having heard : crops genetically engineered to
in males and females : company policies and procedures
is the only system : in one year is
for the following reasons : be present during the
is the best it : the second is by
may be impossible to : that was collected from
and also have the : and the date is
this is the point : the time we are
guess is that the : felt to be a
was in the room : to be works of
is the first of : the following is the
part of the unique : find food for the
is to be a : is the time of
and the time is : the information is the
have the day of : is to the children
is to get you : are we to be
will feature case studies : trampled in the dust
you say you are : the role of design
is an outstanding example : need to start thinking
killed by hostile fire : metric weights and measures
starting in the new : in court was a
played a central role : making false statements in
part of the search : is to avoid a

Table 2.12: 5-grams used in Experiment 3.

5-grams
and to see that the : to the case on the
win friends and influence people : was rumoured that he had
at the end of each : to change the lives of
is the purchase of a : or for the development of
about what can happen to : and learn everything there is
the beginning of the next : of what the year has
all water under the bridge : is at least four times
thank you so much for : always ready to help you
which they have already received : had a distinct impact on
of their registered owners and : serve as a guide to
there are plenty of opportunities : given over a long period
is less like an annoying : are paying close attention to
if we did not know : gives us a sense of
is the name of a : of the city by the
support the full range of : be able to accept a
couple of weeks or so : a very active forum for
gave birth to a beautiful : help you organize your home
used as a kind of : all of whom had the
that the changes in the : and that he is a
data that can not be : in the front or back
it did not seem to : to help you prepare for
be implemented in the future : but good enough for a
so you can find out : a chance of showers and
play an active role in : safer to keep it here
was sentenced to six months : opportunity to introduce ourselves as
here and there in the : and at the beginning the
preparation for life in the : the result of arbitrary and
has nothing to do with : ask to speak to a
with an interesting story or : occurred early in the project
ways to get rid of : at least one year after
that all words are spelled : we propose to carry out
would like to see this : were also of the opinion
appear within a few moments : keep in mind when picking
finally took the plunge and : stable at room temperature for
of the ability of his : to know and keep the
going to have to get : the various properties of the
and can be used for : the first step in the
of the last day of : may not be on the
is a leader in the : and now the process of

2.7 Distribution of *n*-gram frequency ratios in Experiment 3

Table 2.13: Descriptive statistics for 2-grams used in Experiment 3 (log-transformed).

	Mean	SD	Min	Max
N-gram Ratio	1.13	1.01	0.00	5.35
First Word Ratio / wf1Ratio	0.18	2.44	-6.56	7.00
Second Word Ratio / wf2Ratio	-0.01	2.62	-6.00	7.17

Table 2.14: Descriptive statistics for 3-grams used in Experiment 3 (log-transformed).

	Mean	SD	Min	Max
N-gram Ratio	2.50	2.24	0.01	8.82
First Word Ratio / wf1Ratio	0.45	3.01	-6.12	5.60
Second Word Ratio / wf2Ratio	-0.76	3.20	-9.89	4.19
Third Word Ratio / wf3Ratio	0.42	3.00	-6.82	7.58
First Bigram Ratio / bf1Ratio	0.54	2.97	-6.48	8.28
Second Bigram Ratio / bf2Ratio	0.78	2.55	-4.28	7.70
Third Bigram Ratio / bf3Ratio Split-gram: w1, w3	0.41	5.29	-13.15	14.03

Table 2.15: Descriptive statistics for 4-grams used in Experiment 3 (log-transformed).

	Mean	SD	Min	Max
N-gram Ratio	1.91	1.22	0.05	4.58
First Word Ratio / wf1Ratio	-0.28	2.24	-5.19	8.33
Second Word Ratio / wf2Ratio	1.34	3.00	-5.08	6.47
Third Word Ratio / wf3Ratio	0.09	3.05	-8.36	6.51
Fourth Word Ratio / wf4Ratio	-1.08	2.70	-6.39	4.58
First Bigram Ratio / bf1Ratio	1.21	2.53	-3.52	7.40
Second Bigram Ratio / bf2Ratio	1.43	4.29	-13.23	11.92
Third Bigram Ratio / bf3Ratio	-0.26	3.07	-5.25	5.91
Fourth Bigram Ratio / bf4Ratio Split-gram: w1 ,w3	-0.12	4.32	-11.62	7.30
Fifth Bigram Ratio / bf5Ratio Split-gram: w2 ,w4	0.12	3.52	-6.85	7.97
Sixth Bigram Ratio / bf6Ratio Split-gram: w1 ,w4	-1.27	4.75	-11.83	6.45
First Trigram Ratio / tf1Ratio	1.62	2.83	-3.84	9.30
Second Trigram Ratio / tf2Ratio	1.36	2.34	-4.75	6.30
Third Trigram Ratio / tf3Ratio Split-gram: w1 , w2 ,w4	0.94	4.77	-9.96	12.54
Fourth Trigram Ratio / tf4Ratio Split-gram: w1 , w3 ,w4	-0.50	4.86	-8.73	11.63

Table 2.16: Descriptive statistics for 5-grams used in Experiment 3 (log-transformed).

	Mean	SD	Min	Max
N-gram Ratio	2.11	2.10	0.00	7.51
First Word Ratio / wf1Ratio	-0.51	2.76	-7.03	5.97
Second Word Ratio / wf2Ratio	0.72	3.38	-8.30	5.79
Third Word Ratio / wf3Ratio	-0.15	4.11	-6.60	7.53
Fourth Word Ratio / wf4Ratio	0.89	3.28	-5.36	7.16
Fifth Word Ratio / wf5Ratio	-0.89	2.55	-7.75	4.13
First Bigram Ratio / bf1Ratio	0.41	2.96	-6.34	7.86
Second Bigram Ratio / bf2Ratio	0.29	3.03	-8.15	7.52
Third Bigram Ratio / bf3Ratio	1.21	3.45	-8.92	8.28
Fourth Bigram Ratio / bf4Ratio	0.32	3.39	-5.98	8.73
Fifth Bigram Ratio / bf5Ratio Split-gram: w1 ,w3	-1.17	5.71	-10.24	12.30
Sixth Bigram Ratio / bf6Ratio Split-gram: w1 ,w4	0.56	5.12	-12.67	12.24
Seventh Bigram Ratio / bf7Ratio Split-gram: w1 ,w5	-0.38	4.37	-6.65	15.59
Eighth Bigram Ratio / bf8Ratio Split-gram: w2 ,w4	1.45	5.59	-9.21	16.26
Ninth Bigram Ratio / bf9Ratio Split-gram: w2 ,w5	-0.13	5.48	-15.95	10.38
Tenth Bigram Ratio / bf10Ratio Split-gram: w3 ,w5	-0.37	4.46	-13.69	9.69
First Trigram Ratio / tf1Ratio	0.13	3.23	-7.43	7.95
Second Trigram Ratio / tf2Ratio	1.70	2.96	-4.37	7.27
Third Trigram Ratio / tf3Ratio	0.89	3.42	-6.43	7.52
Fourth Trigram Ratio / tf4Ratio Split-gram: w1 ,w3 ,w4	0.81	5.96	-10.73	12.22
Fifth Trigram Ratio / tf5Ratio Split-gram: w1 ,w4 ,w5	0.14	4.82	-10.41	10.43
Sixth Trigram Ratio / tf6Ratio Split-gram: w2 ,w4 ,w5	0.67	6.44	-12.99	15.26
Seventh Trigram Ratio / tf7Ratio Split-gram: w1 ,w2 ,w4	1.72	4.16	-6.58	10.98
Eighth Trigram Ratio / tf8Ratio Split-gram: w1 ,w2 ,w5	0.18	5.95	-11.23	15.10
First Quadgram Ratio / qf1Ratio	2.03	3.00	-2.87	9.20
Second Quadgram Ratio / qf2Ratio	1.78	2.45	-2.45	6.52
Third Quadgram Ratio / qf3Ratio Split-gram: w1 ,w3 ,w4 ,w5	-0.26	5.98	-11.60	14.02
Fourth Quadgram Ratio / qf4Ratio Split-gram: w1 ,w2 ,w4 ,w5	-0.26	4.65	-9.92	10.96
Fifth Quadgram Ratio / qf5Ratio Split-gram: w1 ,w2 ,w3 ,w5	1.69	5.44	-8.84	11.63

Chapter 3

Probabilistic information influences eye movements when reading trigrams.

In the previous chapter we learned about our abilities to remember *n*-grams. How does this *n*-gram knowledge influence comprehension? *N*-grams contain a wealth of probabilistic information that is potentially useful to readers. To better understand which information is relevant and how it is used by the language system, we examined eye movements of participants while they read 1000 trigrams that were sampled strategically from an extremely large, diverse set of trigrams selected from a corpus of English web pages. Each stimulus was sampled from one of 1000 different combinations of *n*-gram frequency bands, providing unprecedented coverage of the probability space of trigrams. An examination of reading times for *n*-grams showed that having a higher frequency has a generally facilitatory effect, but there were also complex, non-linear interactions between *n*-gram frequency and other probabilistic measures. The probability of making one or more regressive saccades during reading was found to be best predicted by a model that contained non-linear interactions that included frequency and other measures of information. The way we read a trigram is intimately linked to the probabilistic information that it contains.

3.1 Introduction

If one thinks of written language as a stream of words, then reading is nothing more than identifying those words as quickly and accurately as possible. There is a wealth of information contained in this stream, and experienced readers may be taking advantage of this information to improve the efficiency of their reading. In particular, the probabilistic information contained in word transitions and multi-word transitions allow a reader to develop some expectations about what words will appear ahead in the stream. There is a natural bias towards looking at words as the key unit of languange in this stream, but there is no clear reason for this bias. The contents of the language stream could be equally well thought of as a stream of groups of words — multi-word units known as *n*-grams. The psycholinguistics of *n*-gram processing is a new and vibrant area. Results from initial studies of n-gram reading offer evidence from a wide variety of paradigms suggesting that readers are sensitive to n-gram information (see Shaoul & Westbury, 2011 for a review). The question I will attempt to address in this chapter is: which different types of probabilistic information are contributing to *n*-gram reading efficiency and how do they all interact? By looking at the timing of eye motion when reading trigrams, I hope to explore this complex problem.

Why is this problem complex? Corpus-based sources of probabilistic information entail an explosion in the number of potential predictors for each stimulus. In particular, for every new predictor found, the number of possible ways that this predictor can interact with other known predictors needs to be considered. In this chapter I will look at trigrams and their component words and bigrams. The number of potential interactions that could be included in a model increases quickly. Yet trigrams are less complex than larger *n*-grams. I chose to study trigrams because they are complex enough to get at the interplay of probabilistic information while being computationally tractable.

Our goal in this exploration was to seek out empirical confirmation that both *n*-gram and and inter-lexical probability measures contribute to the way that we read sequences of words. These inter-lexical predictors, such as bigram

entropy, are new and, to my knowledge, no other studies have looked at the impact of probability and entropy on eye movement when reading *n*-grams. In new areas of psycholinguistics it is unwise to limit in any way the scope of the investigation by imposing *a priori* experimental hypotheses about how my measures of probability will influence behavior. This exploration of trigram reading is guided by a desire to explore the interactions of all possible inputs to the reading system, and by finding which statistical model best fits the data. By exploring as much of the stimulus space as possible I hope to stimulate the development of new theories of reading.

Why are eye-movements an interesting entry-point for studying *n*-gram effects? The theoretical issue is whether cognition is rational or not. Anderson (1990) proposed approaching cognitive problems by defining the goals and constraints involved, and then developing a model that implements an optimal solution within those constraints. After comparing the model's performance with that of humans we can think about the similarities and this will help us see if the mind is actually using the best method of solving the problem.

The problem faced by the mind when trying to read is that if connecting visual information with linguistic meaning. One part of this very complex process is the placement of visual fixations, the timing of these fixations and saccades. Recent work by Levy, Bicknell, Slattery, and Rayner (2009) has provided evidence that the reading system makes use of uncertainty about context when processing new words, showing that the if information is available that will improve the ability to process the meaning of text, the visual system will take advantage of this information. The theoretical question that will be addressed in the chapter is: does the visual system take advantage of the *n*-gram information that has amassed as it has accumulated many years of experience?

Let's now delve into the relevant work on the reading system's relation to a wide variety of probabilistic information contained in streams of language. The most relevant work on *n*-gram processing from Chapter 1 will be highlighted and described in more detail below.

Some interesting findings have been come from in experiments on read-

ing multimorphemic words (derived words or compound words). Multimorphemic word reading research is relevant to my research because the way that morphemes make up multimorphemic words may have similarities to the way words make up *n*-grams. Studies of eye movement when reading multimorphemic words have produced advances in our understanding of compound word reading. These models are allowing researched to model and predict the time-course of compound word reading. Kuperman et al. (2008), Kuperman, Dambacher, et al. (2010), Kuperman et al. (2009), and Juhasz and Berkowitz (2011) have all found interactions between probabilistic variables. All these studies found that *morphological complexity* or *relative entropy* predicts processing times for compound words. The consensus from these studies is that there are multiple routes to processing compound words: a whole word route, a lexeme route and perhaps other routes. They also find trade-offs and interactions between these routes. The Probabilistic Model of Information Sources (PROMISE) model proposed by Kuperman et al. (2009) is one of the first models to explain the evidence from multimorphemic word reading — it is a model that includes probabilistic information as a predictor of ease of lexical access. Lexical processing is easier and faster for words that carry more information. It is a parallel, interactive model that allows for the simultaneous processing of information at multiple levels (from letters up to words). Due to this dynamic, interactive nature, they state that "the effect of virtually any single information source on the speed of word recognition can range from facilitatory to negligibly small to inhibitory depending on the effects of other such sources and the likelihoods that those other sources are available for processing." (Kuperman, Bertram, & Baayen, 2010, p. 95). They go on to point out that considering any one information source in isolation from others by keeping the values of other information sources constant (i.e. by matching stimuli) is bound to miss the essentially interactive use of information in multimorphemic word recognition. I will heed this warning as I study *n*-gram reading, and avoid the use of factorial designs.

As noted in Chapter 1, there is evidence that *n*-gram processing is influenced by the frequency of that *n*-gram, independent of the frequency of its

component words and bigrams. Bannard and Matthews (2008), Matthews and Bannard (2010) and Arnon and Clark (2011) found this effect in children's production of language. Arnon and Snider (2010), Tremblay and Baayen (2010) and Tremblay et al. (2011) found it in the reading speed and memory retrieval pattern of adults. Tremblay and Tucker (2011) found it in the production of *n*-grams by adults.

One study has used eye movement to look at *n*-gram frequency effects. Siyanova-Chanturia et al. (2011) presented both L1 and L2 subjects with two types of 3-grams: binomial phrases (*bride and groom*) and those same phrases reversed (*groom and bride*). These two types of *n*-grams are naturally very closely matched on many lexical variables, and they proposed that any differences in processing must arise from effects of *n*-gram frequency. The binomial 3-grams had an average frequency in the BNC that was 10 times that of the reversed 3-grams (2.473 per million versus 0.274 per million). Thirty 3-grams of each type were embedded in sentences and read by participants in the eye tracker. Measuring eye movements they found that binomial phrases were read faster than reversed phrases by L1 speakers but not L2 speakers. They also found that phrasal frequency facilitated reading even after taking into account the effect of phrase type, providing more evidence that increased exposure to an *n*-gram contributes to its entrenchment. The stimuli used in this experiment were culled from a very small subset of the full range of 3-grams, which helped eliminate unwanted variability, but limits the applicability of the results to non-binomial phrases.

Columbus, Bolger, and Baayen (2010, 2011) also looked at the effect of frequency and idiomaticity on word fixation times when reading *n*-gram embedded in sentences. They found that for semantically transparent, non-idiomatic *n*-grams, there were shorter first fixation durations for words in *n*-grams with higher frequencies. Kliegl, Nuthmann, and Engbert (2006) describes the massive impact of sentence context on word reading, and for this reason I chose to use bare trigrams rather than trigrams embedded in sentences. This type of stimuli offers the possibility of consistent fixation of the first word of the trigram, and eliminates the affect of previous context on the way the trigram

is read.

Our goal in this study was to look at the way that people read groups of three words, which I refer to as *trigrams*. One major methodological issue with the research done so far is that the number of stimuli in each experiment was relatively small, and therefore the portion of frequency space covered by these experiments was also small relative to the size of the frequency space of all trigrams used in English. Undoubtedly there are parts of the trigram frequency space that were left untouched by all of these experiments. The stimuli in my experiment were selected to provide the broadest possible coverage of all of the frequency spectra with the aim of finding if the relationship between whole trigram frequency and reading time would hold true for this representative sample. Another issue that arises with the research done so far is that the temporal resolution of the data collected has in general been low. Collecting response times for the reading of the whole *n*-gram, as well as reading times for each word, it becomes possible to see how the reading of *n*-grams progresses. How much of an impact will there be of whole-*n*-gram measures on the reading of the parts? In compound word reading Kuperman et al. (2008) found that whole word frequency influenced first fixation durations, so there is a potential for this type of effect to be found in *n*-grams as well.

3.2 Probabilistic Information

In the work discussed above several measures of probability and information have been used in the analysis of *n*-grams. *N*-gram frequencies are clearly the probability of an *n*-gram occurring in a language stream, and have been found to predict familiarity ratings (see Section 2.2.4). To move beyond simple frequency, I will look at other measures that are derived from frequency counts. The first derived measure that I will be looking at is Pointwise Mutual Information (PMI, Fano & Hawkins, 1961), which comes from the field of information science. It is used to measure the degree to which words in a stream occur together more frequently than would be expected by chance. Increased PMI means a stronger association between the words, while a lower

PMI means that the co-occurrence of the words is more likely due to chance. To put it another way, high PMI n-grams are those with much greater coherence than is expected by chance (N. C. Ellis & Simpson-Vlach, 2009; Bell et al., 2009; Pluymaekers, Ernestus, & Baayen, 2005b; Gregory, Raymond, Bell, Fosler-Lussier, & Jurafsky, 1999). I calculated the PMI in the following way for the trigrams:

$$PMI_{trigram} = log\left(\frac{P_{trigram}}{P_{w_1} \times P_{w_2} \times P_{w_3}}\right) \tag{3.1}$$

where P_{w_1} is the probability of the first word of the trigram. For all of my calculations I used the log transformed frequency of the item in the Google Web1T corpus (Brants & Franz, 2006) as an estimate of the probability of the item. There have been two studies so far that use PMI in relation to *n*-gram processing. N. C. Ellis and Simpson-Vlach (2009) looked at reading times of *n*-grams and used PMI and *n*-gram frequency to try to predict reading time performance. They found that n-gram frequency was not a strong predictor of reading time, whereas PMI was.

PMI has been called a measure of *lexical association* or *coherence* by N. C. Ellis and Simpson-Vlach (2009). There have been criticisms of the PMI measure, in particular by Kilgarriff (2005), who noted that PMI is problematic because it assumes that the distributions of words in an *n*-gram are independent. His argument is that this cannot be true because non-random contextual relationships are what define language, so words never occur at random. For this reason, PMI can be said to over-estimate the true amount of association between words in an *n*-gram. Another aspect of *n*-gram frequencies and *n*-gram PMI measures is that they are not conditional in any way. They provide valuable information about the nature of an *n*-gram, but they do not take into account local context and local predictability. If my theory of *n*-gram processing is truly information-centric, it should include contextual measures. One way to measure contextual regularities is to use *entropy* or *information content*.

The amount of information in a trigram can be looked at in several ways. I looked at three types of information in particular: the final bigram information

content and the second and third words' information content. Information content is the average amount of information conveyed by a particular *n*-gram, and it is a function of the probability of each *n*-gram given each possible context C. This measure is similar to entropy in that it measures the disorder, or the unpredictability of an *n*-gram; the more contexts an *n*-gram frequently appears in, the less predictable it will be from contextual cues. I calculated the average information content for *n*-grams in the same way that Piantadosi, Tily, and Gibson (2011) did for words. The average negative conditional log probability of an *n*-gram given its contexts, or its average surprisal, is estimated from corpus frequencies in the follow way:

$$IC_{Ngram} = -\frac{1}{N}\sum_{i=1}^{N} logP(Ngram|C = c_i) \tag{3.2}$$

$$IC_{Ngram} = -\frac{1}{N}\sum_{i=1}^{N} log\left(\frac{P(Ngram \bigcap C = c_i)}{P(C = c_i)}\right) \tag{3.3}$$

where *Ngram* is a bigram or word, N is the number of times the *Ngram* appears in the corpus and C is a context word or bigram. To account for the lack of access to the original Web1T corpus to calculate the IC, a token-count weighted average was used across all the contexts.

The IC can be thought of as the average surprisal, averaged across all of the occurrences of the *n*-gram in the corpus. To make the idea of IC clearer, I will use as an example the trigram *was so ludicrously*, one of the stimuli that will be used later on in this Chapter. I use the acronym **w3IC** to refer to the IC for the third word in this trigram, *ludicrously*. The w3IC for *ludicrously* is 12, the highest in the stimulus set. In contrast, an example of a final word with a small w3IC would be the final word in the trigram *front porch of* (*of*, $w3IC = 1.23$). Essentially the w3IC captures, on average, how surprising it is to see the third word, given the first two. High average surprisal means that the third word is highly informative. To capture the informativeness of the entire *n*-gram, I calculated the sum of all the average information content values w2IC, w3IC and b2IC. I call this number total information content (TIC). This summing allows me to capture the overall local informativeness of the *n*-gram and its components in one variable.

I am not the first to look at information and PMI when studying *n*-grams. Tremblay and Tucker (2011) have recently used probabilistic measures including PMI and conditional probability while studying *n*-gram processing. They asked their participants to read 462 quadragrams out loud. When choosing their quadragrams, they did not put restrictions on the completeness of the *n*-grams, allowing both complete ones (*I don't really know*) and incomplete ones (*at the age of*). They looked at *n*-gram frequency, conditional probability, and PMI to see how these measures interacted in their models to predict onset latency and the production duration. Their measure of conditional probability is similar to my measure of information content, allowing me to easily compare my results to theirs. For onset latency they found a facilitative effect of phrasal completeness, and they found interactions between frequency and conditional probability. For the production durations, *n*-gram frequencies were dominant, with less contribution from conditional probability and PMI. I will measure the total duration of reading the trigram, which is comparable to the onset latency discussed in Tremblay and Tucker (2011). If I find a negative effect of phrase completeness on reading time, a negative effect of frequency, and interactions between PMI and information content, then I will have replicated their findings.

3.3 Methods

3.3.1 Participants

Twenty-one native speakers of Canadian English took part in the eye movement experiment. All participants were recruited from introductory Linguistics courses offered by the Department of Linguistics at the University of Alberta. Participants received partial course credit for their participation. All participants had normal or corrected-to-normal vision. Ethics approval for this experiment was obtained from the University of Alberta before data was collected.

3.3.2 Materials

I used the Google Web1T *n*-gram data (Brants & Franz, 2006) as my source of the internal frequencies for all 258,599,481 lower-case trigrams in the corpus. Each trigram has seven frequencies associated with it: w1f (first word), w2f (second word), w3f (third word), b1f (first bigram), b2f (second bigram), b3f (split gram of w1 and w3) and the whole trigram frequency. I then selectively sampled from the full set of approximately 300 million trigrams in such a way that there was a representative trigram from each of the possible combinations of values of the log_{10} of the frequency of each of the seven components. An example of one of these frequency band combinations is shown in Figure 3.1.

Any trigram that satisfied this set of inequalities:

$0.1 \leq \text{w1f} \leq 1$

$1 \leq \text{w2f} \leq 10$

$0.01 \leq \text{w3f} \leq 0.1$

$0.001 \leq \text{b1f} \leq 0.01$

$0.0001 \leq \text{b2f} \leq 0.001$

$10 \leq \text{b3f} \leq 100$

and

$0.00001 \leq \text{Trigram Frequency} \leq 0.0001$

Figure 3.1: One of the 1000 frequency bands that were used to select stimuli.

The total number of unique frequency bins that contained one or more trigrams from the full set of 300 million trigrams was about 27,000. If I used one trigram from each of these bins, the number of stimuli to present would be unreasonably large. I instead randomly sampled 1000 frequency bands from this set of 27,000 frequency bands, and then randomly sampled one trigram from all the trigrams in each of the 1000 frequency bands.

Lemke, Tremblay, and Tucker (2009) asked subjects to produce 4-grams and found that *n*-grams that were full constituents (*I don't want to*) were produced faster that those that were not (*in the middle of*). Most of the trigrams in this experiment were not syntactic constituents (*and into the*), but

there were a minority that could conceivably be uttered in conversation (such as *chopped fresh marjoram*). To avoid any confounds that could be linked to constituency, I measured the subjective completeness of all of the stimuli. I recruited 29 new participants (all graduate students) who did not participate in the eye-tracking experiment to rate the completeness of 250 stimuli (the set of 1000 stimuli was divided into 4 parts). Each participant chose the word YES or NO for each stimulus as an answer to the question "Could this phrase be used on its own?". The average of these ratings was calculated for each stimulus, and as expected, the majority of items were judged by the majority of the raters to be incomplete. 72% of the items had an average rating below 0.5 (mean: 0.31, median: 0.22, standard deviation:0.3). The mean rating is included in the list of predictors in Table 3.1 as *mncmplt*.

All analyses were done using the R language and environment, version 2.13 (R Development Core Team, 2009), the **mgcv** package (Wood, 2006), the **lme4** package (Pinheiro & Bates, 2009) and the **languageR** package (Baayen et al., 2008).

3.3.3 Procedure

The stimuli were presented on a CRT monitor using the Experiment Builder ™ software (SR Research Ltd., Mississauga, Ontario, Canada). The data were collected using an Eyelink II ™ head-mounted eye tracking system (SR Research Ltd., Mississauga, Ontario, Canada). Eye movements were collected using pupil-only sampling at a rate of 500Hz.

Each experiment session was preceded by 10 practice trials, with rest and recalibration breaks occurring after each block (approximately every 5 minutes). Both the practice and 1000 experimental items were randomly ordered for each participant. Participants were seated at a comfortable distance from the screen (approximately 70 cm). They were asked to silently read the phrases for meaning as quickly and as naturally as possible. The trigrams were presented in white Courier font on a black screen following a fixation cross. All stimuli were presented on the left of the screen, halfway from the top. The participant read the trigram, and then cued the next trigram by moving their

gaze to an invisible boundary (100 pixels wide) on the right side of the monitor. The gaze-contingent cue was used in this experiment to prevent participants from moving their eyes down to a keyboard, foot-pedal or mouse before the trial ended, a likely outcome given that only three words were presented at a time. Another benefit was that the gaze-contingent cue prevented re-reading of the trigram.

Since some of my pilot data suggested that, given the opportunity, subjects tended to move their eyes to the invisible boundary without actually fixating on any of the words, I added an extra task to ensure that participants processed the meaning of the stimuli. On 5% of the trials (50 trials in all randomly interspersed with the other trials) participants were asked to create sentences using the most recently seen trigram and a cue word. These cue words were presented in the top left region of the screen after the trigram was removed from the display. An example of this type of trial would be: seeing the trigram *in the road* then seeing the cue word *fork*. A plausible utterance would be *There was a fork in the road.* Responses were manually scored by the authors as grammatically plausible (0), partially plausible (-0.5), or implausible (-1) by the experimenters. All participants had scores above 90% on the sentence creation task, so none were excluded. These sentence creation trials were removed from the data files and were not included in any of my analyses.

The stimuli were presented in white, fixed-pitch font (Courier New) on a black screen, following a fixation cross presented between two and three character spaces into the first word of the trigram. All stimuli were presented from the centre left of the screen, while instructions were presented centrally from the top, and key words for the sentence creation task were presented in the top left region of the screen. Data from two participants was not included due to technical errors during data acquisition.

All participants completed the task in under two hours, and none reported any adverse effects due to the duration of the experiment.

3.3.4 Data Preparation

I gathered information from the Google Web1T data set and other information from my experiment to create my initial list of predictors, listed in Table 3.1. Following a convention from Kuperman, Bertram, and Baayen (2010), when the abbreviation for a predictor is preceded by the letter **s**, the predictor has been centred and scaled (i.e. **PMI** becomes **sPMI**).

As noted by Matthews and Bannard (2010) and Tremblay and Tucker (2011), measures of lexical and n-gram probability are inevitably highly inter-correlated (for example, a trigram with a high frequency initial bigram will almost always have a high frequency word in it). The statistical analysis of experiments with highly inter-correlated predictors can be problematic, particularly when using multivariate regression. When left untouched, predictor co-linearity can cause estimates of slopes to be suppressed or enhanced (Friedman & Wall, 2005). Therefore, before proceeding with any inference, I analyzed the degree of multi-collinearity in the predictors [1]. Many of the predictors are highly correlated; in particular sPMI is correlated with the frequency of the first and second words and the frequency of the first bigram. The intra-experimental word frequency measures are also problematic: approximately 60% of all the words used in the trigrams I presented were seen more than once and so I expected the experimental frequency of the words in the 1000 trigrams to be strongly related to their corpus frequencies, which they were ($r > 0.73$, $p < 0.001$ for all three). I measured the degree of the problem by calculating the condition number, κ (Belsley, Kuh, & Welsch, 2004), which was $\kappa = 5.5e + 11$, where a condition number greater than 30 indicates a dangerous amount of collinearity. My approach to reducing multi-collinearity was to isolate a subset of predictors that were highly inter-correlated and were never going to be analyzed individually in my regression models. The these eleven predictors that I chose are listed in Table 3.1. I used Principle Component Analysis (PCA) to extract orthogonal principle components from this set, and after the application of PCA I found that the first five principle components

[1] A visualization of the multi-collinearity is given in Figure 3.7 in Section 3.10

explained 95% of variance in the predictors. The new set of predictors, with eleven predictors replaced by the five principle components, is shown in Figure 3.8. Strong correlations persisted between PMI and two of the principal components (PC1 and PC2) ($r = 0.50$ and $r = 0.59$). Despite these intercorrelations, the condition number fell to an acceptable level of $\kappa = 17$. The change in the pattern of intercorrelations after PCA is shown in Appendix 3.10.

Name	**Abbreviation**	**Description**	**Included in PCA**
N-gram Frequency	ngramfreq	The corpus frequency of the *n*-gram.	No
Pointwise Mutual Information	PMI	The discrepancy between the probability of the words occurring together given their joint distribution and the probability of their occurring together based only their individual distributions, assuming independence.	No
Second Word IC	w2IC	The average amount of information contained in the second word of a trigram, given the first word of the trigram as context.	No
Third Word IC	w3IC	The average amount of information contained in the third word of a trigram, given the first two words of the trigram as context.	No
Second Bigram IC	b2IC	The average amount of information contained in the second bigram of a trigram, given the first word of the trigram as context.	No
Total IC	TIC	The sum of the above three IC values for each trigram.	No
Closed Class Word	cc1, cc2, cc3	Closed class word in positions 1, 2 or 3	No
First Word Frequency	w1f	The corpus frequency of the first word.	Yes
Second Word Frequency	w2f	The corpus frequency of the second word.	Yes
Third Word Frequency	w3f	The corpus frequency of the third word.	Yes
First Bigram Frequency	b1f	The corpus frequency of the first bigram.	Yes
Second Bigram Frequency	b2f	The corpus frequency of the second bigram.	Yes
Third Bigram Frequency	b3f	The corpus frequency of the third bigram, the split-gram.	Yes
First Word Exp. Frequency	xfq1	The within-experiment frequency of the first word.	Yes
Second Word Exp. Frequency	xfq2	The within-experiment frequency of the second word.	Yes
Third Word Exp. Frequency	xfq3	The within-experiment frequency of the third word.	Yes
Ngram Length	length	Length of the *n*-gram in letters, including spaces.	Yes
Completeness	mncmplt	Mean rating of completeness.	Yes

Table 3.1: Inputs to the statistical models. All frequencies are log-transformed.
.

3.4 Statistical Methodology

In all of the following analyses, I used the following methodology. First, I tested for non-linear relationships between the predictors of interest. If there were any non-linear relationships, I selected to use Generalized Additive Models, or GAMs (Wood, 2006), to better understand my data (for an extensive exploration of GAMs in psycholinguistic experimental analyses, see Baayen, Kuperman, & Bertram, 2010). If there were no non-linear relationships, I used Linear Mixed Effects models instead (Bates, in preparation; Pinheiro & Bates, 2009).

I also checked each dependent measure for temporal inter-dependencies and whenever I found a strong correlation between the measurement of the dependent variable in a trial and same measurement in the previous trial, I added the previous trial's duration into all models. I also checked for practice effects, and when they existed, I included the standardized trial number in my analyses. This allowed me to explain any variability in the data due to experimental position.

A stepwise model selection process was applied during these analyses in which two nested models were compared to see which one had the best balance between fit and complexity. Predictors were added (or subtracted) one by one and only retained more complex models that were improvements over simpler models.

Finally to see if the models were being unduly impacted by outliers, I applied a model criticism method in which I refitted the model to a subset of the dataset that contained all data points that had residuals that were over 2.5 standard deviations larger or smaller than the mean of the residuals. In all of the following analyses neither the direction nor the reliability of the effects differed between the original model and the model with the data points that caused the residual outliers removed.

3.5 Results

In the eye-tracking paradigm both the location and timing of fixations is recorded, meaning that there are many possible dependent measures that I could attempt to predict with my set of predictors. My interest is the sensitivity of the reading system to probabilistic information contained within the trigrams, and so I chose to analyze measures that were likely to reflect the influence of this type of information. I first looked at the sum of all the fixations when reading the trigram, the most directly comparable measure to reading time as measured in many other studies. I then looked at the existence of regressive saccades in each trial as well as the total number of fixations in a trial. Finally, I broke the total time into two shorter intervals: the gaze time for the first word and the gaze time for the first and second words.

3.5.1 First Analysis: Total Duration

Our first analysis was of the total duration of the trial, the sum of the durations of all of the fixations. Based on the previous studies (Arnon & Snider, 2010; Bannard & Matthews, 2008; Matthews & Bannard, 2010; Tremblay & Baayen, 2010; Tremblay & Tucker, 2011), I predicted that there would be a facilitation for trigrams that were more frequent. The stimulus set was larger than the sets used in these previous studies, and contained a broader sampling of n-grams.

The reading durations in the data set were not normally distributed, violating the normality assumption of the statistical models. The standard log and inverse transformations did not produce normally distributed transformed reading durations. To find a better way to transform the reading times so that they would be approximately normally distributed, I used the method described by Box and Cox (1964) and found that by exponentiating the RTs to the power of 0.18 the RTs were approximately normally distributed. After transforming the reading times, the skewness, g_1, of the distribution was reduced from 1.2 to 0.3. In all references to the total reading time in the following analyses, transformed reading times were used.

Table 3.2: Model Comparisons for models predicting total reading time for a trigram. Δ_{AIC} denotes the change in AIC between two models.

	AIC	Δ_{AIC}	Relative Model Likelihood
Model 1: Random Intercepts for Participant and fixed effects of PC1-PC5, sTrial and PrevTrialDur	-13617		
Model 2: Model 1 + cc2	-13699	-83	8.4e+17
Model 3: Model 2 + cc3	-13721	-22	5.8e+04
Model 4: Model 3 + *n*-gram frequency $\otimes$ sPMI	-13974	-252	6.1e+54
Model 5: Model 4 + sTIC $\otimes$ sPMI	-14043	-70	1.3e+15

I created a base model, Model 1, which was made up of an additive combination of the sTrial, PrevTrialDur, and the principal components PC1 to PC5 (cc1 did not contribute in any of the models and was therefore removed from all models). This model also contained random intercepts for participants. I then compared this base model to models that were identical except that it included additive and non-linear interactive effects for *n*-gram frequency, information content and PMI. I compared these models to Model 1, and when the new model was superior to the simpler model as judged by a Log Likelihood Ratio Test (LLRT), it was retained.

A comparison of the relative fitness of all the models is shown in Table 3.2. To make clearer the amount of improvement in these models, I transformed the difference in the AIC values for these nested models (Δ_{AIC}) into a measure of relative model likelihood. Relative model likelihood is calculated using the difference in the AIC values for the models, as described by Burnham and Anderson (2002).

$$\text{Relative Model Likelihood} = e^{\frac{-\Delta_{AIC}}{2}}. \tag{3.4}$$

In the first step, Model 2 improved on Model 1 by adding linear effects for the lexical class of the second word, cc2. Model 3 added the effect of cc3, and then in Model 4 I added a non-linear interaction (called a *tensor product smooth* in GAMs[2]) between *n*-gram frequency and pointwise mutual information. In Model 5, the best model, I added a tensor product smooth for TIC and PMI. All other non-linear interactions were investigated, but none increased the

[2]This is a smooth surface that is a three-dimensional version of a smooth curved line. The symbol used to denote the tensor product is $\otimes$.

model likelihood. TIC improved the fix of the best model, and later additions of w2IC, w3IC and b2IC did not, so they were left out of all models.

	β	SE(β)	t	$p_{\|t\|}$
Intercept	2.577	0.030	85.07	0
PC1	0.023	0.002	11.24	3.4e-29
PC2	0.037	0.002	19.08	2.2e-80
PC3	-0.011	0.001	-9.65	5.4e-22
PC4	-0.022	0.002	-12.76	4.1e-37
PC5	-0.026	0.002	-13.18	1.8e-39
sTrial	-0.042	0.001	-32.49	5.8e-225
PrevTrialDur	0.271	0.007	40.35	0
cc2	0.033	0.004	8.07	7.7e-16
cc3	-0.017	0.004	-4.14	3.5e-05

Table 3.3: Coefficients for linear predictors in the best fitting GAM for the total reading duration of the trigrams.

The parametric coefficients for Model 5 are shown in Table 3.3. The principal components PC1 to PC5 all contributed to the model fit, showing that word and bigram frequencies, length and *n*-gram completeness were involved in determining reading time. PC4 is heavily loaded on the completeness rating with a positive correlation, and since the coefficient for PC4 in this model is negative, trigrams that were rated higher in completeness were read faster. There were strong effects of trial number (negative, showing a practice effect) and reading time on the previous trial (positive, showing the a spillover effect of reading time to the next trial). The open/closed class category of the first word was eliminated during model selection, but the lexical class of the second and third words remained in the final model. Having a closed class word in the second position of a trigram caused the reading time to increase while having a closed class word in the third position caused it to decrease. There were no linear effects of *n*-gram frequency effect, PMI or TIC, but there were non-linear interactions for all three, as we shall see.

Two non-linear interactions were found during model selection. The estimated degrees of freedom for these tensor product terms and the smooth

Table 3.4: Model Coefficients for smooth predictors in the best fitting GAM for the total reading duration of the trigrams. The ⊗ symbol denotes the tensor product.

	Estimated Df	Estimated Residual Df	F	$p_{bayseian}$
N-gram Frequency ⊗ sPMI	14.83	17.26	8.30	8.4e-22
sTIC ⊗ sPMI	11.05	13.33	5.98	1.8e-11
Random Intercept for Subject (Spline)	17.92	18.00	167.10	0

term for the random intercepts for participants are shown in Table 3.4. The first non-linear interaction was between *n*-gram frequency and PMI, visualized in Figure 3.2A, the largest slowdown occurring when the log transformed frequency was between -2 and -4. This pattern is suggestive of an interference effect: if PMI reflects an *n*-gram's cohesiveness, then a lack of cohesiveness interfered with the facilitation that comes from exposure and entrenchment, except for the lowest frequency trigrams, where high PMI caused a slight slowdown.

The second non-linear interaction, shown in Figure 3.2B, was between TIC and PMI. From a visual inspection of the contour plot, it appears that low TIC *n*-grams were read more slowly at the extremes of the range sPMI. When sPMI was not extreme (between -1 and 1), trigrams were read the fastest, independent of their TIC. When the sPMI was smallest (-1 to -2), both high TIC and low TIC trigrams were read the slowest. When ssPMI was high (greater than 2) and sTIC was low (less than 0), there was another slowdown. The forces driving these relationships will be explored further in the last part of this chapter, Section 3.6.

3.5.2 Discussion

This result provides evidence that my model predicting *n*-gram reading time had to be more complex than those presented by Arnon and Snider (2010), Siyanova-Chanturia et al. (2011) and Tremblay and Tucker (2011). Both *n*-gram frequency and TIC interacted with PMI in my model. This is evidence that all of these probabilistic measures are important to the reading system — leaving any of them out would create an incomplete model. The way that

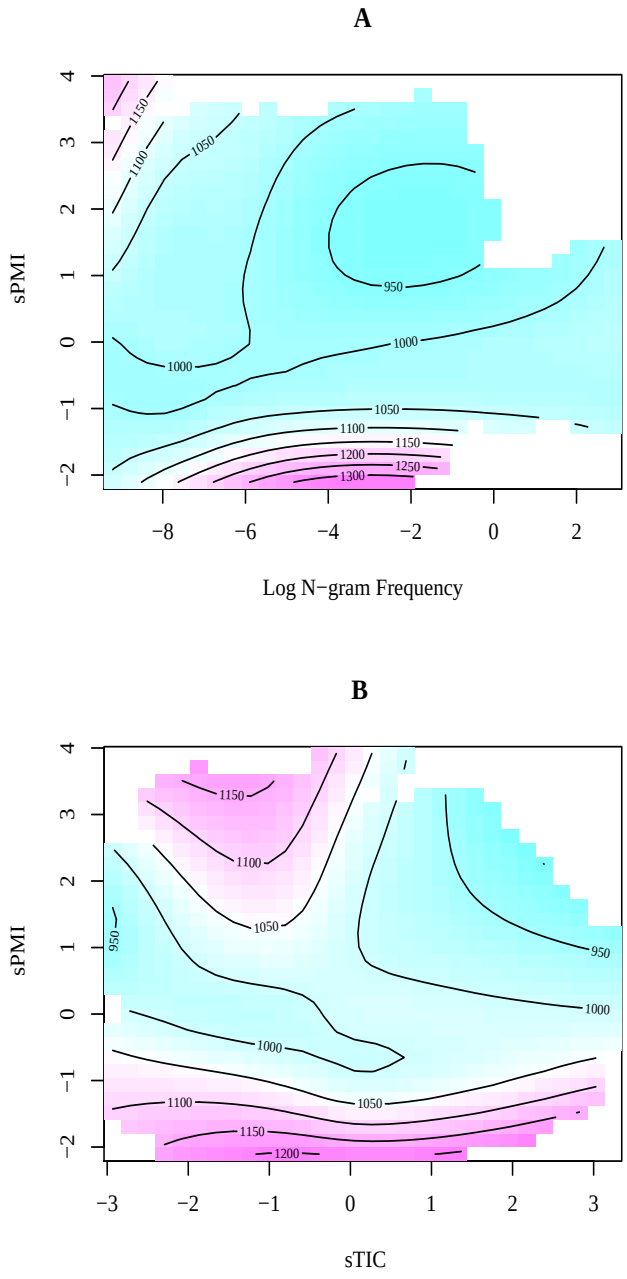

Figure 3.2: A) Partial Effects of *N*-gram Frequency and Pointwise Mutual Information Tensor Product Smooths on Total Reading Time. B) Partial Effects of Total Information Content and Pointwise Mutual Information Tensor Product Smooths on Total Reading Time. The dependent measure has been transformed back to milliseconds before plotting, using the reverse of the Box-Cox transformation.

they interact can provide insight into how information, cohesiveness and entrenchment come together to produce the observed behaviour. Importantly, these effects were part of a multivariate model that included linear effects for the other predictors: whether the n-gram is judged to be a constituent or not, the word and bigram frequencies, and the length of the n-gram.

From total reading time I will move on to the second dependent measure, regressive saccades.

3.5.3 Second Analysis: Regressive Saccades

The existence of regressive saccades (both intra-word and inter-word) within a trial was another outcome I examined. Regressive saccades occur when the first pass at reading does not give sufficient information to the reader (Vitu & McConkie, 2000). The location of the beginning and end point of the regressive saccade are also interesting to analyze. Probabilistic information about the n-grams might be able to help predict the location of the regressive saccade, but due to the complexity of the analysis, it was not attempted. Would my measures of probability and information content be predictive of the existence of regressive saccades? As with my analysis of reading duration, I proceeded with a similar forward stepwise model selection analysis, attempting to add all of my predictors to increasingly complex models. The dependent measure was a binary variable that I set to 1 if there were one or more regressive saccades in a trial. Most of the trials (74.9%) contained no regressive saccades, 21.6% contained one, 3.1% contained two, 0.4% contained three and 0.1% contained four, meaning that 25.1% of the trials had one or more saccades.

Since the outcome is binary in nature (no regressive saccades vs. one or more regressive saccades), I proceeded to use logistic GAMs with a logit link function instead of the Gaussian type. I also included a random intercept for each subject in all of the models, but did not include random intercepts for each item as this did not improve the fitness of any of the models.

To begin the stepwise forward model selection, I built a base GAM model, Model 1, using only the parametric model elements (PC1-PC5, cc1-cc3, sTrial

Table 3.5: Model Comparisons for models predicting probability of one or more regressive saccades in a trial. ΔAIC denotes the change in AIC between two models.

	AIC	Δ_{AIC}	Relative Model Likelihood
Model 1: Random intercepts for participant, PC1-PC5, cc3, sTrial and Previous Trial Regressive Saccade	17363		
Model 2: Model 1 + *n*-gram frequency ⊗ sTIC	17342	-21	4e+04
Model 3: Model 2 + sPMI ⊗ sTIC	17268	-74	1e+16

and Previous Trial Regressive Saccade) and the random intercepts for each subject. I then added interactions of interest one by one, and only retained models that were superior to a simpler model when a Log Likelihood Ratio Test (LLRT) was performed. A listing of all these nested models and their AIC measure of model fitness is shown in Table 3.5. After each new term is added, I report the relative model likelihood of each model when compared to the previous model. Each of the non-linear interactions added caused the new model to be over 100 times more likely than the previous model, confirming the relevance of all of the interactions to improving the final model. Model 2 added the non-linear interaction of trigram frequency and total information content to Model 1. Finally, Model 3 added the non-linear interaction of pointwise mutual information and total information content to Model 2. First I will explain the parametric effects of the predictors shown in Table 3.6, and then move on to the non-linear interactions listed in Table 3.7.

Each of the principal components made a contribution, accounting for the effects of unigram frequencies, bigram frequencies, length and completeness. The effect of having a closed class word in the first or second positions increased the probability of making a regressive saccade, whereas having a closed class word in the third position was inhibitory for regressive saccades. There was also a decrease in the probability of making a regressive saccade as participants progressed in the experiment, likely a practice effect. There was an increase in the probability of making a regressive saccade when a participant made one or more regressive saccades on the immediately preceding trial, an inter-

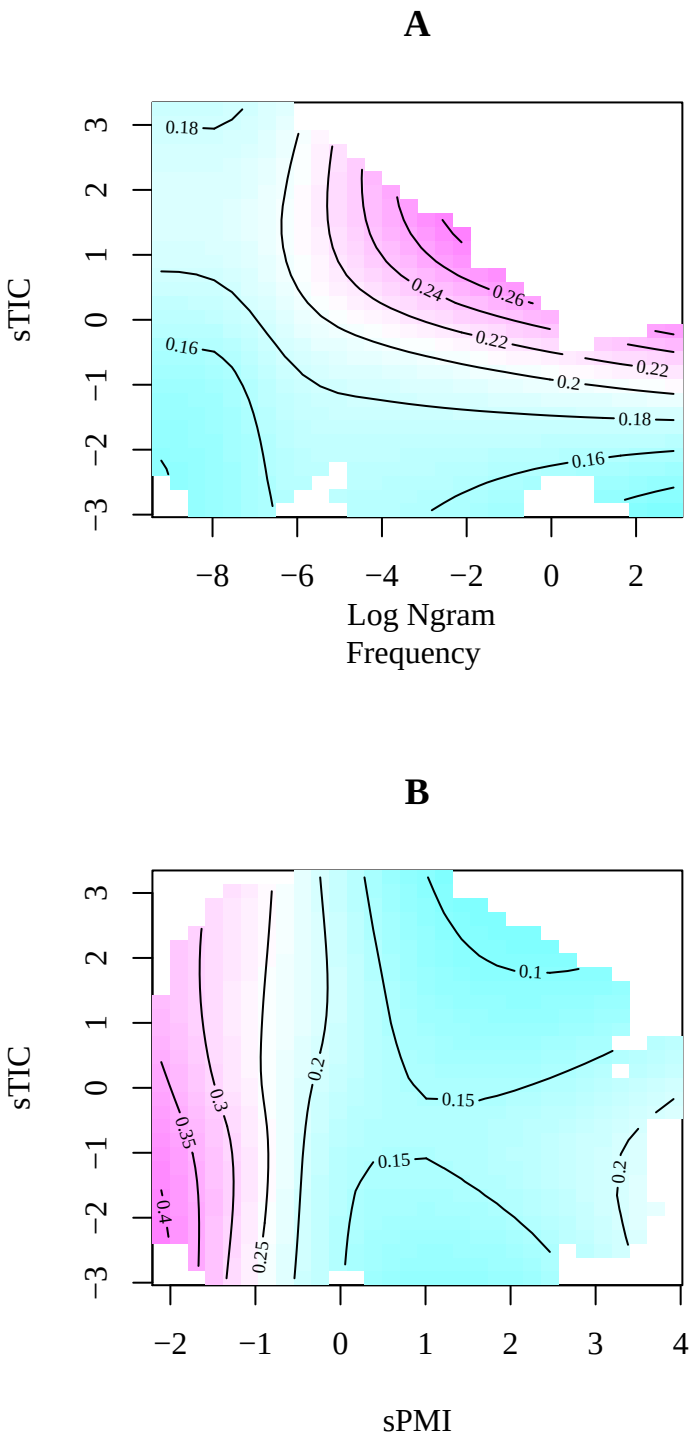

Figure 3.3: Partial effects on the probability of a regressive saccade during a trial for A) Interaction between *n*-gram frequency and sTIC. B) Interaction between sPMI and sTIC. The dependent measure was transformed from logits to probabilities before plotting.

	$\bar{\beta}$	SE(β)	z	$p_{\|z\|}$
Intercept	-1.453	0.194	-7.48	7.7e-14
PC1	0.138	0.025	5.62	1.9e-08
PC2	0.134	0.026	5.17	2.4e-07
PC3	-0.091	0.020	-4.54	5.7e-06
PC4	-0.195	0.025	-7.78	7.2e-15
PC5	0.263	0.029	8.93	4.1e-19
cc1	0.239	0.061	3.92	8.8e-05
cc2	0.388	0.062	6.22	5e-10
cc3	-0.485	0.062	-7.80	6.2e-15
sTrial	-0.369	0.020	-18.93	7e-80
Regr Sacc on Prev Trial	0.238	0.043	5.58	2.4e-08

Table 3.6: Model Coefficients for linear predictors in Regressive Saccade probability GAM

trial spillover effect. The predictors related to the probabilistic measures for the whole *n*-gram frequency and information content were only predictive when entered into non-linear interactions. The first interaction that was added to the base model was an interaction between *n*-gram frequency and TIC. Visualized in Figure 3.3A, this surface has a peak in regressive saccade probability when the sTIC is high (between 0 and 3), with a peak for trigrams that have a log frequency of around -2. The probability of a regressive saccade was lower for very low frequency words (log frequency $\leq$ -6) across the range of TIC values.

	Estimated Df	Estimated Residual Df	F	$p_{bayseian}$
N-gram Frequency $\otimes$ sTIC	6.01	7.37	64.34	3.2e-11
sTIC $\otimes$ sPMI	7.78	10.25	78.72	1.2e-12
Random Intercept for Subject	17.79	18.00	1331.96	5.7e-272

Table 3.7: Regressive Saccade Model Parameters for smooth predictors in the GAM

The next non-linear interaction detected was between PMI and TIC. This surface, shown in Figure 3.3B, has a peak probability when sTIC is less than -2 and the sPMI is below -1. When sPMI is between -0.5 and -2, the probability falls for all values of sTIC. Less cohesive trigrams that are less predictable induced regressive saccades more often than other trigrams.

3.5.4 Discussion

Again, as I saw with the total duration analysis, TIC had an interactive relationship with PMI. Unlike in the total duration analysis though, TIC did interact with frequency in the best model of predicting regressive saccades. Interactive relationships trumped simple relationships during my stepwise modelling and frequency, PMI and TIC were both involved. *N*-grams with a higher TIC and a medium or high frequency were more likely to cause regressive saccades. The familiar *n*-grams with higher average surprisal may have forced a reassessment of the stimulus. Low PMI generally increased the likelihood of a regressive saccade, with a small modulation based on the TIC.

How good is my model at discriminating the observed existence of saccades using the model's predictions? The Somers' D_{xy} Rank Correlation for my model is 0.5, 95% CI: 0.49 - 0.52 and the ROC-AUC (Receiver Operating Characteristic - Area Under Curve) is 0.75, 95% CI: 0.74 - 0.76, a fair level of discrimination.

The next dependent measure I will analyze is the total number of fixations in each trial.

3.5.5 Third Analysis: Number of Fixations

The number of fixations made during the reading of a trigram is another measurable variable of interest. Would my measures of probability and information help me predict how many fixations each participant would make on each trigram? The analysis of the fixation counts follows.

To make sure that my analysis was not biased by the number of regressive saccades in each trial, I subtracted the number fixations after regressive saccades in each trial from the total number of fixations. The median number of fixations was 5 and the standard deviation was 1.57. The distribution of these fixation counts was still skewed, and so I applied a log transformation to the fixation counts. After transforming the fixation counts the skewness, g_1, of the distribution was reduced from 1.2 to 0.093. This is the dependent variable

Table 3.8: Model Comparisons for models predicting total fixations for a trigram. ΔAIC denotes the change in AIC between two models. All random slopes were for the random effect of subject.

	AIC	Δ_{AIC}	Relative Model Likelihood
Model 1: Random intercepts for Participants and Items with random slopes for PrevTrialDur and sTrial	-214		
Model 2: Model 1 + random slopes for sPMI and sPMI	-248	-34	3e+07
Model 3: Model 2 + cc2	-272	-24	1e+05
Model 4: Model 3 + *n*-gram frequency	-282	-10	1e+02
Model 5: Model 4 + sPMI × cc2	-292	-10	1e+02

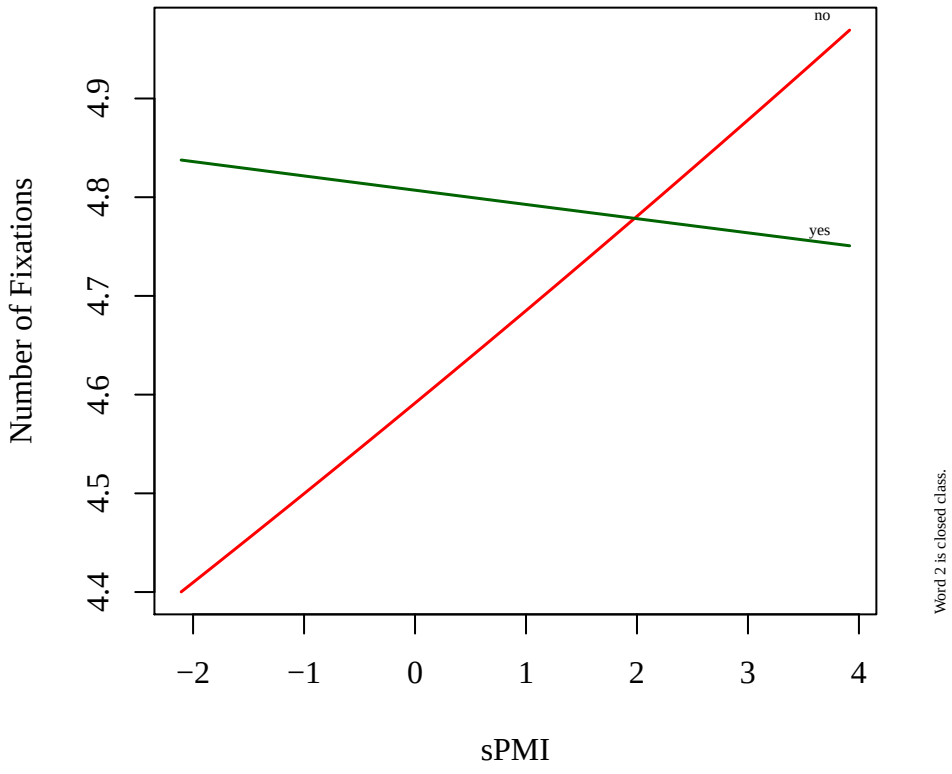

Figure 3.4: Partial effect of the interaction between Pointwise Mutual Information and class of second word in predicting the number of fixations.

that I entered into all of the models.

I added random slopes per subject for predictors along with the predictors themselves to find out if the effect was generalizable. In this process, I retained three new random slopes for each subject: Pointwise Mutual Information (sPMI), longitudinal effects (sTrial) and previous trial duration (PrevTrialDur). In Model 3, the closed class status of the second word was added, causing an increase in model likelihood. In Model 4, *n*-gram frequency was added, and in the final model, Model 5, an interaction between sPMI and cc2 was added, which was the best model of all.

Table 3.9: MCMC-based estimates for the coefficients for the fixed effects in the linear mixed effects model fitted to the observed total fixations.

	Estimated β	β_{MCMC}	HPD lower	HPD upper	p_{MCMC}
Intercept	0.7299	0.7256	0.6296	0.8252	0.0002
n-gram frequency	-0.0072	-0.0071	-0.0117	-0.0027	0.0012
PC1	0.0226	0.0227	0.0156	0.0290	0.0002
PC2	0.0452	0.0454	0.0372	0.0529	0.0002
PC3	-0.0146	-0.0146	-0.0194	-0.0100	0.0002
PC4	-0.0225	-0.0225	-0.0288	-0.0151	0.0002
PC5	-0.0633	-0.0634	-0.0717	-0.0553	0.0002
sTrial	-0.0344	-0.0344	-0.0569	-0.0130	0.0040
PrevTrialDur	0.2084	0.2096	0.1843	0.2361	0.0002
sPMI	0.0202	0.0199	0.0040	0.0363	0.0176
cc2	0.0459	0.0459	0.0284	0.0631	0.0002
sPMI × cc2	-0.0232	-0.0232	-0.0360	-0.0114	0.0002

The fixed effect coefficients for Model 5 are shown in Figure 3.9. As with total duration, PC1-PC5, sTrial and PrevTrialDur all contributed to explaining variability. Above and beyond all of these predictors, the two predictors of greatest interest are cc2 and *n*-gram frequency. There was an increase in the number of fixations when the second word was a closed class word. Finally there was a decrease in the number of fixations for *n*-grams of higher frequency. The partial effect of the interaction term in the model is shown in Figure 3.4.

The estimated standard deviations for all of the random effects in my model are reported in Table 3.10 along with the 95% highest posterior density credible intervals from the MCMC simulations. All of the standard deviations are within the 95% HPD intervals.

Table 3.10: MCMC-based estimates for the random effects in the linear mixed effects model fitted to the observed number of fixations.

	Standard Dev	HPD lower	HPD upper
Random Intercept: Item	0.072	0.060	0.070
Random Intercept: Subject	0.049	0.014	0.057
Random Slope: PrevTrialDur for Subjects	0.044	0.032	0.063
Random Slope: sTrial for Subjects	0.014	0.009	0.023
Random Slope: sPMI for Subjects	0.209	0.087	0.197
Residual	0.232	0.230	0.235

3.5.6 Discussion

Our best model predicting the number of fixations showed linear effects of frequency and PMI, but no interactions between frequency and PMI, and no involvement of any information content variable at all. Why did none of information content predictors improve the quality of the model during the stepwise model selection? It seems logical that more predictable trigrams should have fewer fixations, but I did not find any such effects. This is quite different from the situation with both total duration and regressive saccades.

There was more efficient reading (fewer fixations) for high frequency *n*-grams, a replication of the effect found by Siyanova-Chanturia et al. (2011). The interaction of PMI and cc2 is of primary interest because when the second word was a closed class word, PMI had almost no effect on the number of fixations. When the second word was an open class word, higher PMI trigrams had more fixations. It is unclear why this should be but I can speculate that the coherency of trigrams created the need for caution during the planning of the saccades. The care taken during the reading of the coherent trigrams may have influenced the number of fixations.

The final group of variables I analyzed were the sub-gazes. Considering the trigrams as unitary wholes, the full gaze time should include all of the fixations made during the reading of that trigram. I divided this gaze into sub-gazes in much the same way that other researcher have done when looking at the sub-gazes made within a compound word (Kuperman et al., 2008). I hypothesized that there would be an unfolding of information that could be detected by modeling the sub-gazes. I defined two sub-gazes, the first sub-gaze (SG1) and the second sub-gaze (SG2). SG1 is the sum of the fixations on the first word

in the trigram before there are any intra-word regressive saccades. SG2 is the sum of the fixations on the first and second words in the trigram before there are any inter-word regressive saccades or intra-word regressive saccades.

3.5.7 Fourth Analysis: First Sub-gaze

The analysis of the sub-gaze measures differs slightly from the previous analyses in that new predictors for word length. All of my previous analyses included the number of letters in the whole trigram, but I wanted to account for the effect of the length of the first word in my analysis of SG1, the sum of fixations on the first word. I added this predictor to my standard set predictors and called it SG1Len, with the centered and scaled version called sSG1Len. All the other predictors in these models are the same predictors described in Table 3.1. To confirm that I did not increase the amount of multi-collinearity by adding this predictor, I re-calculated κ for all of my predictors, and it was 50, which is high and could lead to an increased risk of suppression or enhancement.

Friedman and Wall (2005) have looked at regression when predictors are highly correlated, and they note that it is often beneficial to include predictors that are inter-correlated in a regression model. They note that "it is reasonable to consider highly correlated independent variables." (Friedman & Wall, 2005, p.135). To test for any impact of suppression and enhancement due to the collinearity of sSG1Len and PC3 ($r = 0.62$) on my regression coefficients, I compared all models with a sub-model that did not contain PC3. In all cases neither the direction nor the reliability of the effects of the remaining predictors in the models changed, indicating that the collinearity was acceptable.

I found a correlation between the first sub-gaze time and and the total reading time on the trial before it (r =0.13, 95% CI: 0.11 - 0.14). I did not find a relationship between SG1 and the position in the stimulus list (r =-0.0022, 95% CI: -0.017 - 0.014). I added random slopes for each subject for both of these predictors in all of the models, and despite the lack of a correlation in the aggregate, the random effect of sTrial for subjects was a beneficial predictor in all models. The fixed effect of sTrial, though, did not contribute, and was

Table 3.11: Model Comparisons for models predicting SG1 for a trigram. ΔAIC denotes the change in AIC between two models. All random slopes were for for the random effect of subject.

	AIC	Δ_{AIC}	Relative Model Likelihood
Model 1: Random intercepts for Participants and Items, sSG1Len and random slopes for same	10036		
Model 2: Model 1 + random slopes for sTrial	9831	-205	3e+44
Model 3: Model 2 + sPMI and random slopes for same	9818	-14	9e+02
Model 4: Model 3 + PrevTrialDur and random slopes for same	9691	-127	3e+27
Model 5: Model 4 + PC1, PC2, PC3, and PC5	9298	-393	2e+85
Model 6: Model 5 + cc1	9287	-11	3e+02
Model 7: Model 6 + Ngram Freq	9274	-13	7e+02
Model 8: Model 7 + sPMI × cc1	9255	-18	8e+03

dropped during model selection.

Table 3.12: MCMC-based estimates for the coefficients for the fixed effects in the linear mixed effects model fitted to the observed SG1.

	Estimated β	β_{MCMC}	HPD lower	HPD upper	p_{MCMC}
Intercept	5.0802	5.0800	4.9636	5.2034	0.001
sPMI	0.0653	0.0649	0.0479	0.0825	0.001
cc1	-0.0395	-0.0390	-0.0587	-0.0187	0.001
n-gram frequency	-0.0089	-0.0087	-0.0131	-0.0045	0.001
PC1	-0.0407	-0.0405	-0.0474	-0.0333	0.001
PC2	-0.0167	-0.0166	-0.0255	-0.0085	0.001
PC3	0.0353	0.0353	0.0285	0.0415	0.001
PC5	0.0321	0.0321	0.0228	0.0414	0.001
sSG1Len	0.0978	0.0980	0.0828	0.1164	0.001
PrevTrialDur	0.0780	0.0781	0.0670	0.0899	0.001
sPMI × cc1	-0.0323	-0.0321	-0.0463	-0.0201	0.001

Before adding any fixed effects, I added random subject effects for certain predictors. In this process, I retained four new random slopes for each subject: the effect of the length of the first word (cSG1Len), the effect of the position in the experiment (sTrial), the effect of the trigram's pointwise mutual information (sPMI) and the duration of the previous trial (PrevTrialDur). As can be seen from Table 3.11 the addition of these random slopes greatly improved the nested models. I continued to add predictors one by one, but for brevity's sake, I report a smaller number of models here, grouping similar predictors. In Model 5, I added the first set of fixed effects: the effects of first word length, previous trial duration and closed/open class category of the first word (cc2 and cc3 did not contribute anything). These three predictors improved the

Table 3.13: MCMC-based estimates for the random effects in the linear mixed effects model fitted to the observed SG1.

	Standard Dev	HPD lower	HPD upper
Random Intercept: Item	0.069	0.057	0.069
Random Intercept: Subject	0.081	0.021	0.138
Random Slope: sPMI for Subjects	0.021	0.015	0.033
Random Slope: sSG1Len for Subjects	0.037	0.027	0.057
Random Slope: sTrial for Subjects	0.015	0.008	0.025
Random Slope: PrevTrialDur for Subjects	0.011	0.001	0.016
Residual	0.307	0.305	0.312

model, despite the fact there were already random slopes for PrevTrialDur and sSG1Len in the model. Also, the position in the experiment, sTrial, did not improve the model, and so it was left out. The second of the standard predictors to be dropped during the stepwise forward modeling was PC4. In Model 6, PC4 did not contribute to improving the fitness of the model. To help understand this fact, I point to the loading for this principle component, shown in appendix 3.8. PC4 is most strongly correlated with the mean completeness rating. In the context of the first sub-gaze, the lack of a contribution from the completeness of the trigram is sensible, as the participants have not yet seen much of the second or third word. In Model 7 I added only n-gram frequency, PMI and first bigram information content, as all the other information-related predictors did not improve the model fitness. The final model, Model 8, added an interaction between the PMI for the trigram and the class of the first word. This was the best fitting model found, and I will now report the parameter estimates for this model.

In Table 3.12, the estimated coefficients for all of the fixed effects are shown along with the MCMC simulation results. All of the 95% highest posterior density credible intervals did not contain zero, showing that none of the fixed effects were null effects. The partial effects of all of these predictors (except for PC1, PC2, PC3 and PC5) are shown in Figure 3.5. Due to the combination of effects in each of the principle components, I can only report that my model accounted for variability due to the inputs to the PCA: word and bigram frequency, trigram length and completeness. The effect of increasing PMI was an increase in sub-gaze time. When the first word was a closed class word, all

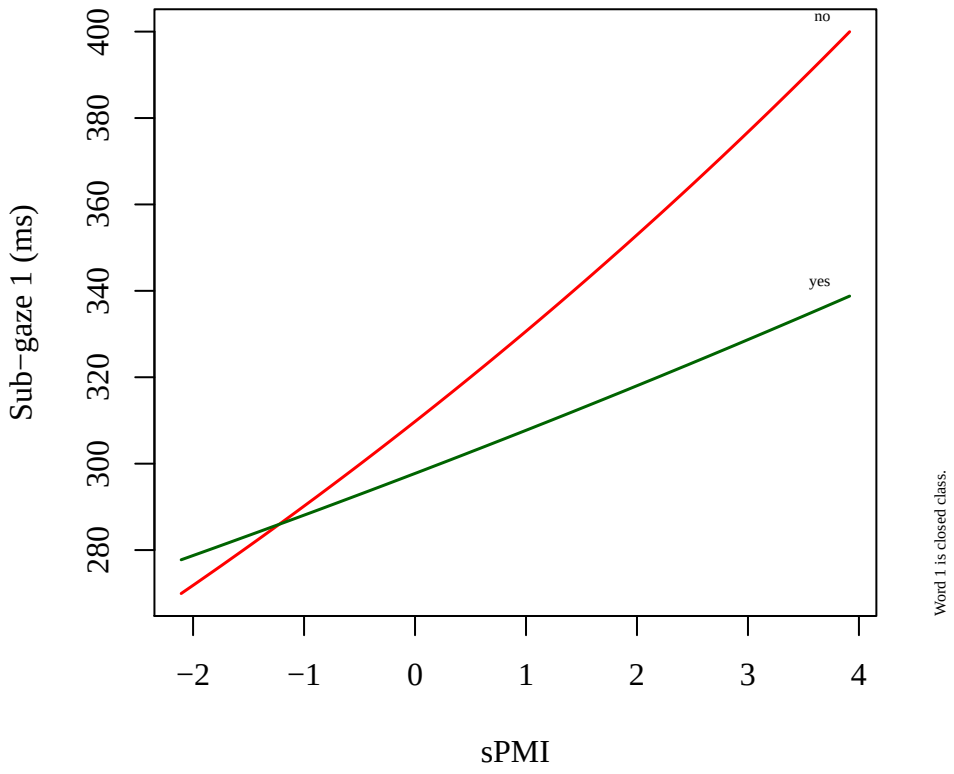

Figure 3.5: Partial effect of the interaction between Pointwise Mutual Information and class of first word in predicting SG1.

other things, including word length, being equal, subgaze time was shorter.

Another early effect of a whole *n*-gram property was found: the first subgaze was shorter for trigrams that were more frequent. The largest of the fixed effects was the length of the first word, with shorter words having a shorter sub-gaze than longer words.

There was also an inter-trial temporal dependency, with a slowdown in the first sub-gaze in trials that were preceded by a slow trial.

The final fixed effect was the interaction between the class of the first word and the PMI of the trigram. For trigrams with high PMI, the trigrams that started with a closed class word had a shorter first word sub-gaze than those that started with an open class word. The sub-gaze times were not influenced by the class of the first word for low-PMI trigrams (Figure 3.5).

The estimated standard deviations for all of the random effects in my model are reported in Table 3.13 along with the 95% highest posterior density credible intervals from the MCMC simulations. All of the standard deviations are will within the 95% HPD intervals.

3.5.8 Discussion

The early impact of probabilistic measures for the whole sequence (*n*-gram frequency and PMI) show how sensitive the visual system is to the context around words. The amount of information about the second word that is available during the reading time for SG1 is not large, and yet I found whole-trigram effects. The PMI of the trigram influenced the reading time of the first word in the trigram, even after taking into account the frequency and lexical class of the first word. In the same way, a trigram with a greater frequency had a faster reading time of the first word than one with a lower frequency. These results imply that coherence, entrenchment and predictability of the trigram have a very early influence on the reading of the trigram.

The interaction between cc1 and PMI implies a very early interaction between lexical class and coherence. *N*-grams with open class words in the first position had a larger slowdown due to PMI than *n*-grams with closed class words in the first position. As with the fixation data, the coherency of tri-

grams made the readers proceed with caution.

I now move on to the final dependent measure, the second sub-gaze, SG2.

3.5.9 Fifth Analysis: Second Sub-gaze

The second sub-gaze, SG2, included all the fixations on the first and second words in the trigram before fixating on the third word or before any regressive saccades. My methodology was identical to the one used for SG1.

The fixed effect of sTrial did not contribute to the model, and was dropped during model selection. I also checked to see if the amount of multi-collinearity had changed when I added the two new predictors, sSG2Len and sW3Len, and as with SG1, the condition number was above the acceptable amount ($\kappa = 50$). As with SG1, I tested for any impact of suppression and enhancement due to the collinearity of sSG2Len and PC2 ($r = 0.78$) on the regression coefficients. I compared all models with a sub-model that did not contain PC2. As with SG1, neither the direction nor the reliability of the effects of the remaining predictors in the models changed when PC2 was removed, indicating that the collinearity was not distorting my results.

The steps in my model comparison are listed in Table 3.14. As with SG1, my first model contained no fixed effects, only the random intercepts for subjects and items and random slopes for the centered measure of the number of letters in the first two words, sSG2Len. The next three models added random slopes for sTrial, PrevTrialDur and the standardized length of the third word, sW3Len. All of these random effects improved model fitness. Model 5 added the fixed effects of sSG2Len, PrevTrialDur and sW3Len, and these fixed effects explained variability above and beyond the random effects of these predictors. The next model included fixed effects for cc1 and cc2. The predictor cc3 was dropped at this point as it did not improve model fitness, perhaps due to the lack of lexical access for the third word. PC2, PC3 and PC5 were also added at this point, but PC1 and PC4 were dropped from the model. Since all the words in the trigram had not been read yet, the reason for the lack of an impact for PC4, which was loaded on phrasal completeness, seems clear. Why did PC1

Table 3.14: Model Comparisons for models predicting SG2 for a trigram. ΔAIC denotes the change in AIC between two models. All random slopes were for for the random effect of subject.

	AIC	Δ_{AIC}	Relative Model Likelihood
Model 1: Random intercepts for Participants and Items, Random Slopes for sSG2Len	8106		
Model 2: Model 1 + Random Slopes for sTrial	7817	-289	5e+62
Model 3: Model 2 + Random Slopes for PrevTrialDur	7558	-259	2e+56
Model 4: Model 3 + Random Slopes for sW3Len	7023	-535	1e+116
Model 5: Model 4 + sSG2Len, PrevTrialDur and sW3Len	6880	-144	1e+31
Model 6: Model 5 + cc1, cc2, PC2, PC3, and PC5	6742	-138	8e+29
Model 7: Model 6 + sb2IC	6730	-13	6e+02
Model 8: Model 7 + sPMI × cc2	6676	-53	4e+11
Model 9: Model 8 + sPMI × *n*-gram freq	6666	-10	1e+02

drop out at this point? Perhaps it can also be explained by the loadings: PC1 loaded most heavily on variables related to the frequency of the third word in the trigram, w3f and xfq3. Since the third word has not been read yet there is no way for its frequency to impact the reading time.

Table 3.15: MCMC-based estimates for the coefficients for the fixed effects in the linear mixed effects model fitted to the observed SG2.

	Estimated β	β_{MCMC}	HPD lower	HPD upper	p_{MCMC}
Intercept	5.4725	5.4721	5.3259	5.5965	0.002
sPMI	-0.0126	-0.0130	-0.0355	0.0126	0.284
n-gram frequency	-0.0008	-0.0008	-0.0054	0.0040	0.764
sb2IC	-0.0197	-0.0197	-0.0297	-0.0077	0.002
cc1	-0.0438	-0.0437	-0.0652	-0.0239	0.002
cc2	0.0350	0.0351	0.0146	0.0564	0.004
PC2	0.0332	0.0334	0.0234	0.0428	0.002
PC3	-0.0167	-0.0165	-0.0239	-0.0080	0.002
PC5	0.0262	0.0259	0.0105	0.0411	0.002
PrevTrialDur	0.1387	0.1390	0.1192	0.1569	0.002
sSG2Len	0.1347	0.1345	0.1130	0.1534	0.002
sW3Len	-0.0817	-0.0822	-0.1013	-0.0664	0.002
sPMI × *n*-gram freq	-0.0056	-0.0056	-0.0087	-0.0028	0.002
sPMI × cc2	-0.0629	-0.0628	-0.0800	-0.0494	0.002

The next model included a crucial predictor, the second bigram's informa-

tion content. This number represents the predictability of the second bigram (word 2 and word 3) based on the first word. All the other information content predictors were dropped at this stage, including TIC, as they could not explain any more variability. The final two models added interactions, the first being an interaction between PMI and cc2. The second interaction was between sPMI and the *n*-gram frequency. Both of these interactions improved the models, bringing us to the final model, Model 9. The coefficients for this model are shown in Table 3.15. Only two of the of the 95% HPD intervals contain 0, the main effects of sPMI and *n*-gram frequency. These two fixed effects were left in the final model because both of these predictors were involved in interactions.

The amount of information in the second bigram facilitated the reading of the first two words. This result supports the argument that the predictability of words is influencing the performance of the reading system, even when many other inputs are being taken into account. Even though the eyes had not yet fixated on the third word, the ability of the first word to help predict the next two words was having an effect on the reading time up to that point. The effects of the lexical class of the first and second words are also clear – a closed class word in the first position led to a shorter SG2, whereas a closed class word in the second position let to a longer SG2, mirroring the results for total duration and fixation count. As in my previous analyses, there was an inter-trial spillover effect, with trials that were preceded by slowly processed trigram having a longer SG2. The effects of the combined length of the first and second words, SG2Len, was strong: trigrams with a larger SG2Len took longer to read. The effect of the length of the third, as yet unseen, word on SG2 was in the opposite direction, with shorter third words leading to slower reading of the first two words (this effect may have supplanted the effect of the cc3 variable, which dropped out during model selection). As for the interactions, the first one between PMI and trigram frequency is plotted in Figure 3.6A. As the frequency of the *n*-gram increased, the direction of the PMI effect changed from positive to negative. For very rare trigrams, the higher PMI trigrams were read more slowly than the lower PMI trigrams. For common trigrams,

the higher PMI trigrams were read faster than the lower PMI trigrams. The final interaction between the lexical class category of the second word and the PMI is shown in 3.6B. When there was a closed class word in the second position, the trigram's PMI had a facilitatory effect on SG2, whereas when the second word was not a closed class word, PMI had an inhibitory effect on SG2.

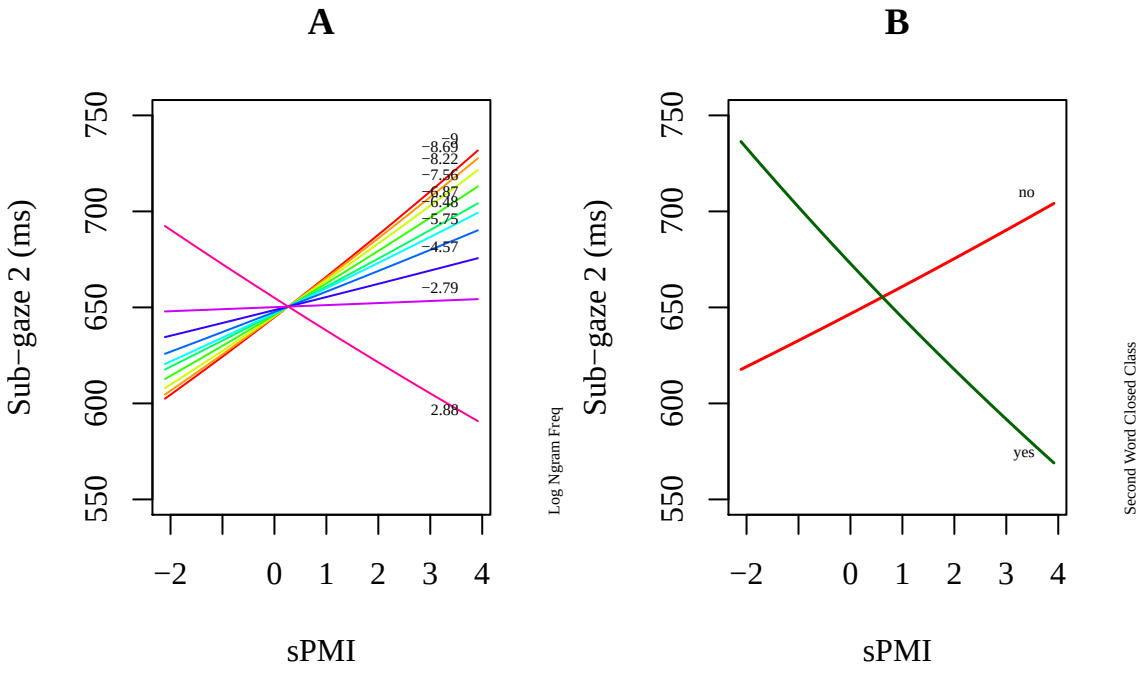

Figure 3.6: Partial effects of predictors in linear mixed effects model predicting SG2. A) Interaction between Pointwise Mutual Information and log n-gram frequency, B) Interaction between Pointwise Mutual Information and class of second word.

Discussion

As with the total duration data, the SG2 data showed an interaction between frequency and PMI. Even when the participants had not yet fixated on the final word of the trigram, the coherence and the entrenchment of the n-grams

Table 3.16: MCMC-based estimates for the random effects in the linear mixed effects model fitted to the observed SG2.

	Standard Dev	HPD lower	HPD upper
Random Intercept: Item	0.088	0.074	0.084
Random Intercept: Subject	0.313	0.120	0.284
Random Slope: sW3Len for Subjects	0.027	0.021	0.046
Random Slope: sSG1Len for Subjects	0.027	0.021	0.041
Random Slope: sTrial for Subjects	0.029	0.021	0.043
Random Slope: PrevTrialDur for Subjects	0.040	0.006	0.040
Residual	0.281	0.280	0.286

begin interacting. As with SG1, this effect points to fast, parallel dynamism in the reading system. After all of the other variables were entered in the model, these interactions were still relevant. I left the simple effects of PMI and *n*-gram frequencey in the final model because they were involved in interactions, but, just as I saw in the total reading duration analysis, there were no simple effects of PMI or frequency. As with SG1, the completeness information contained in PC4 did not contribute, and so PC4 was dropped out of the model. This shows some of the temporal unfolding in the reading of these trigrams: completeness ratings only entered into models that allowed for the viewing of the full *n*-grams.

As with the fixation counts and SG1, there was an interaction between lexical class and PMI, with the same increase of reading time for higher PMI trigrams when the second word was an open class word. The predictability of the last two words based on the first word, b2IC, also entered into this model, which again shows how the scope of SG2 gives preference to a more germane measure (b2IC) over a more global measure (TIC).

To help make sense of all the evidence presented here, I will now attempt to synthesize what have learned in the next section.

3.6 Conclusion

The current study presented university students with 1000 trigrams that covered a large swath of the frequency space for three-word combinations in English to examine the effect of probabilistic information on their eye movements. The statistical models included many measures including word frequencies, bi-

gram frequencies, completeness ratings, longitudinal effects, temporal dependencies and word class. Participants showed sensitivity across the full range of stimuli to *n*-gram frequency, information content and pointwise mutual information after taking into account all the other measures already included in the models. In contrast with other research on *n*-gram reading, the effect of *n*-gram frequency was not purely facilitatory. My experimental design and methodology allowed me to find regions of frequency space where lower frequency *n*-grams were read faster than higher frequency *n*-grams.

Some of the previous experiments that looked at the effects of *n*-gram frequency on *n*-gram reading did not include in their models all of the probabilistic measures I have used here. This leaves open the question of whether there were uncontrolled sources of variance driving performance, or whether the results could be generalized beyond the restrictive set of stimuli they drew from. The number and variety of covariates presented here is unparalleled, and even when all the covariates were accounted for, *n*-gram frequency effects were still in the best models.

Another criticism that could be made of earlier studies is the size of the stimuli sets and the selective sampling of stimuli. The size of the stimuli set I have used in this study and the method of randomly sampling stimuli from an extremely large list of trigrams allow me to say that the effects I have found generalize to the majority of trigrams. *N*-gram frequency was not the only probabilistic measure to aid in predicting eye movements. PMI and IC also made a contribution in some of these models. I will now discuss the contributions of each of these measures and relate them to my view of reading trigrams.

3.6.1 Contribution of Frequency

A linear effect of *n*-gram frequency was included in the best models for only two of the five measures: total fixations and SG1. In these two cases, *n*-gram entrenchment improved processing efficiency: reading was faster with fewer fixations. As for the other three dependent variables, *n*-gram frequency was part of an interaction term in the models of total duration, regressive saccades

and SG2. In both total duration and SG2, there is evidence for interference between entrenchment (frequency) and coherence (PMI). This interference may arise from a clash of expectations: high frequency *n*-grams that are incoherent interfere with fluent reading, as we see from a trigram with low PMI and high frequency, *increase in the*. There were no simple effects of frequency in any models for the other measures, but there were consistent interactions with other variables: frequency interacted with PMI in the models for total duration and SG2, and frequency interacted with TIC in my model for regressive saccades. The non-linear interaction was strongest in low PMI *n*-grams in my total duration data (Figure 3.2A). This frequency PMI interaction appeared again in the reading time for SG2, with a similar pattern: low frequency *n*-grams were read faster if they were coherent, whereas the same was not true for high frequency *n*-grams (Figure 3.6A). Why did Arnon and Snider (2010) and Siyanova-Chanturia et al. (2011) find a linear frequency effect on reading efficiency where I did not? One possible explanation is the lack of PMI in their experimental design. Since my stimuli contained *n*-grams with great variety of frequencies, and since I was looking for interactions between frequency and other probabilistic information, I was able to detect the interaction. The frequency-PMI interaction I found may have absorbed all the variability that could have been explained by frequency or PMI, making the main effects of frequency and PMI superfluous (as seen in the coefficients for frequency and sPMI in Table 3.15). The results presented here suggest that the effect of *n*-gram entrenchment (as measured by orthographic frequency) is consistently modulated by other aspects of the *n*-gram— the coherence and predictability.

How could whole trigram frequency influence SG1? Looking to computational models of gaze may provide some clarity. Two parallel reading models, SWIFT (Engbert, Nuthmann, Richter, & Kliegl, 2005; Kliegl et al., 2006) and context-sensitive Mr. Chips (Bicknell & Levy, 2010), predict such benefits. SWIFT is a spatially distributed processing model of eye movement that takes into account properties of the oculomotor system and the process of word recognition. Mr. Chips (Legge, Klitz, & Tjan, 1997) is another model of eye movement, an ideal observer model that yields optimal performance within

the constraints of the human visual system. Bicknell and Levy (2010) recently modified the Mr. Chips model to enable the inclusion of contextual information in predicting eye motion. By using a bigram language model, that is a model that includes the contextual effects of the previous word, Mr. Chips is able to more accurately predict saccade size than a model without contextual information. Any speed-up in the reading time of the first word of a high frequency trigram would be explained by the parafoveal preview of the entire trigram from the first word (Rayner, Inhoff, Morrison, Slowiaczek, & Bertera, 1981). The further implication of this result is that there were parafoveal-on-foveal effects[3] for not only the second word but also the third word (Kliegl, Risse, & Laubrock, 2007). Angele and Rayner (2011) have questioned the existence of these parafoveal-on-foveal effects, and the EZ-reader model of eye movement does not predict such effects (Rayner, 2009). The evidence presented here is compatible with SWIFT and the modified Mr. Chips, that is to say, models that take statistical, probabilistic information into account at multiple grain sizes.

3.6.2 Contribution of PMI

PMI, my measure of phrasal cohesion, entered into many of the models I presented here. It interacted with *n*-gram frequency in predicting total reading time and with b1IC and w1IC in predicting regressive saccades. It interacted with cc2 in predicting the number of fixations. It interacted with cc1 in my model for SG1, which is an early stage of reading. PMI interacted with *n*-gram frequency and cc2 in my model predicting SG2. If there is a general pattern to be found in the contribution of PMI, it is that, like frequency, it is highly interactive. The data show how PMI modulates or is modulated by other inputs. For both total reading time and SG2, the modulation took place at the extremes of the range of PMI. The reasons I can propose for this pattern are speculative, but they have to do with the nature of semantic coherence.

By looking at items with extreme PMI values we can begin to see what

[3]These are effects due to the parafoveal information from fixation on a word being combined with the foveal information from the next fixation

could be going on. The two trigrams with the highest PMI scores are *covetous griping miser* ($sPMI = 3.91$) and *obstructive pulmonary disease* ($sPMI = 3.11$) while the two trigrams with the lowest PMI are *for provided the* ($sPMI = -2.08$) and *reserved of the* ($sPMI = -2.1$). The trigram with the median PMI in the stimulus set was *a glorious new* ($sPMI = -0.14$). When reading these trigrams the subjective feeling of cohesion increasing with PMI is clear. Incoherent *n*-grams were harder to read in most cases, but when they were uninformative, coherent *n*-grams were also hard to read (see Figure 3.2B). A similar pattern arose in the regressive saccade model (Figure 3.3B). PMI interacted with the lexical class of the second word in all relevant models for number of fixations and SG2 — in these interactions, the direction of the PMI effect flipped depending the class of the second word. I discuss this pattern further below when I discuss cc2.

3.6.3 Contribution of TIC

The holistic information content measure, TIC, entered into the best models for predicting total reading time and regressive saccades. In both cases there was an interaction between coherence (PMI) and informativeness (TIC). In both total duration and regressive saccade analyses, the combination of both low sPMI and an sTIC between -2 and 2 was followed by slower reading times and a greater chance of regressive saccades. An example of this kind of stimulus can be seen in the trigram *in got the* ($sPMI = -1.92$, $sTIC = -1.12$). When both PMI and TIC were larger, reading was more fluent, even when all other aspects of the *n*-gram were taken into account. An example of one these coherent, highly information rich trigrams is *freight train rumbles* ($sPMI = 2.09$, $sTIC = 1.86$). When length, component *n*-gram frequencies and all the other variables were taken into account, there was a reading advantage (less regressive saccades, faster reading) for *n*-grams like *freight train rumbles*. TIC did not contribute to predicting the total number of fixations, SG1, but b2IC did contribute to predicting SG2. This is another example of how the informativeness of the second bigram is having an impact on the reading of the first bigram.

3.6.4 Contribution of cc2

The class of the second word in the *n*-gram was a predictor that entered into many models. Closed class words in that position increased the total reading time, SG2, and the total number of fixations [4]. These findings suggest that words I marked as closed class words (shown in Appendix 3.9) added a cognitive load to the task when they were in the middle of trigram. Since the effect of cc2 is seen in models that have already taken into account the impact of phrasal completeness, the effect of cc2 must be independent of completeness. Why else would closed class words in the second position have such robust and widespread effects on trigram reading? I speculate that it may be because closed class words rarely occur at the end of a phrase and therefore prime the language system to expect more words ahead with more information. When there are two more words (cc1), this expectation is satisfied in a situation where there are no prior words that might be re-assessed. When there are no more words (cc3), the lack of further input may stop the system from making predictions altogether. Only when a single word follows (cc2) does the system find the expectation of more information unsatisfied while also having the possibility of a reinterpretation of a previously seen word.

Our results were similar in many ways to those of Tremblay and Tucker (2011). Their measure of onset latency and my measure of total duration of reading both had simple effects of length and *n*-gram completeness and all of these effects were in the same direction. Tremblay and Tucker (2011) found interactions between the frequency of the second word in their quadragram and information content and mutual information, which I did not find in the data. They attribute these effects to their experimental paradigm, in which their fixation marker appeared near the second word of each quadragram, allowing the frequency of the second word to influence the reading of the whole *n*-gram. As Baayen et al. (2010) point out, these empirical interactions between the whole and the parts imply interactive, dynamic processing of information in the brain. Sequential, modular models will not predict this amount of interaction.

[4] It is likely that it did not enter into the model for SG1 due to the fact that the second word had not been seen yet.

I too found interactions, both non-linear and linear, in all of the analyses. My replication of these interactive effects supports the idea that probability and information content are concurrently impacting the reading of word groups.

Are these results generalizable to reading more than trigrams? The ecological validity of reading trigrams out of context is debatable, but so is the ecological validity of reading single words. The task that I asked the participants to execute is uncommon, but I believe that it is still ecologically valid. In everyday, real world situations, such as jumping to the first line of a new page in a book, we often encounter groups of words that are mid-sentence. Any of the trigrams that I used as stimuli could be at the beginning of the line. In these situations we are still able to continue reading – our reading systems are able to make do with this truncated input. Furthermore, the participants were able to consistently extract the meaning from the trigrams in the sentence production trials and create valid sentences, and I feel there is not reason why they could not have done so for every item. I take their success at completing the sentence production task as evidence that they were processing the meaning of the trigrams, as fragmented as they were. This is why my results speak to reading in general, not merely to reading bare trigrams. The advantage obtained from avoiding embedding the trigrams in longer sentences is the elimination of the longitudinal effects of previous context. By isolating trigram, the processes taking place during the reading of short n-grams can be teased apart.

There is a connection between this evidence and evidence from lexical processing research. This body of work suggests that information accumulates as we listen to speech or read (Rayner & Pollatsek, 1989; Elman, 1990, 2011). Some recent models of reading proposed by Norris and Kinoshita (2008), Dilkina et al. (2010b) and Baayen et al. (2011) look at this process of accumulation as a way of explaining experimental evidence from reading. One of these models, the Naive Discriminative Reader (NDR, Baayen et al., 2011) is particularly interesting because it does not contain representations for word forms or n-gram forms, but rather shows the emergence of morphological and lexical effects using nothing but sub-lexical probabilistic information. This model has

been applied to modeling the frequency effects when reading n-grams. Baayen and Hendrix (2011) used the NDR model to predict reading times for the stimuli used by Arnon and Snider (2010). The NDR model predicted the reading time from the model's knowledge of the statistical properties of patterns of letters and letter bigrams in the input. The data I have presented here would be an interesting challenge for these types of models. In particular, if the interactions between frequency, PMI and TIC emerge spontaneously from the sub-lexical patterns in each speaker's language experience, it would be a very persuasive argument for the relevance of non-lexical models of language.

3.7 Final Thoughts

In the introduction to this chapter, I explained my theoretical motivation for studying eye movements while reading n-grams: to further explore the possibility that the reading system is rationally using all the information available to read words in context more rapidly. The complex interactions between the many probabilistic measures included in the models have explained otherwise unexplainable variation in reading performance. This is evidence for the position that the reading system is using n-gram information in these micro-contexts to better process words. Any theoretical framework that chose to not include this information in a reading model would be sub-optimal by definition.

The interactivity between the many sources of information is a telltale sign of a dynamic system (Kuperman et al., 2008). Models of language processing that don't allow for the free flow of information throughout the language system (from ocular control to semantic integration) do not make sense in the face of my evidence. The processing of the first word in the trigram was influenced by the probabilistic relationships between all the words in trigram, a truly fascinating result. They eye movements of our readers were optimized by the effects of linguistic usage and experience. This implies that theories of language performance will need to take into account the contexts that a person is exposed to, as well as the local context. The way we read *emerges* from the interactive, simultaneous combination of many complex inputs.

The breadth of the stimuli set used in this study has allowed me to delve into the complexities of reading isolated trigrams. I found that total reading time was predicted by a combination of many factors with unique contributions made by the interactions between *n*-gram frequency, TIC and PMI. The size and scope of my trigram sample and the sensitivity of my analytical methods have provided the first detailed perspective on how we read three-word groups. All three types of probabilistic information were involved in predicting eye movements, all helping to predict the efficiency of reading. My results support the position that this sensitivity stems from the probabilistic nature of language and the probabilistic processing that it enables. It also shows that despite the presentation of the trigram outside of its normal linguistic context, the holistic effects are strong.

3.8 Appendix: Variable loadings for principal components PC1 to PC5

	PC1	PC2	PC3	PC4	PC5
w1f	0.02	-0.35	**-0.49**	0.03	-0.05
w2f	-0.06	**-0.44**	0.39	-0.05	-0.01
w3f	**-0.52**	0.17	0.01	0.13	-0.15
b1f	0.06	**-0.5**	-0.03	-0.05	**-0.5**
b2f	-0.35	-0.14	0.34	0.36	-0.13
b3f	-0.34	-0.08	-0.35	0.39	0.14
xfq1	0.05	-0.28	**-0.48**	-0.01	-0.18
xfq2	-0.05	-0.4	0.35	-0.16	-0.01
xfq3	**-0.48**	0.17	-0.04	-0.04	**-0.41**
length	0.32	0.34	0.05	-0.01	**-0.69**
mncmplt	0.38	-0.04	0.12	***0.81***	-0.05

Table 3.17: Loading of the first 5 principal components in the PCA solution for frequency, length and completeness. Correlations over 0.4 are shown in boldface. Correlations over 0.7 are shown in italics.

3.9 Appendix: Closed class word list

a	about	after	against	all	alongside	an	and	any	are
around	as	aside	at	atop	away	back	be	because	been
being	between	but	by	can	did	do	even	ever	few
for	forward	from	get	got	has	have	he	here	herself
his	if	in	into	is	it	just	may	me	more
much	no	not	of	off	on	only	or	other	our
out	per	please	rather	same	shall	should	so	something	than
thanks	that	the	their	them	they	this	those	throughout	thus
to	too	under	upon	us	was	we	well	were	what
when	where	whether	which	who	whose	why	will	with	without
would	you	your							

Table 3.18: The list of words used to classify closed class words in our stimulus list.

3.10 Appendix: Intercorrelations of predictors before and after PCA.

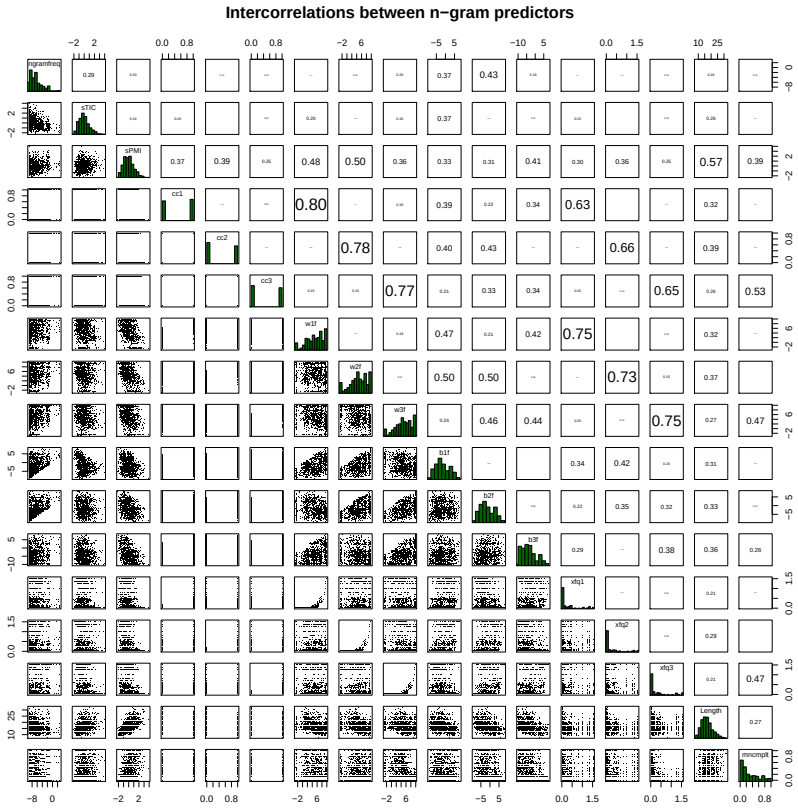

Figure 3.7: Matrix of correlations for all predictors before orthogonalization. The lower triangle contains scatterplots for all the predictor relationships. The stimulus sampling technique used enabled us to have the broad coverage seen in the scatterplots for all the corpus frequency measure (n-gram frequency, bigram frequencies and word frequencies). The upper triangle contains pairwise Pearson correlations for all the predictors, with the size of the font used showing the size of the correlation. The diagonal contains histograms for each predictor.

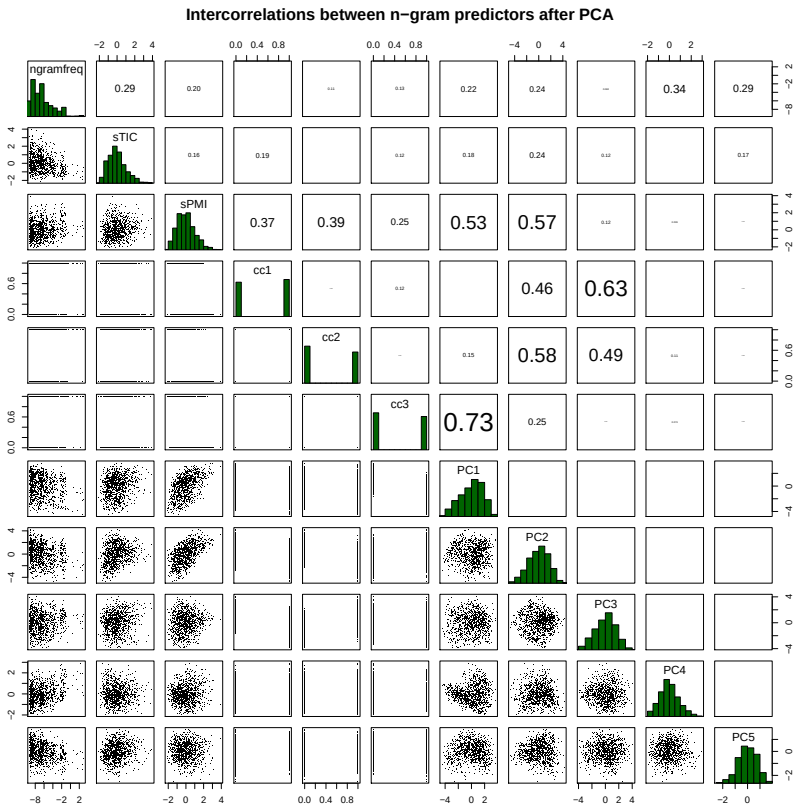

Figure 3.8: Matrix of correlations for new set of predictors after orthogonalization.

Chapter 4

N-gram probability effects in a cloze task.

We now know more about our memory for *n*-grams and how we read and comprehend *n*-grams. This chapter will delve into how we produce words in context, how we create *n*-grams. The paradigm I will used is the fill-in-the-blank task, or *cloze task*, is common in educational materials and psychological tests. Until recently, addressing the question of how to predict which words people will choose in a cloze task, that had not had its cloze probability measured previously, was impractical. In this paper I will harness the richness of large-scale corpora to look at the influence of lexical micro-context on word choice in the cloze task. In two experiments I asked young adults to complete short phrases called *n*-grams. The probabilistic properties of the *n*-grams were predictive of the frequency with which each word was produced. Furthermore, in the second experiment, the order in which the words were generated was predicted by the conditional probability of the word occurring in that micro-context in a corpus. These results suggest that the cloze task is essentially a memory task and the choice of certain words over others is driven by the probability of the word given the micro-context.

4.1 Introduction

When we read or hear language, we are concurrently processing what is perceived and predicting what will happen next. Landmark studies of anticipation in language perception have allowed us to understand how we react to expected versus unexpected words (Kutas & Hillyard, 1984; DeLong, Urbach, & Kutas, 2005). Bar (2007) goes as far as to say that "memory-based predictions/association-based predictions" are one of the unifying principles of the brain (p. 280). In this paper I will look beyond the influence of concurrent

prediction in linguistic comprehension and investigate the use of anticipation in linguistic production. In particular, I will expose the participants to incomplete fragments of language and invite them to complete them, a type of cloze task (Taylor, 1953). I hypothesize that contextual, probabilistic prediction will play a part in completing these phrases because there is strong evidence that we anticipate upcoming words during sentence processing (Kamide, 2008). A similar process may be taking place during the completion of a cloze task. I propose that the word chosen in a cloze task is the first one provided by the language prediction system, working within the constraints of the context.

This raises a larger theoretical question: what is a language prediction system and what does it base its predictions on? The fundamental theoretical shift that is happening in psycholinguistics is a shift towards theories of language that posit the existence of very simple statistical learning processes at the heart of language acquisition and processing. The theories proposed by Frank and Bod (2011), Elman (2011) and Baayen et al. (2011) rest on the idea that complex language behaviour emerges from a very simple process, and the statistical structure of the input is sufficient to explain our linguistic abilities. Looking at our language prediction system from this vantage point, all that is needed to predict words in a language stream is exposure to the statistical structure of a language. Using that probabilistic information, a person can begin to predict and anticipate words from context. However, there is an alternative possibility, one that cannot be dismissed outright. That position would be one in which prediction from context is not the dominant influence — sources of information other than the micro-context might be the main factor in word choice in an *n*-gram cloze task. This is the theoretical question I hope to address in this chapter: how does contextual probabilistic information take part in the process of choosing a word in a cloze task?

A first step in getting closer to answering this question is to define the term *cloze task*, first proposed by Taylor (1953). The word *cloze* was chosen because it harkened back to the gestalt principle of *closure* in the visual sense — if people are exposed to a partial drawing or photograph they perceive a whole by filling in the missing information (W. Ellis, 1999). The linguistic

cloze task was originally devised as a method to measure the readability of texts. A certain number of words in a text were replaced with blanks, and naive readers were asked to fill in the blanks with the words that made sense in the contexts given. If the accuracy on the cloze task was high, the text was considered very readable, implying that texts that were more predictable were easier to read.

Before the coining of the term *cloze task*, in 1951, Claude Shannon calculated the entropy of English using a letter cloze task and a small corpus (Shannon, 1951). He was not addressing psycholinguistic questions, but his influential paper spurred psychologists to look at information, entropy and redundancy in language. One of the first studies to use the cloze task as a window into the psychology of language was carried out by Fillenbaum, Jones, and Rapoport (1963). They deleted words from transcripts of spoken English at 5 different rates (every 2nd, 3rd, 4th 5th or 6th word) and measured subjects' accuracy for all of these deletion rates. They also counted how often the word was replaced with a word different from the original word but of the same lexical class. They found that performance was better than chance at all the different deletion rates, showing the powerful impact of context on performance. Despite this, there was a great deal of item variability in the accuracy data, with some contexts enabling greater item accuracy than others. The key discovery in this study was that context strongly constrains what subjects will produce when asked to fill in the blanks, paving the way for computational studies of cloze task completion.

Once the computational resources became available to build digital corpora, find the word frequencies in these corpora (Francis & Kucera, 1982), and calculate probabilistic measures from these corpora, researchers began to look at the cloze task as an experimental behaviour that could be predicted. Up to this point, human rating norms were the only way to calculate the cloze probability. The second wave of studies attempted to understand the results of cloze experiments based on the distributional properties of language as observed in corpora.

Finn (1977) was the first to use the ideas of entropy and information from

Shannon (1948, 1951) to analyze a cloze task. He used data from a cloze experiment done by Bormuth (1966) to calculate the information, taking into account both the orthographic frequency of the words and the number of completions given by subjects. He found that if the words replaced by blanks had high amounts of information (and were low frequency) they were less likely to be correctly chosen in the cloze task. Low information/high frequency words were more likely to match the original word. The reason for this was that the constraints on the high frequency words (which were mostly closed class words) by the context of content words (open class words) was strong. In contrast, the constraints placed by open class words (such as *cat*) on the preceding closed class word (*a cat?*, *the cat?*, *some cat?*) were found to be weak. This attempt to understand the relevant sources of variation in the cloze task was valiant, but was limited by the quality of the frequency info that was available at the time.

Beattie and Butterworth (1979) did ground-breaking work on the interactions between lexical frequency and cloze probability. They looked at pauses of more than 200 ms in spontaneous speech and noted the frequency of the words after the pauses. Judges were then asked to fill in a transcription of the speech data where the words after the pauses had been replaced by blanks. Beattie and Butterworth found that the corpus frequency of the missing word was correlated with the cloze probability given by the number of judges choosing that word. As in all the other research on the cloze task, the cloze probability of a word in a certain context was defined as the probability that a panel of judges will choose that word when asked to fill in the blanks. If 10 out of 25 judges pick a certain word in that context, the cloze probability is said to be 0.4. This was the only way to analyze data from experiments that used the cloze task, by comparing the new results to cloze norms. Two problems exist with this definition of cloze probability: 1) The inconsistency of human judges can render cloze norms too noisy to be useful, and 2) When there are a large number of meaningful completions possible, a panel of judges, each providing its solution to the cloze riddle, cannot provide enough completions to give probabilities for all possible meaningful completions. The theoretical signifi-

cance of this correlation between orthographic frequency and cloze probability was not appreciated at the time, but this result was evidence for frequency effects in the process of generating a response in a cloze task.

McKenna (1986) proposed that the responses in a cloze task could be predicted based on the associative strength of the words in their semantic categories. He built a computational simulation that searched through lists of word categories from the norms given by Battig and Montague (1969). His explanation of the process of completing a cloze was that "At the semantic level, a single schema associated with a key constraint is used as the basis of a memory search. Sub-schemata in the form of individual words are examined in the order of associative strength until a word that meets additional constraints (if there are any) imposed by context is encountered." (McKenna, 1986, p.493) This model is far removed from my theoretical stance, but the core concepts of memory retrieval within constraints remains important.

There has also been research on a verbal production variant of the cloze task that directly addresses the interplay between frequency and contextual constraint. Griffin and Bock (1998) asked participants to complete a sentence with a word that was cued with a drawing. They manipulated the spoken frequency of the word cued by the picture and the amount of constraint created by the context. They found that when the context was highly constraining the effect of frequency on the word chosen was diminished. When the context was less constraining or even incongruous, high frequency words were chosen over low frequency words. This was evidence that both verbal and written cloze tasks are influence by orthographic frequency of the response word.

Smith and Levy (2011) asked subjects to complete sentence-initial 4-grams and compared their responses with the most frequent continuations form the Google Web1T corpus. They found that subjects response were sensitive to corpus probabilities, and the responses from the subjects were more variable than the corpus. They then asked a different group of subjects to read some of these 5-grams that they found in the first experiment, and calculated reading times for these critical 5th words. They found that the a model without control covariates found an effect for cloze probability, but once the covariates (lexical

frequency, concreteness, contextual diversity and length) were added, these effects were no longer significant. The question of what drives word choice in a cloze task was left unresolved, but there was a tantalizing possibility proffered: that real-word experience and prediction from experience is behind it.

Cloze norms

The value of cloze probabilities to experiment designers in many fields is high. There have been several popular lists that have been used to help design psycholinguistic experiments where the predictability of a word in context is critical such as event-related potential (ERP) experiments that measure expectancy violations (Kutas & Hillyard, 1984). Bloom and Fischler (1980) produced one of the first set of norms, a set of 329 sentences. Recently a larger set of 498 norms has been released by Block and Baldwin (2010), and the N400 effect was validated for these contexts. The goal of these norms is to find the most highly semantically constrained contexts possible, such as the sentence *She could tell he was mad by the tone of his* _____, which their subject completed with *voice* 99% of the time. 400 of the 498 sentence have a top completion that is dominant (defined as a cloze probability between .67 to .99). They achieved their goal of finding and norming many highly constrained sentences, but all of these norming studies do not delve into the sources of constraint, nor do they try to understand the source of the variability in their data. The issue with the analysis of cloze norms is the generalizability of the data. To know the cloze probability of an arbitrary piece of language it would be necessary to collect more human judgements, impractical for large amounts of text.

Are sentences and paragraphs the only type of stimuli that make sense in a cloze task? In many types of reading activities we are not exposed to a sentence-worth's of context. For example, reading a narrow column of text will inevitable cause a group of three or four words to be cut off, and the reader will have to look down to the next line to continue the sentence. These short groups of words are what I will call *n*-grams. There has been a recent growth in the number of studies investigating the processing of *n*-grams. Arnon and

Snider (2010) showed that more frequent *n*-grams were read faster than less frequent *n*-grams. Tremblay and Tucker (2011) measured how long it took subjects to read an *n*-gram and also how long it took them to produce the *n*-gram. Probabilistic predictors were able to explain much of the variability in reading the *n*-grams aloud (see Shaoul & Westbury, 2011 for a comprehensive review).

In this study I will look at how *n*-gram statistics from a corpus can predict our choices when we complete a linguistic fragment, an *n*-gram. The *n*-grams that I will use are 3 or 4 words long, much shorter than sentences used in most previous research on the cloze task. The shorter length reduces the amount of context and increases the number of possible completions of an *n*-gram cloze task. These differences force us to look at other types of psycholinguistic paradigms for theoretically relevant work.

One type of research that may be germane is work on free association (Nelson, McEvoy, & Dennis, 2000). The classic free association task is to provide a cue (such as *bread*) to many subjects and count the frequency of the various responses (such as *butter*). Nelson et al. (2000) characterize free association as a memory task, and point to evidence from cued recall experiments (Nelson, McKinney, Gee, & Janczura, 1998) and false memory experiments (McEvoy, Nelson, & Komatsu, 1999) that support the predictive power of associative strength in memory tasks. They conclude that the probability of a response being given is a manifestation of its associative strength in memory, noting that the "strength of a response reflects the number of its instances in memory, with stronger associations reflecting larger numbers of instances" (Nelson et al., 2000, p.896).

How similar is the free word association task to the *n*-gram cloze task? The context of a single word is less than that of a 3-gram, but the process of producing "the first word that comes to mind" may be similar for both free word-word association and free *n*-gram-word association. When presented with a short *n*-gram, such as *third most popular*, it is conceivable that a pool of associates emerges, and that the strongest associate is chosen first. One question I hope to address in this chapter is what information is producing

this emergence of candidates and what factors determine the order that they are selected from this pool.

Another way to view the cloze task is as a strategic test of memory retrieval. Pickering and Garrod (2007) argue that language comprehension involves making simultaneous predictions at different linguistic levels and that these predictions are generated by the language production system. They report increased muscle activity in the lips and tongue when listening to speech but not when listening to non-speech noise as evidence for an automatic forward-modeling system. In their system comprehension and production are tightly coupled and the motor production system facilitates language comprehension just as the comprehension system facilitates production. Pickering and Garrod (2007) is persuasive when he argues that language knowledge, stored in memory, exerts its contribution to behaviour by way of predictions. The advantages of a constant simulation or *emulation* of the external is becoming a foundational idea in psycholinguistics (Willems & Hagoort, 2007). In looking at how we process and complete *n*-grams, an emulation framework may help us explain how completions are chosen. To understand how written production systems choose a word, I will look at what kind of information could be used by a hypothetical emulator to predict an upcoming word in a stream of words.

The context found in an *n*-gram puts certain constraints on what words can fill an empty slot. Semantic and syntactic constraints are the most studied constraints. The constraint that I hope to add to this list is the constraint of memory: if there is a implicit or explicit memory of seeing or writing an *n*-gram, that *n*-gram should be accessible during the completion of a cloze task. Conversely if there are no memory traces for an *n*-gram, the likelihood of predicting a completion using that *n*-gram is much lower. I will assume in these experiments that the probabilistic measures of the *n*-grams are correlated with the participants' language experience, and the effects of the constraints of experience will be seen in the responses the participants give.

New tools have become available in the current era of psycholinguists that allow for new approaches to the question of constrained language production. In particular, corpus data-driven research has begun to let us investigate the

probabilistic nature of language in new ways. One asset is the enormous corpora of electronic texts, and the computational resources to calculate lexical statistics across these corpora. A case in point is the Google Web1T corpus (Brants & Franz, 2006) that I will be using exclusively as my source of probabilistic information about language in this paper. It is a corpus made up of one trillion words of English web page text. Using the immense computing infrastructure at Google, the authors were able to count all the occurrences of all the word groups or *n*-grams from two to five words long (but only those that occurred more than 40 times per trillion). These frequencies are a rich source of information about the word grouping patterns of English and give us an unprecedented ability to estimate the probability of word co-occurrence on the web. There are, undoubtedly, differences in the language experience of individuals that are not captured by the broad coverage of the web corpus, but the immense number of *n*-grams included in the Web1T corpus make it invaluable to those seeking to understand the influence of *n*-gram probability on lexical processing. By using the information in this large corpus I hope to better understand performance on cloze tasks.

4.2 Statistical Considerations

It has been duly noted that there are individual differences in cloze tasks and verbal fluency tasks, even among non-pathological populations. There will also be individual differences seen in the data collected in my experiments. There is a need to account for the theoretically uninteresting variation caused by individual differences between subjects and between items. Baayen (2008) recommends using mixed effects models with crossed random effects to account for the variability in these types of experiments. I will use various statistical tools in this paper, and I will use mixed models whenever there is theoretically irrelevant within-subject or within-item variability.

In all of my analyses I will use model selection to perform statistical inference. Models that had the best balance between fit and complexity will the ones I report, and this balance is always measured by using the Log Likelihood

Ratio Test (LLRT) or the Akaike Information Criterion (AIC), which penalize overly complex models.

The multi-colinnearity of the predictors in all of my models were checked using the condition number measure, κ. When necessary, I applied principle component analysis (PCA) to produce a set of orthogonal predictors that explained the same variability as the original predictors.

Finally, I performed model criticism on all of the models presented in the paper: I located those observations that led to residuals greater than 2.5 standard deviations away from the mean of the residuals and temporarily dropped them from the data. I then reanalyzed the data using this subset and checked to see if there were any large fluctuations in the size or direction of the effects. I only present models here that were stable during model criticism.

4.3 Experiment 1

The original cloze task was characterized by a long passage of around 200 to 300 words with a certain percentage of the words removed (Taylor, 1953). Later psycholinguistic studies shortened the stimuli to single sentences, sometimes forcing the blank to always be in the same position (sentence-final for example) (Schwanenflugel & LaCount, 1988). As the amount of context shrinks, the task itself changes. One question I will address in this study is: What happens when the context is reduced almost to the minimum, to two or three words? The poverty of context should reduce the constraints on the number of meaningful ways to complete the fragment and allow a greater variety of completions than for longer sentences or passages. Most of the n-grams I will use in my studies are not complete constituents and some will remain fragmentary even after filling in the blank (e.g.*nothing to do with the*).

If the statistical properties of the English language, in particular the frequencies with which words occur together, are reflected in the memories of the participants, the frequency of n-grams captured in the Web1T corpus should help predict the likelihood that the participants will choose a certain completion in this cloze task.

4.3.1 Participants

2110 undergraduate students at the University of Alberta participated in this study as part of a larger group of web-based surveys. 74% of the participants indicated that English was their first language. All participants had university-level English abilities and so all 2110 were included in the dataset.

4.3.2 Materials

Two sets of n-grams were created. This first set was the letter-completion task stimuli. 22 n-grams from the Web1T corpus, 12 trigrams and 10 quadragrams, were chosen at random, and in each of these n-grams a critical letter was removed. The n-grams were chosen in such way that the number of possible completions attested to in the corpus for the critical letter varied in the corpus from 2 to 11, $\overline{\mu} = 4.9$, $\overline{\sigma} = 2.7$.

In an identical way I chose 23 5-grams from the same corpus, deleting one word from each n-gram in such a way as to create variety in the number of possible completions. These varied from 3 to 5792, $\overline{\mu} = 384$, $\overline{\sigma} = 1194$.

4.3.3 Procedure

The survey was administered using custom web-based software. All participants completed the web survey at a time and location of their choosing. In the pre-survey instructions they were requested to find a quiet location where they would not be disturbed before starting to complete the survey.

Participants were asked to take note of the first phrase that came to mind when reading each of the incomplete phrases. They were instructed to fill in the blanks with the missing letter (letter completion) or word (word completion) for that phrase. They were also requested to take care and make sure that the completed phrase would make sense. Almost all of the participants fully completed the survey: out of 92,840 responses, only 1247 (1.3%) were left blank. For the word completion survey the fields did not allow more than 12 characters to be entered, making the maximum word length for those responses 12 letters.

4.3.4 Results

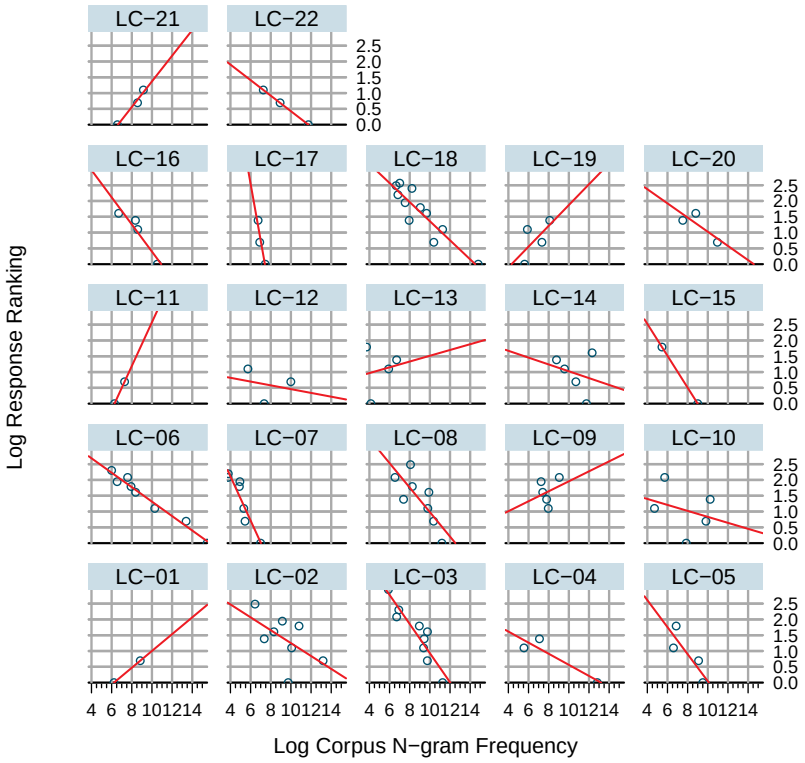

Figure 4.1: Item level scatterplots for all items in the letter completion task with best fitting linear regression line. To make the graphs easier to read, the log-transformed rank is shown on the y-axis. The outcome of the analyses are identical when the untransformed rank is used.

To reduce the impact from idiosyncratic responses in the word completion task, I removed all responses that were given by less than three participants for each item. This entailed removing 8445 responses (18.2% of the data) from the word completion dataset, and no responses from the letter completion dataset. I also dropped all responses which created *n*-grams that had a frequency of less than 40 times per trillion on the Web1T corpus (the lower bound of the frequency range in that corpus). By doing this I removed most of the non-sensical responses that were given more than 2 times, such as *I have get the*. In letter completion data, I dropped 8,878 responses (19% of the data), and

for the word completion dataset, I dropped 11,578 more responses (25% of the data). After removing all the idiosyncratic, nonsense responses I was left with 36,991 responses for the letter completion task and 33,219 responses for the word completion task. This was the final data set that I used in the following analysis.

First I will look at the results from the letter completion task. The frequency of each response was tabulated, and those frequencies were then assigned a rank, the dependent measure of interest. In Figure 4.1, the relationship between rank and corpus frequency is shown for each item. Out of 22 items, 15 had negative relationships, but the number of response types for some items was very small. For inferential purposes all the items were pooled and a linear mixed effects model was used to analyze the effect of frequency when the random effect of item was included in the model. Compared to a model with no fixed effects and the random effect of item, a second model with the fixed effect of corpus frequency and the random effect of item was a better model ($\chi^2(1) = 30.6$, $p =$ 3.2e-08). In completing these words with a letter, the frequency of the word completed is of concern. For examples in the case of the stimulus *the _at is*, the frequency of the *n*-gram*the cat is* in the corpus is higher than the frequency of *the fat is*, but the frequency of the word *cat* is lower than the frequency of the word *fat*. To investigate the additional impact of word frequency, I created a third model that included both *n*-gram frequency and word frequency (and the random effect of item). It was no better than the simpler model that did not include word frequency ($\chi^2(1) =$ 1.2 , $p = 0.27$). Word frequency was not predictive of the rank of the cloze completion.

As with the data from the letter completion task, I calculated the frequencies of each of the responses in the word completion task to use as the outcome variable. After a visual inspection of the relationship between response rank and corpus frequency, shown in Figure 4.2, I noted that 21 of the 22 items had negative slopes. To confirm this negative relationship I entered all of this data into a linear mixed effects model, first with no fixed effects and a random effect

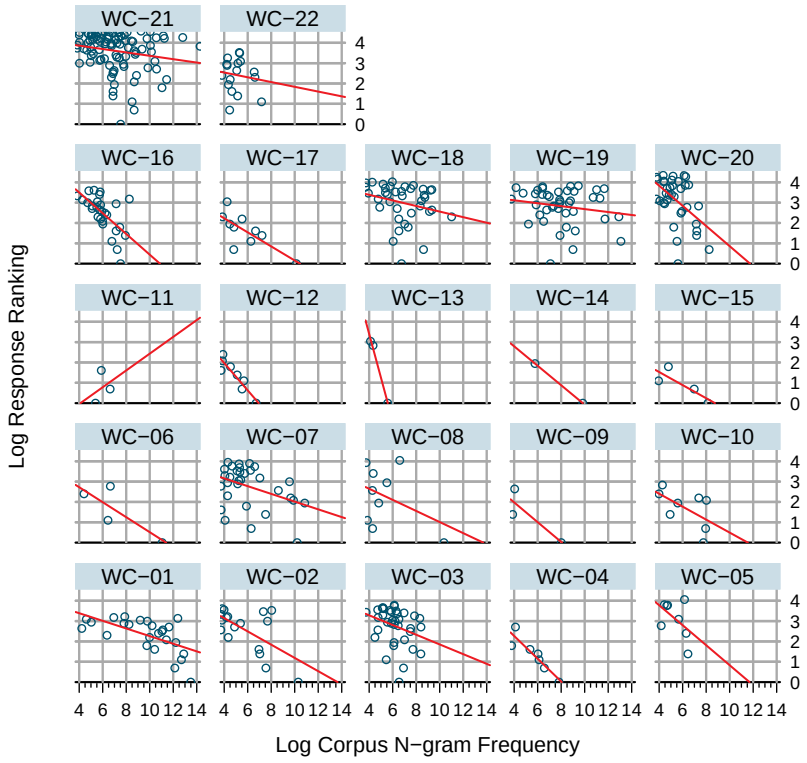

Figure 4.2: Item level scatterplots for all items in the word completion task with best fitting linear regression line. To make the graphs easier to read, the log-transformed rank is shown on the y-axis. The outcome of the analyses are identical when the untransformed rank is used.

of item. I then created a model with the fixed effect of corpus frequency and the random effect of item. It was a better model than the first model. ($\chi^2(1) = 22.3$, $p = 2.3\text{e-}06$). Again, a third model was created with the addition of the fixed effect of word frequency, and it was no fitter than the model with only *n*-gram frequency ($\chi^2(1) = 0.53$, $p = 0.47$).

4.3.5 Discussion

The *n*-gram frequency, but not the word frequency, aided in predicting how participants would complete *n*-grams in a close task. The raw frequency of a *n*-gram is merely a count of occurrences in a corpus. *N*-grams with a high corpus frequency were likely to have been seen more often by our participants, making these contexts more familiar, and this could explain my results, but I feel that there is more to this story.

McDonald and Shillcock (2001) and Baayen (2010a) have shown that pure repetition is not what gives high frequency words their advantage in lexical processing tasks. By necessity, orthographic frequency is correlated to many other probabilistic measures that are more psychologically relevant than exposure. Baayen argues strongly that context effects are more important than pure repetition effects. Also, McDonald and Shillcock (2001) found that *contextual distinctiveness* (CD) can subsume orthographic frequency in predicting behaviour. CD intimately linked to the micro-context, the *n*-gram, because it is the relative entropy between a word's micro-context and a distribution that ignores context. CD can be thought of as a word's informativeness about its contexts. Baayen (2010a) has looked at the contribution to predicting lexical decision RT of 17 lexical variables from many categories : frequency, genre distribution, CD, syntactic entropy, morphological entropy, and orthographic features. He found that once other predictors were entered into a model of RT, very little variability was left for orthographic frequency to explain. This suggests that frequency effects are epiphenomena. Further buttressing the argument for the frequency-effect-as-epiphenomenon position, a computational model, the Naive Discriminative Reader (NDR, Baayen et al., 2011) has found frequency effects in simulated lexical access without having any lexical rep-

resentations, but rater only sub-lexical, letter *n*-gram representations. These sub-lexical context effects that mimic frequency effects show how the illusion of orthographic frequency effects comes to be.

This view of frequency is relevant to this experiment because *n*-grams are almost always embedded a context, the sentence, and *n*-grams that are more frequent will get a boost from their contextual distinctiveness in much the same way that words do. McDonald and Shillcock (2001) and Baayen (2010a) restricted their discussion to frequency effect for single words, but the similarity of frequency effects in the processing of words and *n*-grams has been demonstrated in multiple experimental paradigms as described in Chapter 1. The *n*-gram frequency effect seen in my experiment was not simply a consequence of repetition-based learning but rather a consequence of reduced processing effort arising from the contextual distinctiveness of the *n*-grams.

If we think of frequency as an epiphenomenon that measures how we learned to link *n*-gram form to *n*-gram meaning, then it is actually *n*-gram learning that drove the choices in the cloze completion. The participants relied on their memory of learned mappings, and chose the first letter or word that they remembered. The Web1T corpus captured the combined linguistic experience of many people writing in many registers, and the response distribution reflected this diversity: less people associated with the lower frequency *n*-grams because their personal experience falls to the tails of the distribution of experience.

This experiment provided evidence of a memory-based production process, but the small number of items and the single response per item by each subject limited the number of questions I could ask of the data. To probe this *n*-gram completion process more deeply I followed it up with a second experiment.

4.4 Experiment 2

To address the limitations of a single completion per subject per item in Experiment 1 I used a different methodology in Experiment 2 to allow each subject to list multiple completions. I posited that if participants were given the opportunity to generate many completions as the could (up to 20), allowing us

to probe differences in the productivity of *n*-grams. I would now be able to analyze the order in which the completions were generated, giving me a window into the sources of relative accessibility of the responses. Finally, to simplify the experiment, the position of the blank in the stimuli was limited to the beginning of the stimulus (prepending) or the end of the stimulus (appending). By limiting the location of the responses to either the beginning or the end of a trigram I am able to calculate the conditional probability of each response in the Web1T corpus allowing me to investigate how the predictability of a word in context can influence behaviour in a cloze task.

A relevant body of work to Experiment 2 is the study of verbal fluency. In verbal fluency studies the subject is presented with a categorical cue and asked to generate as many members of that category as possible in a fixed amount of time, usually one minute (Ruff, Light, Parker, & Levin, 1997). The categories are often letter categories ("Words that begin with the letter F.") or semantic categories ("Animals"). The letter fluency task is in some ways similar to the *n*-gram completion task, where I asked the subjects to generate up to twenty members of a category without time limits. My categories are slightly different ("Words that can be joined to the 3-gram *chocolate chip cookie*") but the response space is essentially the same as that in the letter cloze task: a subset of all words in the language. Owing to these similarities, I will also attempt to understand my results in the context of research done on verbal fluency. Unfortunately most studies of verbal fluency depend on a manual coding of the responses, looking for clusters of similar responses (*fast, faster, fastest, fasting,...* (Troyer, 2000). I did not attempt these types of analyses, but I was able to measure the total number of responses produced, which is also one of the main outcomes of verbal fluency studies.

One of the few studies that used a task similar to ours was one by Owens, O'Boyle, McMahon, Ming, and Smith (1997). They were interested in speech recognition systems and had built a weighted average *n*-gram language model that predicted words based on their probability given the previous context. They used the 1 million word Brown corpus (Kučera & Francis, 1967) to train the model, then used the model to calculate the top 30 completions for 768

fragments. The same fragments were given to 8 human participants asking them to provide a ranked list of completions for fragments of text such as *the republicans must hold a _____under the county.* The results of the statistical model and the humans were compared with the intention of measuring the quality of the model. They found that their model was almost as good as the humans in picking the word that had been deleted from the passage (21% versus 26% correct) and was almost as good for getting the correct answer within the first three tries (72% versus 80% correct). From a psychological point of view it is fascinating that humans performing this task produce such similar responses to a statistical model of memory built on n-grams. My corpus is larger than the one used by Owens et al. by a factor of 10^6 and my methodology is different but my hypothesis is that my n-gram probabilities will have a similar predictive power for the participants' performance in a multiple-response cloze task.

Another study of interest by Crowe (1998) looked at the change in the responses to a verbal fluency task over time. Subjects were given one minute to complete letter and semantic fluency tasks, and the number of responses was counted over each 15 second interval.Crowe noted that the largest number were produced in the first 15 seconds and progressively smaller numbers of responses were produced for each of the following three 15-second periods. Also, the orthographic frequency of the words was higher for the first words produced, and lower for the later words. This task is similar to my task, and I will be able to analyze the order of production to see if the same pattern appears in my results.

Unsworth, Spillers, and Brewer (2010) conducted a verbal fluency study and found that working memory capacity (WMC) was the most effective predictor of individual differences in verbal fluency, with some additional contribution of vocabulary size. They hypothesize that WMC is involved in the maintenance of category cues and in the monitoring of the retrieval of responses from memory. According to Unsworth et al. and Rosen and Engle (1997), fluency arises from a combination of strategic (WMC) and associative (vocabulary) processes. In this study I did not collect any data that would

allow me to study individual differences in the participants. I used statistical methods to account for the random effects of participant, and individual differences in WMC is likely one of the sources of subject variability.

4.4.1 Participants

Using a custom web experiment management system, I recruited 864 undergraduate students from the University of Alberta. All were self-described native speakers of English.

4.4.2 Materials

I randomly sampled 240 trigrams from the Web1T corpus to cover a broad range of frequencies. As in Experiment 1, I sampled these trigrams at random without regard for their status as constituents. Most of the trigrams in the corpus are very low frequency, and so a minority of the stimuli may appear to be malformed (e.g. *in the to*). Rather than try to filter out these items using arbitrary criteria I left them in the stimulus set and gave them no special treatment, but the inclusion of these items may have been the reason for the large number of responses that were eliminated from the data as explained below.

4.4.3 Procedure

The instructions about how and where to complete the survey were identical to the procedure in Experiment 1. The participants were randomly assigned into one of 30 groups and each group was asked to complete a different survey. Each survey consisted of a set of 8 trigrams for a total of 240 trigrams. Participants were asked to type in a word either before or after the trigram, and were given twenty fields to use for each trigram. If all 864 participants had provided 20 completions for each of the 8 trigrams they saw, the maximum total number of observations would have come to 138,240. I did not receive this many responses, perhaps due to the decision not to set any restrictions on the minimum number of responses in the experiment. Participants were

allowed to submit surveys with between 0 and 20 responses per item and still receive credit for their participation. I obtained 77,621 data points, an overall response rate of 56%.

This survey was part of a larger package of surveys that took approximately 50 minutes to complete in total. The other surveys in the package did not contain any tasks that were similar to this task. This survey was always the last of the web surveys to be administered in the package.

The instructions at the beginning of the survey asked participants to fill in a blank with the first word that came to mind, and to avoid changing the order of their responses once they had typed in a word. They were instructed to only type in one response per line, either before or after the *n*-gram. It was explicitly noted that all completed phrases should make sense. They were also asked not to consult books, web pages or other resources when thinking about how to complete these *n*-grams. As in Experiment 1 the maximum word length was 12 letters.

There were some entries that were excluded because participants typed a word both before and after the *n*-gram. After eliminating 7,925 responses because of this type of double entry error, I was left with 69,696 observations.

4.4.4 Results

The first analysis I attempted on this data was similar to the analysis I did for Experiment 1: I tried to predict the frequency of responding, or cloze probability, of each type of response.

To filter out nonsense responses I removed data for responses that did not have a corresponding entry in the Web1T corpus. As has been noted by Hahn and Sivley (2011), there is an issue with the Google Web1T data from Brants and Franz (2006): due to technological constraints, it is very computationally expensive to collect frequencies for very rare *n*-grams. For this reason the corpus only contains data for *n*-grams that occurred 40 times per trillion or greater. Since my trigram stimuli were drawn at random from the Web1T corpus, and since most trigrams are rare, many low frequency trigrams were drawn (the full list of stimuli are given in Appendix 4.7). Finding a sensible

completion to a very low frequency proved to be extremely difficult for the participants. This may explain why most of the participants' 4-gram responses were in the range of 0 to 39 occurrences per trillion, leaving them out of the Web1T 4-gram data. Without a 4-gram frequency measure, I cannot include these responses in my model. A casual inspection of these removed responses showed them to be almost all nonsensical. After removing all responses without Web1T frequency data, the number of types present dropped from 47,438 types to 8,066 types translating to a drop from 69,696 responses to 18,356 responses. Of the 51,340 observations that were dropped, 33114 were singletons. Following Nelson, McEvoy, and Schreiber (1998), all idiosyncratic responses were dropped from the dataset. The number of observations per type was still quite small: the mean number of observations per type was 2.2 ($\overline{\sigma} = 2.6$, range = 2 to 29).

During this process of removing idiosyncratic responses the number of stimuli left in my data set dropped from 240 to 201, meaning that 39 items produced 4-gram responses that were *all* absent from the Web1T corpus (a total of 937 observations). A description of the 39 items that were dropped at this point is provided in Section 4.8.

The final data set was split into two sub-groups: responses that were prepended (7,461 observations of 3,248 types) or appended (10,895 observations of 4,818 types). These are the data sets that I will analyze.

I hypothesized that if many participants produced the same response for an item, that response may be preferred because it has been seen in that context before, and this contextual cue aided in retrieving a memory involving that context. Consequently I expected that the conditional probability of that response in the Web1T corpus should be predictive of the response rank. The conditional probabilities for the responses were calculated as:

$$P(\text{Response Word}|\text{Trigram}) = \frac{P(\text{Response Word} \bigcap \text{Trigram})}{P(\text{Trigram})} = \frac{P(\text{Quadragram})}{P(\text{Trigram})}$$

I entered this variable along with my other predictors and began fitting models to the data.

Each response has Web1T frequencies for all of the contained words, bi-

grams and trigrams, but since they are high inter-correlated, they cannot be entered into my model until the degree of multi-collinearity is reduced. I used PCA to extract the first four principle components from the set of the 12 frequency variables. The four new predictors are uncorrelated with each other, and combined they explain 85% of the variability in the original predictors. Using these four principle components I reduced the collinearity in my model's predictors to an acceptable level. The four principle components are referred to as PC1 to PC4.

I detected non-linear relationships in my models, and so I chose to use Generalized Additive Models (GAMs, Wood, 2006) to better understand the data. I included the random effect of item (as a smooth) in all of my analyses because I needed to take into account the relationship within the group of responses given for each stimulus.

Since the response frequency variable was not normally distributed, I chose to use Poisson distribution and the log link function instead of the Gaussian distribution. These counts are not completely independent because once a response is counted, it is no longer among the possible candidates of words to be counted, but this issue should not influence our results in an experiment of this size.

After following a stepwise forward model selection procedure for both prepended and appended datasets, the best models were almost identical. They both contained the following:

- A random effect for each trigram.
- The effect of the conditional probability of each response word given the trigram (taken from the Web1T corpus)
- PCs derived from the word, bigram and trigram frequencies within the quadragram (but not the quadragram's frequency itself)
- The response length (in letters)

The other predictors that were temporarily entered into models but did not improve the models were: the frequency of the whole response 4-gram, the

stimulus PMI and the response PMI. Frequency may have dropped out of the models because it is similar to conditional probability and did not absorb any new variability above and beyond conditional probability.

The estimates for the model's degrees of freedom in my final models are shown in Tables 4.1 and 4.2. The only linear relationship is the one for the conditional probability of the prepended responses, with an edf of 1. Since the relationships with PC1, PC2 and PC4 are not theoretically relevant, they will not be discussed. To understand the relationships in the GAMs it is vital to study the shapes of the smooths, shown in Figure 4.3. For both the prepended and appended responses, the conditional probability of the response in the corpus had a positive, mostly linear relationship with the participants' frequency of producing that response. To reiterate, the more predictive the context of a response, the more likely more people would choose the response (Figure 4.3 A and B).

The effect of length was negative for the appended responses (Figure 4.3 C). Longer responses were less frequent than shorter responses. For prepended responses, the relation was more complex (Figure 4.3 D). It is unclear why this smooth is this shape, but the only possibility is that there were some numbers of letters that were more common, and each of the peaks corresponds to one of these common word lengths.

	Estimated Df	Estimated Residual Df	F	$p_{bayseian}$
PC1	8.41	8.82	36.17	3.2e-05
PC3	8.00	8.66	57.74	2.6e-09
PC4	7.44	8.15	74.35	8e-13
Response Length (Letters)	6.82	7.80	260.43	7.3e-52
P(First word|Trigram)	3.31	4.16	540.86	1.4e-115
Random Effect of Item	192.14	212.06	1080.23	2.7e-116

Table 4.1: Model coefficients for smooths predicting response frequency in the best fitting GAM for the prepended responses.

	Estimated Df	Estimated Residual Df	F	$p_{bayseian}$
PC1	7.23	7.97	52.26	1.4e-08
Response Length (Letters)	6.82	7.87	95.35	3.2e-17
P(Last Word|Trigram)	4.89	6.00	1221.13	1.3e-260
Random Effect of Item	176.20	200.19	965.14	1.9e-100

Table 4.2: Model coefficients for smooths predicting response frequency in the best fitting GAM for the appended responses.

After taking into account the frequencies of the component *n*-grams and the length of the words, the conditional probability measured in the Web1T corpus was predictive of the frequency of the response, or the cloze probability. This evidence supports the hypothesis that *n*-gram memory drives word choice in my task.

4.4.5 Response Entropy and Family Size

The first question I asked of the data was: What influenced the participants' choices as to whether to place a response before or after the stimulus? To measure this I counted the total number of responses for each stimulus in each position and then divided them to create a position ratio: a ratio greater than 1 for items that had more prepended responses than appended responses and a ratio less than 1 for items that had more appended responses than prepended responses.

In all the following analyses, I created saturated linear models that included all of my predictors derived from the stimuli (whole *n*-gram frequency, component frequencies, pointwise mutual information and entropy), including all two-way interactions. Pointwise mutual information (PMI, Fano & Hawkins, 1961) is defined as the degree to which words in an *n*-gram occur together more frequently than would be expected by chance. For the trigram stimuli it

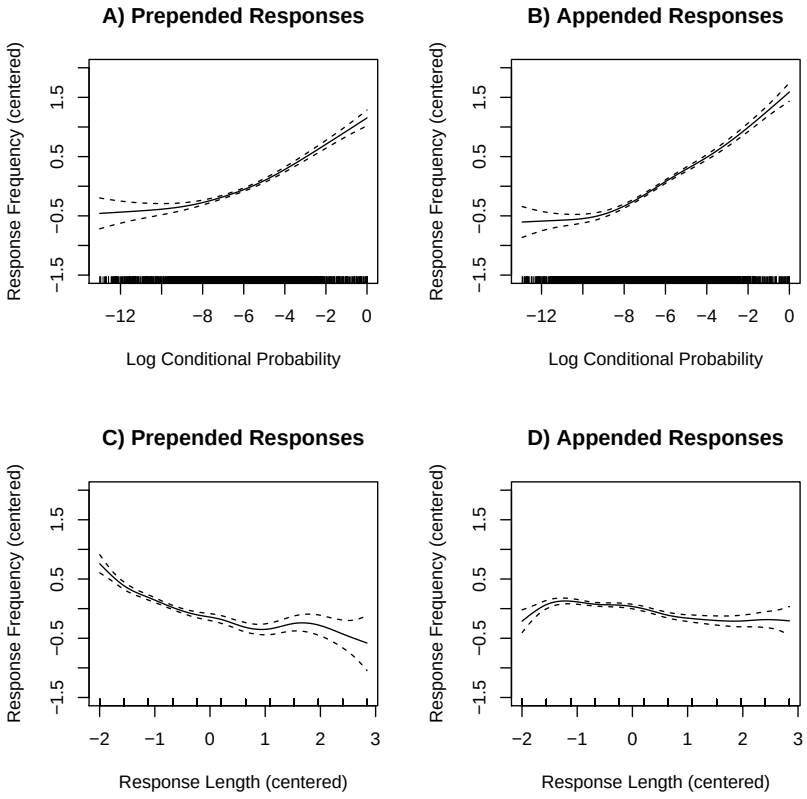

Figure 4.3: Plots of the smooths from the GAM models for response frequency. Relationship between conditional probability and response frequency for prepended (A) and appended (B) responses. Relationship between response length and response frequency for prepended (C) and appended (D) responses.

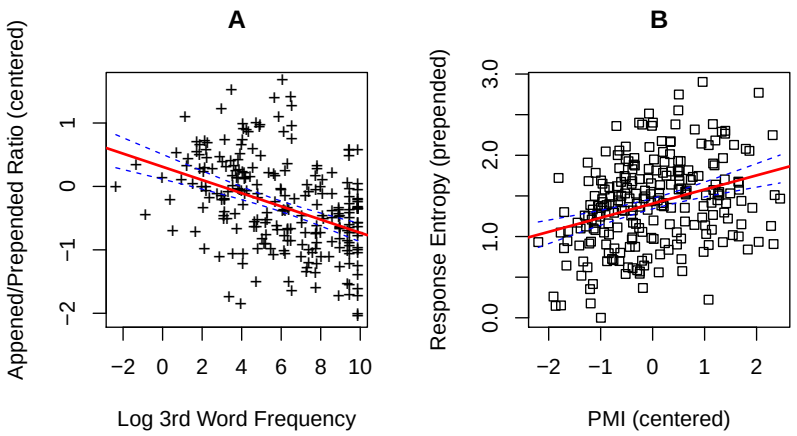

Figure 4.4: Scatter plots including best fit regression lines and 95% confidence intervals for A) the relationship between trigram 3rd word frequency and the ratio of numbers of appended and prepended responses, and B) the relationship between trigram PMI and response entropy (for prepended responses).

was calculated using the following formula:

$$PMI_{\text{trigram}} = log\left(\frac{P_{\text{trigram}}}{P_{w_1} \times P_{w_2} \times P_{w_3}}\right) \tag{4.1}$$

where P_{w_1} is the probability of the first word occurring alone and P_{trigram} is the probability of the three words occurring together. In the field of information theory, entropy, H, was defined by Shannon (1948) to be:

$$H_{N\text{-gram}} = -\sum_{i=1}^{N} P_{N\text{-gram}} log P_{N\text{-gram}} \tag{4.2}$$

where there are N n-grams in a family, each with a probability of $P_{N\text{-gram}}$. I calculated these two values for each of my 240 stimuli. In the case of entropy, I calculated the entropy for both the prepended family and the appended family separately.

During my analysis I used a backward elimination model comparison procedure to eliminate predictors that could be removed without hurting the fit of the model. Using the log likelihood ratio test, I eliminated predictors one by one until I found the model with the best balance of complexity and fit.

The best model for position choice retained only one of my predictors, the frequency of the final word in the stimulus. This model found a negative

linear relationship between the log transformed frequency of the final word of the stimulus and the log transformed position ratio ($\beta = -0.11$, $t(2) = -7$, $p = 2e - 11$). This model was a significant improvement over the null model in a log likelihood ratio test ($\chi^2(1) = 45$, $p = 1.6e - 11$). A visualization of this relationship is shown in Figure 4.4A. Closed class words are much more frequent than open class words therefore this frequency effect may be related to the class of the final word. This interpretation would imply that if a closed class/function word was at the end of one of my trigrams it would increase the proportion of responses at the end of the trigram making the position ratio smaller. An open class word at the end of a stimulus correlated with less appended responses and more prepended responses.

The second analysis was of the response distributions. From a theoretical perspective, production is necessarily constrained by experience. How are the probabilistic properties of the trigram stimuli influencing the distribution of the responses that the participants produced? I counted the number of distinct responses made for each stimulus and called this number the *family size* of the responses for that item. I then calculated the entropy of those responses as well, providing a measure of the amount of order/disorder for the responses to a stimulus. Since the location of the response (before or after the stimulus) will change the constraints that arise from the distribution of completions in the Web1T corpus, I analyzed these two data sets separately. I retained all the idiosyncratic responses for this analysis as they did not impact the results (all of the effects I found were identical when I removed the idiosyncratic responses).

For the responses that were prepended to a stimulus, I found that the none of my models were able to reliably predict the family size of the response set, but there was one model that was able to predict the entropy of the response set. This model found a positive linear relationship between the PMI of the stimulus and the entropy of the responses ($\beta = 0.17$, $t(238) = 5.3$, $p = 3e - 07$). Stimuli that had a higher PMI had higher entropy, or less order, in the set of n-grams created by prepending a word. This model was

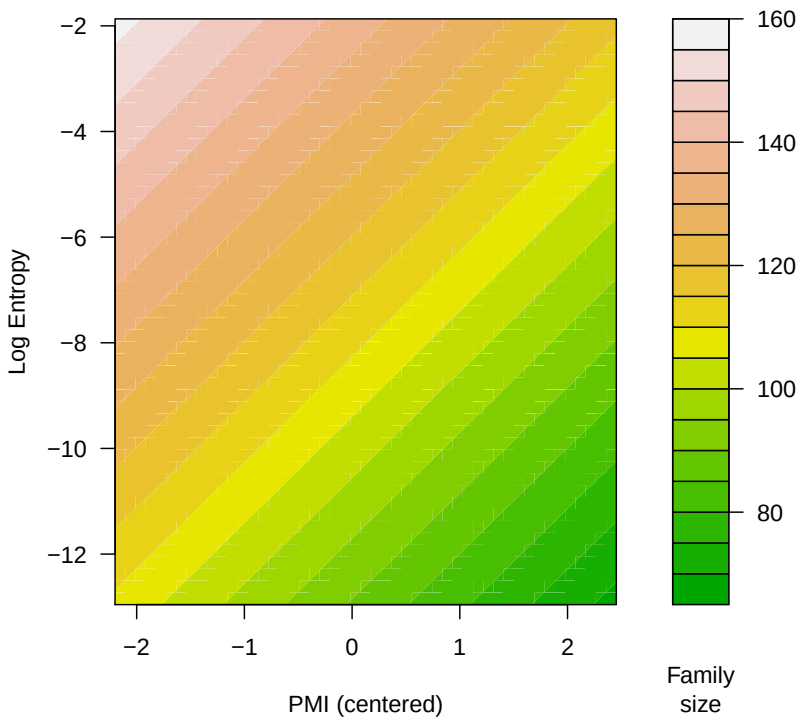

Figure 4.5: Contour plot of the fit for the linear regression model for family size predicted by PMI and Entropy for appended responses.

a significant improvement over the null model in a log likelihood ratio test ($\chi^2(1) = 27$, $p = 2.6e-07$). A visualization of this relationship is shown in Figure 4.4B.

In contrast to the prepended trials, the response entropy data for the appended words did not have any strong relationships with the predictors but there was a relationship found for the family size. The first model included the PMI of the stimulus, and this model was significantly better than the null model ($\chi^2(1) = 9.9$, $p = 0.0016$). This model was compared to one that also included the entropy calculated from the Web1T corpus (the predictability of the final word based on the preceding three words). After including the entropy predictor in the model there was further improvement ($\chi^2(1) = 8.3$, $p = 0.004$). In the final model the coefficient for the effect of PMI was negative ($\beta = -8.7$, $t(237) = -3.1$, $p = 0.0019$), meaning that trigrams with a greater PMI had response families with smaller family sizes. The coefficient for the effect of entropy was positive ($\beta = 4.4$, $t(237) = 2.9$, $p = 0.0042$), and therefore trigrams that were less predictive of the next word in the corpus had a greater variety of responses. A contour plot for the fit of this model is shown in Figure 4.5, and can be thought of as a plane intersecting the space of possible family sizes. The greatest variety of responses was produced for stimuli with high entropy and low PMI, and the least variety of responses was produced for stimuli with low entropy and high PMI.

These results demonstrate the differences in the process of choosing prepended versus appended completions, which depended on the frequency of the final word. The properties of the stimulus influenced the participants productivity as well as the amount of order in the response distributions for each item.

4.4.6 Response Order

In this final analysis I looked at the order in which the responses were generated by the subjects. Each response has a position in the response list from 1st to 20th. I fit models that predicted that position for each response. I used the same set of 18,357 that I used in the previous analysis. I only retained observations for 230 stimuli as 10 of the stimuli were left out under this

criterion.

The dependent measure in this analysis is not a continuous, normally distributed variable. Rather it is an ordinal, categorical variable (participants chose what order to list their responses, and 1 is before 2, etc). Linear models are not well suited to analyze ordinal categorical data; they assume that the outcome is continuous and can make predictions that extend beyond the range of possible outcomes. Cumulative Link Models (CLM), also called cumulative logit models, are a type of ordinal regression model that is well suited to analyzing my data (Agresti, 2010). I used a type of CLM that can also include crossed random effects, Cumulative Link Mixed Models (CLMM). Furthermore, my data satisfy the proportional odds assumption, that the relationship between any two pairs of outcome groups is the same. I used the **ordinal** package (Christensen, 2011) in **R** (R Development Core Team, 2009) for this analysis.

I have two very different types of responses to analyze, responses that were prepended and those that were appended. To better understand the contrast between these two types of responses, I split my dataset in two: 7,464 observations of prepended responses and 10,893 observations of appended responses. I will report the results for these subset separately but the process that I used to build my models was identical.

I used forward stepwise model selection that included all of the predictors as well as interactions between them. All models included crossed random effects for subjects and items. I compared nested models and retained the models that had the best balance of fit and complexity. All models were compared using the log likelihood ratio test.

The best models for the two data segments were:

$$logit(P(Y_i) \leq j) = \theta_j - \beta_{\text{Freq 1st Word}_i} - \beta_{P(\text{First word}|\text{Trigram})_i} - \mu_{\text{subject}_i} - \mu_{\text{item}_i} \tag{4.3}$$

$$logit(P(Z_i) \leq j) = \theta_j - \beta_{\text{Freq 2nd Trigram}_i} - \beta_{P(\text{Last word}|\text{Trigram})_i} - \mu_{\text{subject}_i} - \mu_{\text{item}_i} \tag{4.4}$$

$$Y_i = \text{order of generation for prepended responses}$$
$$Z_i = \text{order of generation for appended responses}$$
$$i = 1, ..., N_{\text{observations}}$$
$$j = 1, ..., N_{\text{categories}} - 1$$

where θ is the vector of response category thresholds, μ_{subject} is the vector of random intercepts for subjects and μ_{item} is the vector of random intercepts for items.

	$\bar{\beta}$	SE(β)	z	$p_{\|z\|}$
Log Response 1st Word Freq	-0.248	0.035	-7.08	1.5e-12
P(First word|Trigram)	-0.151	0.013	-12.12	8.5e-34

Table 4.3: Coefficient estimates from a cumulative link mixed model for response order for prepended responses.

The estimated fixed effect coefficients in my first CLMM for the prepended responses (Equation 4.3) are shown in Table 4.3. The first predictor that remained in the model was the frequency of the critical first word. The frequency of this word, the one the participants chose, had a negative relationship indicating that responses with a high frequency word in the first position were produced before responses with a low frequency word. The second predictor was the conditional probability of the response word given the trigram. The sign of this coeffecient was also negative. The lower the conditional probability, the earlier the response of was generated. Neither bigram, trigram, quadragram frequencies nor PMI were retained during this model selection process as they did not contribute and explanatory power to the model.

	$\bar{\beta}$	SE(β)	z	$p_{\|z\|}$
Log Response 2nd Trigram Freq	-0.156	0.031	-5.10	3.4e-07
P(Last word|Trigram)	-0.184	0.013	-14.24	5.5e-46

Table 4.4: Coefficient estimates from a cumulative link mixed model for response order for appended responses.

For the appended responses, the best model (Equation 4.4) included two fixed effects, shown in Table 4.4. The first predictor was the 2nd (final) trigram and frequency. It had a negative relationship with the order in which the responses were produced. Responses with a higher frequency final trigram (*bottom* ***part of the***) were produced earlier than responses with a lower frequency final trigram (*chocolate* ***chip cookie lover***). The final predictor in this model was the conditional probability of the final responded word given the preceding trigram. This effect was negative, showing that words that were more likely to occur after the trigram in the Web1T corpus given the trigram context were generated earlier. As with the prepended responses, the other *n*-gram frequencies and PMI did not enter the model.

These results provide evidence that the search process the participants used to generate responses in this cloze task was sensitive to the conditional probabilities of the *n*-grams that were created. The predictors that remained in the models can help us understand this search process. For the prepended responses, high frequency words were the first to be generated, in particular high frequency words that were part of a high-probability *n*-gram. For the appended responses, words that had a high conditional probability given the preceding context were generated first, in particular those word that created a high frequency trigram.

Our only *n*-gram frequency effect was position dependent. The frequency of the trigram containing the response word constrained the search process when the response was appended, but not when it was prepended. This result is very similar to the results obtained by Griffin and Bock (1998) in their verbal production task. Their frequency effect was also modulated by the amount of contextual constraint in their sentences.

4.4.7 Discussion

Experiment 2 expanded on Experiment 1 by allowing participants to provide multiple completions for each stimulus. The design of the experiment also allowed me to calculate the conditional probability in the Web1T corpus of each response provided by the participants. I found support for the effects

of conditional probability in the response frequencies and the in the order of generation of the responses. I also found effects of the entropy and PMI of the stimuli on the entropy of the responses and the family size of the responses.

Our results mirror those Crowe (1998) for the prepended responses. As in his letter and semantic verbal fluency tasks, I found evidence that higher frequency words were produced earlier. The same was not true for the appended responses where the word frequency of the responded word did enter into the model because it did not help explain the order data. The amount of constraint created by the letter categories and semantic categories in the verbal fluency tasks is much less than the amount of constraint in the *n*-gram cloze task and this may explain the difference between the appended and prepended responses. Crowe (1998) proposed a mental store of high frequency responses that eventually get depleted and force participants to use a different, slower search strategy to complete the task. My data do not support this model since I found a continuous, linear, negative trend for the order of the subjects' responses, with not change in strategy evident.

I will discuss how the data from Experiment 2 fits into the larger landscape of psycholinguistic theory in the following section.

4.5 Conclusion

I performed two experiments where I asked participants to complete *n*-grams that required either a letter or word to be added to them. The only constraints in this task were the other words in the *n*-gram, the linguistic micro-context. I uncovered evidence that probabilistic measures derived from large samples of language predict which words participants choose to complete these cloze *n*-grams and the order in which they generate them. For the first time I applied frequencies, conditional probabilities, PMIs and entropies extracted from a one trillion word corpus of English to this type of psycholinguistic task. The relationship between the corpus measures and the observed behaviour imply that a similar kind of probabilistic information is available to the language system when choosing a completion in a cloze task. Since the conditional

probability of a word in context can only be learned from linguistic experience, and since experiences are stored in memory, I submit that the process of completing the cloze task was a memory process. I will now incorporate theoretical considerations from research on linguistic memory and cognition to better understand how this probabilistic information is used in the process of completing fragments.

First I will look to the research on memory for theoretical considerations. How much of the cloze task response process can be ascribed to memory, to recall? The answer must be tied to the process of remembering words. Tulving (1985) famously drew a distinction between episodic and semantic memory. Abstract word memory, along with memory for other types of factual knowledge, was categorized as semantic memory. Memories for events were categorized as episodic. Under this assumption of a separate lexical memory store, the first psycholinguistic models of lexical memory were simple affairs — memory was seen as containing localist units for each word, and each of those word memories was said to have a certain strength plus links to associates (Collins & Loftus, 1975). These models did not include episodic memories for words in the lexical memory model. In a break with these ideas, distributed connectionist models have challenged the validity of localist models since the 1980s (M. Seidenberg & McClelland, 1989; Sibley, Kello, Plaut, & Elman, 2008). Elman (2011, 2009) has eloquently shown how lexical relationships can arise from temporal statistical patterns and how meaning can emerge from distributed systems without any local semantic representations in memory. In his dynamic model of lexical memory there is no need to differentiate between episodic and semantic — all information from episodic memory is captured in the probability space representation of the language. Despite the lack of consensus on the representation of language in memory, we can still look to evidence from memory research for some perspective on my results from the *n*-gram cloze task. I found evidence for frequency and contextual probability effects in the responses. If there are similar frequency and context effects in analogous memory tasks, this would support my position that performing the cloze task relies heavily on memory systems.

The statistical properties of language have been thought of as a nuisance variable in many studies because they have been found to influence memory tasks that use words as stimuli. Instead of trying to eliminate frequency effects by matching and counter-balancing average frequency, Criss, Aue, and Smith (2010) systematically manipulated the contextual variability and orthographic frequency of the cues and targets in a paired associated cued recall task. They found that high frequency targets were recalled better, independent of the context variability and frequency of the cue. The frequency of the cue did not influence recall, but cues with low context variability had the effect of improving recall performance. Is this result relevant in light of my results? From the perspective of memory theory, the cloze task could be thought of as a cued recall task without a study list, but with an enormous amount of exposure to the stimuli. The *n*-gram is what the participant uses to probe memory. In light of these similarities, the results from their cued recall task and my cloze task are convergent. As in their cued recall task, high frequency targets were recalled more frequently (our result from Experiment 1). The constraint created by a low context variability cue improved recall, and it also increased the number of responses recalled for items in Experiment 2 (independent of the main effect of PMI, see Figure 4.5). Finally, the frequency of the cue was not a strong predictor of response order for the appended responses, which is similar to a forward recall task. This parallels their finding that cue frequency did not affect their recall probability. The similarity of the pattern suggests that a similar process is taking place during cued recall and the cloze task.

There are similarities between the *n*-gram cloze task and the lexical free association task. Nelson et al. (2000) proposed a theory of the free association process that models word choice as a system that samples a word from a distribution of candidate words when participants are given a certain free association cue. The context of three or four words that were given in the *n*-gram cloze task was greater than that of the single word cue in the free association task. With the increased constraint, it is difficult to see how the sampling process proposed by Nelson et al. (2000) is directly relevant to the *n*-gram

completion process (similar to the difficulty in aligning my results with the results from verbal fluency research). Recently Nelson and McEvoy (2007) have suggested that associations are best modelled as entangled quantum states. These models may have more relevance to the question of *n*-gram processing, but these quantum formalisms are currently in the embryonic stages of development and cannot be used to computationally model behavioural data yet. It appears that there no evidence for convergence

In the discussion of Experiment 1 I noted the importance of contextual distinctiveness in explaining frequency effects (McDonald & Shillcock, 2001). This point was further supported in Experiment 2 where raw *n*-gram frequency was subsumed by conditional probability in the analysis of response frequencies and response order. These results appear to differ with those of Arnon and Snider (2010) and others, who found raw frequency predicted changes in behaviour. I continue to argue that *n*-grams are being processed in a similar way to words. The reason for this apparent discord in the results is due to a confounding of frequency with other important variables. I propose that the *n*-gram frequency benefit that has been found in recent psycholinguistic experiments is not a consequence of the raw count of exposure. Rather, much like the effect of word frequency, the effect of *n*-gram frequency is epiphenomenal. Baayen (2010a) has argued persuasively that the information from the linguistic context is what dictates facilitation, not merely exposure. My results provide further evidence that context, captured in conditional probability, is the a dominant force in processing *n*-grams.

This research is exploratory in nature and not a definitive adjudication. This is the first *n*-gram cloze task experiment that I know of. I have not as of yet fit any computational models to this data but I believe it would be beneficial to attempt to computationally simulate the data I have collected. To spur the development of computational models of word generation I have provided all the raw data from my experiments at the following location: `http://tinyurl.com/ClozeExperiment`.

After minimal modification, some current computational models of multiword reading, such as the NDR (Baayen et al., 2011), the Bayesian Reader

(Norris & Kinoshita, 2008), simple recurrent networks (Mirman, Graf Estes, & Magnuson, 2010) and neural networks (Dilkina, McClelland, & Plaut, 2010c) are potentially capable of modelling this data. All of these computational models are trained on a corpus of text and build a large network of probabilistic relations between form and meaning. The NDR is particularly promising because it has already been applied to multi-word input (Baayen & Hendrix, 2011). The NDR can generate predictions about upcoming words in a stream using the conditional probability of the sub-lexical information which is contained in the models association network. I hope to apply the NDR model to this data to see how well it can simulate the participants' cloze task performance. Simulations by computational models will undoubtedly reveal more about the workings of the word production process. If these prediction-based computational models turn out to be good models of the cloze task I created, it will validate the concept of the automatic forward-modeling system put forward by Pickering and Garrod (2007).

The existence of micro-context effects has implications for models of word selection in language production. My results are only directly applicable to isolated short *n*-grams, but it is conceivable that during the production of longer utterances, and the writing of text, the probabilities from the micro-context of the previous few words has a effect on the next word produced. Once that word is chosen, the micro-context moves forward and begins to influence the choice of the next word, and so on. The question that can now be addressed is: how do the micro-context (the local quadragram, for example) and the macro-context (the sentence or paragraph) interact when language is being produced. The power of human memory to retain *n*-gram experiences helps explain the extraordinary fluency that adults exhibit when speaking and writing. My results may also have implications for theories of verbal Working Memory (WM). Cowan (2008) proposed that WM functions emerge from the temporary activation of domain-specific long-term representation under the guidance of attention. Acheson and MacDonald (2009) feel that serial ordering in language production is intimately linked to verbal WM. My *n*-gram results expand the grain size of long-term representations that must be activated

during language production. Future research into verbal WM may find that *n*-gram memory is involved in many verbal WM tasks.

Future research with cloze tasks will not only help us understand how we read groups of words and choose words during production, but also how we learn the meanings of *n*-grams. There is plentiful evidence that infants use the statistical patterns of language to learn how to speak and understand speech (Saffran et al., 1996). Infants have been shown to use statistical learning when learning both artificial and natural languages (Hay, Pelucchi, Estes, & Saffran, 2011) and recently implicit statistical learning has been seen in adults as well (Conway et al., 2010). The cloze task can provide valuable data about how our word selection processes operate, and computational models can help us link language acquisition theories and language processing models.

There are powerful anticipatory processes at play in single word processing, *n*-gram processing, and sentence processing. The impact of context in the *n*-gram cloze experiments was pervasive and suggestive of a link between contextual memory and word selection. Without the computational infrastructure and extensive *n*-gram frequency data that is now readily available I would not have been able to attempt to understand the processes underlying word choice in a cloze task. As the large data trend continues to progress other previously intractable problems in psycholinguistics may soon become tractable as well.

4.6 Appendix: Stimuli, Experiment 1

Letter Completion	Word Completion
could you _ay	_____to do with the
the _at is	an exception to the _____
in the _one	have no time to _____
with a _ag	keep a firm _____on
going to _reak	go _____in the face
lots of _arts	at some _____in time
the _ay to	share your _____with others
into the _aves	pull out all the _____
and _illing in	the taller of the _____
is a _it	_____album of all time
_ate with the	is _____on thin ice
_ell and then	going to a _____tonight
you want to _eep	the midterm _____will be
what are you _iving	the effects of _____warming
the house is _old	I was _____my bike
had been _ought in	it is a _____day
is a _ap of	no matter what _____is
I have _et the	best interest of the _____
it is the _eat	would you like to _____
got a _an in	the _____is to help
of the _eal was	on the basis of _____
look at the _ash	looking for a _____vacation
	_____found out how to

Table 4.5: Stimuli for the letter and word completion tasks.

4.7 Appendix: Stimuli, Experiment 2

Trigrams	Trigrams (cont)	Trigrams (cont)
a couple weeks	hanging out in	potentially toxic compounds
a minimum threshold	happily ever after	press release can
a more moderate	have more to	preyed upon the
a report detailing	have not been	printable on your
about what the	hereby certifies that	printable telephone numbers
account at official	homeland security threats	quality of and
act on behalf	identify the potential	quality to your
affair with a	if you are	quite a variety
aggregate amount payable	illustrative purposes only	rarely the case
an agent who	image of an	reliable but not
an income stream	implication or otherwise	reliable migration from
an unbelievably low	in into the	relocating overseas canadians
and all my	in the advance	repeats of the
and are of	in the greatest	residential real estate
and in the	in the helping	reverse chronological order
and into the	in the moment	revolves around the
and on behalf	in the to	scripts for the
and the return	in tracking the	shall be certified
are health care	including air conditioning	shines white light
are to facilitate	including more than	shipping rates for
as to claim	increase in the	small engine repair
as to state	instructor led training	sooner rather than
asbestos attorney lawyer	internally powered by	source software community
at game show	into the website	speak not of
authorize appropriations for	is against a	specs of the
backed sequin belt	is approaching a	speeds of up
bacterial flagellum is	is designed in	strictly prohibited without
barriers to entry	is to elucidate	submitting to a
because it did	it just comes	subsequent to this
bird flu virus	its jurisdiction the	talk about what
bottom part of	java mortgage calculator	technical to make
brutally murdered his	judicial paperwork that	text have been
business and for	keen interest in	the entire class
but to a	kiln lime production	the film have
butchering technique of	liberates toxic gas	the on position
by telling the	locally advanced or	the on site
by the thread	mad doctoring skills	the response you
calories you burn	make it the	the same the
can be huge	member at the	the special meaning
can be left	molecular mass of	third most popular
can be required	more rather than	those with an
center offers a	more support than	thoughts with the
chocolate chip cookie	more will determine	to and establish
chronic pain condition	musical or comedy	to be impeached
click here if	my bloody valentine	to force the
combined shipping rates	not can not	to into the
comes back to	obligated to pay	to please contact
comparing store ratings	obstructive pulmonary disease	to rental cars
comply with a	of an underlined	to say goodbye
components is not	of cruises aboard	to that contained
compulsive behaviors which	of it comes	to the years
constitute endorsements of	of members for	to those offered
cooling plant setting	of the what	treasure trove of
credited alongside another	of things past	trusted source for
details of and	of ulcers caused	tucked away in
did not like	of your password	two consecutive years
died last week	on by his	under difficult circumstances
dietary supplements have	on store shelves	unequally yoked with
dietary supplements with	on this page	urinary tract infection
distinction between public	one able to	used to love
ditch and bank	one iota of	usher dashboard confessional
doting grandpa of	one of this	virtually all of
ears perked up	organizations all over	was above the
electric mixer until	our cover showed	way it was
electromagnetic waves of	our staff who	website shall be
explores essences of	outdoor activities such	weight loss vitamin
fits nicely into	outline of your	when he fought
flattens out the	particles in the	wherever they are
floodlit tennis court	payment details and	which equals the
for something new	pending renewal or	which may be
for the most	per day or	who carried out
friends and the	per year of	whoever posted them
from the finest	performance of three	with obtaining the
front porch of	perked me up	with too many
fullest extent of	physical high all	with you can

gallons per day	pill weight loss	workshops held in
genetically modified organisms	place to consider	worthy of more
glutton for punishment	pleads guilty to	years of credited
grain leather upper	polyester blend fabric	you realize that
hang to dry	polyphonic ring tones	you would on

4.8 Appendix: Analysis of items dropped from Experiment 2

The full list of items that were eliminated from the dataset is provided in Table 4.6. Descriptive statistics for these two groups are given in Table 4.7. The dropped stimuli were an average lower in frequency, lower in entropy and higher in PMI than the retained stimuli.

Items Dropped from Exp.2	
account at official	explores essences of
asbestos attorney lawyer	to those offered
backed sequin belt	at game show
cooling plant setting	implication or otherwise
ditch and bank	relocating overseas canadians
homeland security threats	unequally yoked with
java mortgage calculator	constitute endorsements of
judicial paperwork that	of an underlined
kiln lime production	of the what
reliable migration from	our cover showed
technical to make	hang to dry
usher dashboard confessional	internally powered by
in the to	shines white light
printable telephone numbers	as to claim
liberates toxic gas	to rental cars
including air conditioning	the film have
mad doctoring skills	not can not
whoever posted them	to into the
the on site	compulsive behaviors which
of cruises aboard	

Table 4.6: Trigram Stimuli dropped from in Experiment 2 due to lack of quadragram frequency data in the Web1T corpus.

Statistic	Mean for Dropped Stimuli	Mean for Retained Stimuli	Cohen's d', 95% CI
Log Frequency	-4.05	-1.52	1.39 (1.3 , 1.5)
Log Entropy	-8.6	-6.1	1.41 (1.3 , 1.5)
PMI	11.61	7.12	0.85 (0.7 , 1)

Table 4.7: Comparison of dropped and retained stimuli. Bootstrapped 95% confidence intervals for Cohen's measure of effect size, d', are included.

Chapter 5

The nature of *n*-gram processing

Up to this point I have presented my case for an information-centric view of *n*-gram processing based on the work of those before me and the results of my experiments. In Chapter 1 I reviewed the history of *n*-gram research within the context of psycholinguistic inquiry. An explanation of the current state of *n*-gram research led to an exposition on the rationale for my research. In the following chapters I described my results from three distinct lines of research that involved *n*-gram processing. The subjective frequency of *n*-grams was investigated in Chapter 2, the reading of *n*-grams in Chapter 3 and the production of *n*-grams in Chapter 4. These three lines of research had much in common: they all used *n*-grams as the experimental stimuli and they all attempted to uncover relationships between corpus-derived probabilistic measures of *n*-grams and the participant's behaviour. They differed in the type of task that was involved and the dependent variable that was measured.

5.1 General Discussion

In this final chapter I will bring together the conclusions drawn from each of these lines of research and offer my thought about what my research has to say about the nature of *n*-gram processing.

After summarizing my conclusions, I will then address how my research relates to certain ongoing debates in psycholinguistics. The main issues that I will touch upon in this chapter are: the debates on the nature of the lexicon, the debates on the importance of linguistic storage vs. linguistic computation,

and the debates about emergentist models of language.

The evidence I have brought forth in this dissertation paints a complex picture of *n*-gram processing, but there are some general conclusions that I can now make about it.

5.1.1 *N*-grams are more than the sum of their parts

It is not enough to have information about the individual words in an *n*-gram. There is an independent, important contribution to be made by holistic information. I have taken precautions in my experiments to always include the effects of lexical, component *n*-gram **and** whole *n*-gram statistics into my models of experimental phenomena. The finding across the gross majority of my experiments was that the holistic variables invariably made contributions above and beyond the component variables. First, the whole *n*-gram frequency was involved in subjective frequency judgements. Second, the whole *n*-gram frequency, coherency and information content were predictive of the reading time, probability of regressive saccade, and number of fixations. There was also an early effect of *n*-gram frequency on first word reading time. Third, the whole *n*-gram conditional probability was the most influential predictor of cloze performance (Chapter 4, Experiment 2). All of this gives strong support to my assertion that all *n*-grams are full-fledged psychological entities.

This conclusion may appear to obvious to any user of language, but in many ways it is a subtle point. It is indeed obvious that changing the order of any two words in an *n*-gram could completely change how it is processed, despite the fact that the same words are involved. *N*-grams contain not just information from the words they contain, but also information about the order in which they are sequenced. The subtlety here is that for non-compositional *n*-grams (the majority of the stimuli used in my experiments), there is no clear semantic analysis. Yet despite their incompleteness as phrases, they have holistic properties. The generality of our holistic effects are the hallmark of sub-symbolic processes, and not predicted from symbolic, parse-based theories.

This view of non-syntactic processing is not without its detractors. The theory proposed by Jackendoff (2007) or others who subscribe to theories of

syntactic modules and the segmentation of sentences based on parsing would not align with this view. Kuperberg (2007) has looked at the many studies on the time course of the N400 and the P600 syntax signals in ERP data and she takes a conciliatory tone. She finds a modular syntax-then-semantics theory implausible, and lays out a dual process model, with a fast semantic memory-based constraint processing system operating in parallel with a combinatoric system, and with the two systems exchanging information during sentence comprehension. In future ERP studies it will be possible to discern if *n*-grams are engaging in this combinatoric system that is sensitive to morpho-syntactic and thematic-semantic constraints.

5.1.2 *N*-gram processing is an anticipatory process

How do my results help explain the anticipatory nature of *n*-gram processing? In this section I will describe how each line of research provided evidence of prediction.

The first line of research into subjective frequency judgements was not directly investigating how *n*-grams were read, but there is one aspect of these experiments that has an indirect relationship with prediction. I speculated that the reason that *n*-gram frequency was correlated with the subjective frequency ratings was that higher frequency *n*-grams had more diverse contexts. Contextual diversity implies lower relative entropy (Baayen, 2010a) and therefore greater predictability. My data is insufficient to determine if the signal contained in the entropy of the *n*-gram (above and beyond its frequency) is involved in the production of subjective frequency ratings or relative frequency judgements, but future research may allow us to untangle frequency and entropy and better understand this phenomenon. If an experimenter selected matched pairs of *n*-grams which had very similar frequencies but differed in their contextual predictability (or vice versa), new conclusions could be made about the relationship between prediction and subjective frequency processing.

In Chapters 3 and 4, there was a temporal component to each task. In the trigram reading task, the participants have a visual fixation on the first word, and then they read the rest of the trigram, usually without any regressive

saccades, finally moving their eyes off the screen to signal the end of the trial. In the analysis of SG1 and SG2, the unfolding of the process became clear. The reading time for the first word was partially influenced by *n*-gram frequency and TIC, but the reading time for the first two words was not. I take this as a signal that there was an early prediction made by the reading system about how familiar and informative the *n*-gram would be based on the first word and a para-foveal preview of the second word. No more information was gained during the reading of the second word, explaining why the reading time of the first two words was not influenced by the *n*-gram statistics. For the full reading time for all three words, *n*-gram frequency and TIC effects returned. This is perhaps the clearest evidence from my research that the visual system and the reading system are fully interactive and share significant amounts of information, with the lexical prediction system guiding the saccadic system to optimize reading times for trigrams.

Despite the fact that I did not collect any timing data in the cloze task experiments (in particular, Experiment 2), there is a clearly anticipatory aspect to the results. In that experiment, the participants read a trigram and then produced a word that created a quadragram. When they read each trigram, they saw the blanks in the order they read the stimulus, meaning that the position of the blanks became part of the visual experience. Even though the provided instructions asked them to "type in the first word that comes to mind", the order in which they generated their responses was related to the order that they thought of the responses. The statistical relationship between the trigram and the completion word was between conditional probability of the temporal sequence of all four words occurring together in the corpus. The answer that was easiest to anticipate (due to its likelihood) was the produced first. This demonstrates how an unwritten, anticipated word can influence the way the *n*-gram preceding it or succeeding it was processed.

Altmann and Mirković (2009, p. 585) identified four principles that define anticipatory, incremental models of language:

1. Mapping across domains: Structure in language has significance only insofar as it covaries with, and enables predictions

of, structure in the external world (event structure). Sentence comprehension consists in realizing a mapping between sentence structures and event structures.

2. Prediction: "Knowledge" of the language can be operationalized as the ability to predict on the basis of the current and prior context (both linguistic and, if available, nonlinguistic) how the language may unfold subsequently, and what concomitant changes in real-world states are entailed by the event structures described by that unfolding language. Such predictions constitute the realization of the mapping between sentence structures and event structures
3. Context: Concurrent linguistic and nonlinguistic inputs, and the prior internal states of the system (together comprising the context), each "drive" the predictive process, and none is more privileged than the other except insofar as one may be more predictive than the other with respect to the subsequent unfolding of the input.
4. Representation across time: The representation of prior internal states enables the predictive process to operate across multiple time frames and multiple levels of representational abstraction. The "grain size" of prediction is thus variable, with respect to both its temporal resolution and the level of representational abstraction at which predictions are made.

The above quote refers to sentence processing, but the same process is undeniably taking place during *n*-gram processing. The fourth principle is particularly relevant. It is the idea of the flexible grain size that makes these models capable of modelling *n*-gram behavioural data.

Altmann and Mirković (2009) also point out that the idea of mental simulation proposed by Pickering and Garrod (2007) and Glenberg (1997) is compatible with incremental models like Simple Recurrent Network models (SRNs, Elman, 1990). Simulation would be the equivalent of changes to the internal state of the common language and sensorimotor domain state-space. Changes in the state of the real-world or the body could be predicted from events described in language (i.e. *I hit the ....* enabling the simulation of a motor program for the arm and helping to predict the following word, which could likely be *ball.*).

All of these anticipatory aspects of *n*-gram processing support more gen-

eral notions of forward-modeling as a basic function of the language systems in the mind. This is what Bar called the "pro-active brain" (Bar, 2007, 2009). Van Berkum (2008) has used ERP methods to show this process in action during language use. Frank and Vigliocco (in press) have built a model of sentence comprehension that treats sentence processing as mental simulation of the "real world" with anticipation at the core of the model. Kukona, Fang, Aicher, Chen, and Magnuson (2011) have also looked at anticipation fixation on upcoming words in a sentence based on prediction and situational constraints. All this activity is evidence for a strong interest in the anticipatory processing that takes place when using language. Below I will discuss possible future directions for mental simulation research.

5.1.3 *N*-gram Frequency Effects are epiphenomenal

The question of orthographic frequency is another one that I have addressed in this dissertation. What is frequency measuring? In particular, the results presented in Chapters 3 and 4 showed how entropy, information and conditional probability interact with or supplant pure corpus frequency in models predicting performance on various *n*-gram tasks. Following the ideas of Baayen (2010a) and McDonald and Shillcock (2001), I conclude that in much this same way that lexical frequency effects are epiphenomenonal, *n*-gram frequency effects are also epiphenomenonal. *N*-gram frequency indirectly measures contextual richness, and the way that it influences subjective ratings of frequency, reading speed or completion probability in a cloze task is a function of this contextual richness.

Do any theories of language agree on the epiphenomenal nature of frequency effects? One theory of reading that relies heavily on contextual diversity is the Lexical Quality Hypothesis (Perfetti, Hart, Verhoeven, Elbro, & Reitsma, 2002). At the core of this theory is the idea that the richer the representation of the word in an individual, the faster and more accurately he or she will be able to process those words. The quality of the lexical representations is measured by testing a person's ability to spell words correctly. All representations are the result of exposure to words, so this theory predicts that readers

who read more will have more exposure to context and therefore richer representations. If one were to extend the Lexical Quality Hypothesis to *n*-grams, how would the theory look? The theory calls upon two types of processes to explain word reading performance: *top-down* processes (inference from contextual, semantic and world knowledge) vs. *bottom-up* (perceptual recognition processing). Experimental work by Andrews and Bond (2009) has found that expert readers use less top-down processing and more bottom-up processing when recognizing words. The representations required by the bottom-up process in the Lexical Quality Hypothesis are "minimally constrained by semantic context" (Perfetti, 1992). The Lexical Quality Hypothesis tries to understand the interplay between these two types of processes.

It is unclear if the same interplay between top-down and bottom-up processes described by Perfetti (1992) are at play during *n*-gram processing. One way that I speculate *n*-grams are involved in reading is that there are concurrent "sliding windows" for words, bigrams and trigrams. As these windows slide forward, certain bigrams or trigrams that are high in contextual diversity may appear in the window at any point in time. Depending on the individual's experience with that *n*-gram, the individual may recognize the *n*-gram. If there were to be a theory based on this idea, that theory might be called the *N*-gram Quality Hypothesis. It would predict that expert readers would be better at accessing the meaning of those *n*-grams that are "minimally constrained by semantic context". I will discuss this possibility further in the final section of this chapter.

5.1.4 *N*-grams and the existence of a mental lexicon: Storage versus computation

There is no real debate about the purpose of language: language is for conveying meaning. Words have meaning, and words can be listed in a dictionary, with their meanings listed next to them. The real question for believers in a mental lexicon is how is that meaning represented in the brain? In this dissertation I have not explicitly asked my subjects to perform any semantic tasks with *n*-grams, so it may seem odd for me to try to weigh in on the issue

of the mental lexicon and the semantic information that it may or may not contain, but there is a very relevant point to be made here. The critics of the localist models of the mental lexicon have make strong arguments against any type of static representation based on the impact of context on words (Elman, 2009, 2011). In Elman's lexicon-free theory of meaning, the meaning of words correspond to one of innumerable mental states. These mental states are islands of stability in a dynamic state-space, and are an emergent property of a complex, dynamic system. *N*-grams are merely one type of context that can change the meaning of a word, and since any type of stimulus can push the system into a new state, *n*-grams are going to be one of the constructions that will help explain how dynamic models of meaning work. Furthermore, the meaning of *n*-grams will also be a point in state-space that is determined by the context that the *n*-gram was found in. This is the elegance of Elman's idea; the grain-size of meaning is not restricted to the morpheme or the word: as the unit of language grows, the dynamic system still tracks the meaning of that stream at all points in time, both before it is stable (while the meaning is unclear) and when it is stable and the meaning emerges.

Another point to be made is that all the work on single word reading has pointed to lexical access being inseparable from lexical semantic access (Binder et al., 2003). The only logical conclusion we can make for *n*-grams is that they too can be accessed, and so *n*-gram access must also be inseparable from *n*-gram semantic access. The semantics of *n*-grams does not have to be localist: HAL (Lund & Burgess, 1996), LSA (Landauer & Dumais, 1997), HiDEx (Shaoul & Westbury, 2008) and BEAGLE (Jones & Mewhort, 2007) are good examples of emergent models of semantics that show how meaning can arise from co-occurrence. I can only speculate about how *n*-gram semantics work, but there is no reason other than practical, computational hurdles that prevents all of these models from explaining the meanings of *n*-grams from co-occurrence. The BEAGLE model already contains a holographic representation of all *n*-grams in a corpus, and can use this information to predict word transitions. BEAGLE is not billed as a model of *n*-gram semantics, but I believe that with some modification, it could be used to model *n*-gram

semantics.

This brings us to another ongoing debate: are the meanings of complex words stored or computed? What about the meaning of *n*-grams? Are they stored along with the *n*-gram or are they computed on-the-fly as the words roll in to the language module? This fire has been fanned by some of the rhetoric in the initial papers on *n*-gram processing (the rhetoric in Arnon & Snider, 2010 for example). This whole debate between storage and computation is a red herring. When language is seen as being completely dynamic and interactive, the dichotomy of "stored" and "computed" becomes untenable because it is clearly false.

Before the work of Bannard and Matthews (2008), Arnon and Snider (2010) and Tremblay et al. (2011) there was much theoretical discussion but not much empirical evidence for *n*-gram processing. There was an overarching assumption that certain *n*-grams would rise above some arbitrary threshold and achieve a special status (often called lexical bundle-hood), creating a clean way of dividing *n*-grams into two categories: "lexicalized sequences" and "other" (Biber, 1999). This idea fit well with the computations versus storage debate. Lexical bundles would be stored whereas all other *n*-grams would not. The results from my research have clearly shown that there is no evidence for a dichotomy. By using a large corpus and picking *n*-grams at random to create representative samples spanning the full frequency range, I have found that the effect of the prevalence of an *n*-gram is not dichotomous (for more on the issues around factorialization and its consequences, see Baayen, 2010b). This graded effect of probability, for both linear and non-linear effects, is a side-effect of the emergent nature of *n*-gram processing.

Since I have called the debate about storage and computation unnecessary, what about the memory system? My point is not that we do not use our memory to remember words, *n*-grams and their meanings. Rather, the dynamics of lexical access and *n*-gram access are a type of memory process. The results from my studies do shed some light on this. In Chapter 2 the frequency of exposure of an *n*-gram, estimated by its corpus frequency, predicted performance on the subjective frequency tasks. The only conceivable way for

participants to measure their own amount of experience with an *n*-gram is to (implicitly or explicitly) use their memory systems. In Chapter 3, I described how the interplay of word frequency, bigram frequency, TIC, PMI, and trigram frequency influenced how trigrams were read. The key point is that PMI, TIC and trigram frequency are properties of the *whole*, not of the *parts*. Since participants were required to hold each trigram in working memory between trials, the reading time of each trigram also reflects the simultaneous recall of the meaning of each trigram from long-term memory.

Furthermore, I discuss in Chapter 4 the similarity between the task of completing an *n*-gram and the task of paired associate recall, a classic example of an episodic memory retrieval task. There is much more work left to do to bridge the worlds of memory research and psycholinguistics, but the recent growth in memory models that take temporal state into account and deal with complex dynamics shows that there is more common ground now than before (Squire & Kandel, 2000; Raffone & Leeuwen, 2003). One example of this union of memory model and language model has been proposed by Zwaan (2008) who has a theory of experiential memory traces and mental simulation that can explain language comprehension. In this model, multi-model memories, traces of temporal patterns of perceptual or motor activity, can be cues into temporal patterns. The concept proposed is one of *presonance*, a combination of **pre**diction and re**sonance**. This model was applied to explain embodied language cognition, but it shows how episodic experience and memory systems can be linked to explain language processing[1].

There is one more concept from memory research that the work of Zwaan (2008) brings to mind. That is the idea of gist memories and verbatim memories (Brainerd & Reyna, 2002). Their fuzzy-trace theory of memory was developed to explain memory illusions such as the DRM false memory phenomenon (Deese, 1959; Roediger & McDermott, 1995), but also has been applied to decision making and other areas. It is interesting to speculate about the possibility

[1]In an interesting case of parallel terminology, Jones and Mewhort (2007) also use the term resonance to describe how an *n*-gram query pattern (which they call a *probe vector* such as *thomas* _____ is able to retrieve the word *jefferson* from the full set of holographic vectors representing the English language.

that every time we read an *n*-gram we are laying down both a verbatim trace and a gist trace. This idea may help us understand how we are able to process the meanings of so many variations of an *n*-gram (*the child said*, *the brat said*, *the brat screamed* , *a brat screamed*, etc.) without needing to retrieve all of the verbatim traces. Fuzzy-trace theory is not a computational model, but some of the current crop of dynamic models of language and memory may be able to model the effects of gist processing which would make *n*-grams flexible. If an trigram had two words in common with another trigram, and if those words were in the same position, a dynamic model of memory would end up in a certain region of state space that would contain all *n*-grams that satisfied that constraint. This process of *n*-gram comparison is speculative but conceivable as a way to implement gist memories for *n*-grams.

5.1.5 Language as an emergent process

Forster (1979) proposed that word processing has a *functional autonomy*, and that lexical identification was encapsulated and immune from contaminations by context, and Fodor (1983) went further and proposed a highly modular mental architecture. If one were to try to shoehorn the ideas of *n*-gram processing into a series of modular processes, the results would be dismal. There is no discrete set of black boxes operating independently and in a certain order that can explain the massively interactive effects of information and conditional probability in my studies. The only solution is to accept the emergentist school of thought and reject this serialized way thinking. By allowing all levels of processing to fully interact, the observed behavior begins to make sense. *N*-gram recognition processes interact with word recognition processes, transitional probabilities interact with sub-lexical statistics, and new input continuously changes the state of the system.

There are still those who hold tight to the "symbols and rules" approach to language systems. Jackendoff (2002) shrugs off emergentist models as irrelevant to human language comprehension. His most current model is steadfast in its loyalty to sequential, directed processing (Jackendoff, 2007). Meanwhile SRNs and other types of models are quickly beginning to dominate the field.

These detailed computational models are capable of explaining how meaning is constructed as language unfolds. To add to the already long list of models using SRNs, Crocker, Knoeferle, and Mayberry (2010) have created an emergentist model of language based on Elman's SRN, that computes the interaction of utterances and visual attention. Misyak, Christiansen, and Tomblin (2010) have used an SRN to predict human performance patterns in a sequential learning task. This is the first study to provide a link between linguistic and non-linguistic learning of probabilistic sequences [2].

Another emergentist model that uses truly sub-symbolic information to explain lexical and supra-lexical phenomena is the Naive Discriminative Reader model developed by Baayen et al. (2011). This model is a two-layer symbolic network model built using the equilibrium equations of the Rescorla-Wagner model (Danks, 2003). Building a complex mapping between letter bigram cues and semantic representations, it is able to simulate frequency effects, morphological family size effects, and some initial *n*-gram frequency effects (Baayen & Hendrix, 2011). The NDR reader is a truly emergentist model, with all these effects arising simply from the interaction of all the sub-symbolic learning taking place at the letter bigram level.

Are there only two types of models at play, or are there more worth considering? The simpler associationist, network models are more attractive than the complicated, nativist systems. Is there a third way? One family of non-connectionist models are the Bayesian models. The Bayesian Reader (Norris & Kinoshita, 2008) is one such model. Bayesian models represent the world as a set of competing hypotheses, and the task of the models is to choose the most likely hypothesis, given the evidence and the level of uncertainty surrounding it. Tenenbaum, Kemp, Griffiths, and Goodman (2011) argues that Bayesian models are valid and relevant, but Kwisthout, Wareham, and Rooij (2011) have criticized these models on the grounds that they cannot be approximated or computed within a reasonable amount of time.

[2]Incidentally, Christiansen, Conway, and Onnis (in press) found that linguistic and non-linguistic sequential learning, as modelled by Misyak et al. (2010), produce very similar ERP responses, implying that language shares domain-general processing substrata with other sequence processing systems.

There is still a strong resistance to emergentist models in many quarters. Why is it so hard to shift from sequential viewpoints to emergent viewpoints? Some have proposed that the sequential process schema are an ingrained, intuitive first attempt at explaining all phenomena (Chi, 2005; Chi, Roscoe, Slotta, Roy, & Chase, 2011). Many phenomena that we observe our whole lives, such as the way that our blood moves around (directed by the heart, moving from point A to point B) are not emergent. When young adults and even some scientists are faced with emergent phenomena, such as the way ants collect food (without the aid of a queen ant telling them where to look or how to cooperate), they are initially resistant to the notion of emergence. In chemistry classes, for example, the way that ink diffuses in water is confusing to students, even after the emergent explanation is taught. This is one of many examples of emergent phenomena that are often misunderstood. Chi et al. (2011) has listed criteria for identifying emergent phenomena, and they fit the language system to a tee. Using their terminology, words would be the "agents", and they note that (1) in emergent systems the interactions of the entire collection of all the agents together "cause" the observable pattern, not any one special type of agent. (2) All the interactions have equal status with respect to the pattern. (3) Agents' interactions and the pattern can behave in disjoint or non-matching ways. (4) Interactions are undertaken by the agents with the intention of achieving local goals only, without any intention of causing the (changes) in the pattern. The pattern emerges from the local interactions of all the agents. (5) The pattern is caused by the collective summing or net effect of all the interactions at each point in time. (Chi et al., 2011, p. 10). This difficulty in understanding the difference between emergent and non-emergent phenomena in all the sciences (including physics, chemistry and biology) is a ongoing issue, but psycholinguistics should move forward and accept the fact that the mind is a emergent phenomenon and that there is no other way to explain the combined action of all the brain activity we observe.

5.2 Future directions

The sub-field of *n*-gram processing holds great potential for the field of psycholinguistics. Because of their short length, they are amenable to techniques that quickly balloon in complexity when faced with longer stimuli, like sentences. In particular, models that take into account temporal dependency and initial conditions can take advantage of experimental results from *n*-gram experiments and attempt to explain *n*-gram processing phenomena that would be impossible to attempt with current computational resources on longer passages.

After completing these three lines of research described here, many ideas for potential follow-up studies have occurred to me. The individual experience of the feeling of familiarity with words is something that has no compelling computational model. Orthographic frequency can only explain a portion of the variability in subjective frequency judgements. Larger datasets of subjective frequency ratings for *n*-grams should be collected using the representative sampling techniques I developed in Chapter 3. With thousands of *n*-grams rated by hundreds of participants a much more nuanced view of subjective frequency would emerge. *N*-gram entropy effects might become apparent, but other features, such as semantic coherence might also have an influence.

New directions also came to mind after working with the *n*-gram reading eye-movement data. When people read sentences, they may be segmenting them into *n*-grams to better understand their meaning. By looking at fixations in natural reading of longer passages and focussing on which words are skipped and where the eyes pause, it might be possible to detect real-time *n*-gram processing. Concurrent ERP data collection could help detect the markers of *n*-gram detection in the dynamics of the electrical activity. Studying *n*-gram processing using neuroimaging techniques such as fMRI and functional Near Infrared Spectroscopy (fNIRS, Bunce, Izzetoglu, Izzetoglu, Onaral, & Pourrezaei, 2006) is another option. With careful attention to the experimental design it should be possible to create contrasting conditions that would allow the researcher to tease apart the effects of different anticipatory processes.

There remains much more to be done in *n*-gram production research as well. Instead of asking participants to produce one word to complete a 4-gram, why not ask them to produce two or even three words? This reduction in the amount of constraint could allow a greater variety of responses to be collected, and, using the statistical methods that I proposed in Chapter 4, the probability of each word selection could be calculated in the context of each word. This data would also be valuable to researchers who are creating dynamic models of language production, as the initial conditions would be the same across all subjects for each trial, but the outcome would often be very different.

All of my research presented here would not have been possible without the corpus of web documents made available by Brants and Franz (2006). The size of the corpus (1 trillion words) and the quality of the *n*-gram counts inspired me to attempt this research and I am grateful to Google for their generosity. In the future I see larger and better corpora being produced for us by the scientific community. Already Google has released 100 million word corpora for non-English languages (Brants & Franz, 2009). In the cloze task research, I noted that there were many 4-grams that were plausible, but not included in the Web1T database and this fact leads me to the conclusion that better corpora are vital to the continuation of *n*-gram processing research.

Corpus quality is not only related to size. The content of the corpus also matters. It is obvious that the Web1T corpus has a specific web-centric bias to some of its frequency data (one of the most frequent trigrams in the database is *all rights reserved*, which is most definitely not that frequent in everyday usage). A challenge for future *n*-gram research will be to find large, relevant corpora to improve the selection of stimuli and to improve the quality of the computational models that are trained on the corpus and are then used to simulate behaviour (Recchia & Jones, 2009). Corpora that incorporate more spoken language and that approximate the reading patterns of people over their lifespan would be best. Until such corpora are available, psycholinguists will have to use resources such as the Google Web1T corpus which provides the broadest coverage currently available.

Another untouched area of *n*-gram processing is the study of individual differences: Are expert readers (those who have better reading speed and comprehension) better at *n*-gram processing tasks than normal readers? If they are, then *n*-gram processing skills may be one of the key skills, along with word identification, that should be practiced to improve reading. As Stanovich (2000) notes in his review of reading research, there have been ongoing "Reading Wars" over the best method of teaching children to read, and these wars will continue until we have a better understanding of how the psychology of reading works. Quick and accurate *n*-gram identification could help poor readers become better at reading. Identification of *n*-gram processing deficiencies in those with speech or language impairments could be beneficial in offering them the best type of therapy. For those with certain types of aphasia, *n*-gram related therapy may also aid them in improving speech and understanding. Much work will need to be done on *n*-gram processing.

Language acquisition research can now be expanded to include the acquisition of *n*-gram knowledge. Bannard and Matthews (2008) and Matthews and Bannard (2010) have already show how young children are sensitive to *n*-gram probabilities. Using all the techniques available to developmental psycholinguists it will be possible to probe for the first detectable *n*-gram understanding, and how sensitivity to context changes over the lifespan. There is already one emergentist model of language acquisition, the DevLex model, that has been developed by Li (2009). It uses SRNs and self-organizing maps (SOM, Kohonen & Somervuo, 1998) to build a dynamic model of lexical learning and language acquisition. One possible way forward would be to extend the DevLex model to handle *n*-grams, another would be to use the NDR model (Baayen et al., 2011) and expose it to child-directed speech corpora during training to see if it can simulate the word-explosion that takes place in small children.

Finally, the most important challenge: building accurate, psychologically plausible computational models. I have repeatedly expressed my preference for distributed, non-symbolic models because I think they have the greatest likelihood of capturing the complexity of *n*-gram phenomena. They can capture the temporal dynamism of the language stream, its dependence on initial

conditions, its non-linearity and its massive interactivity.

After studying *n*-gram processing with three different paradigms and synthesizing the evidence, I feel that I have contributed novel and valuable insights into how *n*-grams are processed. It remains to be seen if the current trend of active interest into *n*-grams continues at its current pace, but if it does, *n*-grams will help us understand how the mind is capable of the incredibly difficult task of understanding and producing language.

References

Acheson, D., & MacDonald, M. (2009). Verbal working memory and language production: Common approaches to the serial ordering of verbal information. *Psychological bulletin*, *135*(1), 50.

Agresti, A. (2010). *Analysis of ordinal categorical data*. Hoboken, NJ, USA: John Wiley & Sons Inc.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Altmann, G., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583–609.

Anderson, J. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369–406.

Anderson, J. (1990). *The adaptive character of thought*. Lawrence Erlbaum.

Andrews, S., & Bond, R. (2009). Lexical expertise and reading skill: Bottom-up and top-down processing of lexical ambiguity. *Reading and Writing*, *22*(6), 687–711.

Angele, B., & Rayner, K. (2011). Parafoveal processing of word n+2 during reading: Do the preceding words matter? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1210.

Arnon, I., & Clark, E. V. (2011). Why Brush Your Teeth Is Better Than Teeth – Children's Word Production Is Facilitated in Familiar Sentence-Frames. *Language Learning and Development*, *7*(2), 107.

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

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baayen, R. H. (2010a). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *5*(3), 436–461.

Baayen, R. H. (2010b). A real experiment is a factorial experiment? *The Mental Lexicon*, *5*(1), 149–157.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baayen, R. H., Feldman, L., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*, 290–313.

Baayen, R. H., & Hendrix, P. (2011). Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. *Proceedings of the Annual Meeting of the Linguistic Society of America*.

Baayen, R. H., Kuperman, V., & Bertram, R. (2010). Frequency effects in compound processing. In S. Scalise & I. Vogel (Eds.), *Compounding*. Amsterdam/Philadelphia: Benjamins.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3.2*, 12-28.

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

Baetes, E., & Elman, J. (1993). Connectionism and the study of change. *Brain Dovelopement and Cognition*, 420–440.

Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C. hui, Morgan, N., et al. (2009). Developments and directions in speech recognition and understanding, part 1 [DSP education]. *IEEE Signal Processing Magazine*, *26*(3), 75–80.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, *29*(4), 639–647.

Bandura, A. (1997). *Self-efficacy: the exercise of control*. New York, US: W.H. Freeman.

Banks, W. (1977). Encoding and processing of symbolic information in comparative judgments. *The psychology of learning and motivation*, *11*(101-159).

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), 241–248.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*(7), 280–289.

Bar, M. (2009). The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1235.

Bates, D. M. (in preparation). *lme4: Mixed-effects modeling with R*. Springer.

Battig, W., & Montague, W. (1969). Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of Experimental Psychology*(392).

Beattie, G., & Butterworth, B. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, *22*(3), 201.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, *60*(1), 92–111.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2004). *Regression diagnostics:*

Identifying influential data and sources of collinearity. Hoboken, NJ, USA: Wiley-Interscience.

Biber, D. (1999). Lexical bundles in conversation and academic prose. *Language and Computers*, *26*, 181–190.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405.

Bicknell, K., & Levy, R. (2010). Rational eye movements in reading combining uncertainty about previous words with contextual probability. In *Proceedings of the 32nd annual conference of the cognitive science society*.

Binder, J., McKiernan, K., Parsons, M., Westbury, C., Possing, E., Kaufman, J., et al. (2003). Neural correlates of lexical access during visual word recognition. *Journal of Cognitive Neuroscience*, *15*(3), 372–393.

Block, C., & Baldwin, C. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior research methods*, *42*(3), 665–670.

Bloom, P., & Fischler, I. (1980). Completion norms for 329 sentence contexts. *Memory & Cognition*, *8*(6), 631–642.

Bloor, D. (1983). *Wittgenstein: A social theory of knowledge*. London, UK: MacMillan.

Bod, R. (2009). From exemplar to grammar: A probabilistic Analogy-Based model of language learning. *Cognitive Science*, *33*(5), 752–793.

Bormuth, J. (1966). Readability: A new approach. *Reading research quarterly*, 79–132.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, *26*(211-252), 57.

Brainerd, C., & Reyna, V. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, *11*(5), 164–169.

Brants, T., & Franz, A. (2006). *Web 1T 5-gram version 1*. Philadelphia, PA USA: Linguistic Data Consortium.

Brants, T., & Franz, A. (2009). Web 1t 5-gram, 10 european languages version 1. *Linguistic Data Consortium, Philadelphia*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Bunce, S., Izzetoglu, M., Izzetoglu, K., Onaral, B., & Pourrezaei, K. (2006). Functional near-infrared spectroscopy. *Engineering in Medicine and Biology Magazine, IEEE*, *25*(4), 54–62.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, *30*, 188–198.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. New York, NY, USA: Springer Verlag.

Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition*, *24*(02), 215–221.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of con-

stituency: the reduction of 'don't' in English.(Statistical data included). *Linguistics: an interdisciplinary journal of the language sciences*.

Chi, M. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *The Journal of the Learning Sciences*, 161–199.

Chi, M., Roscoe, R., Slotta, J., Roy, M., & Chase, C. (2011). Misconceived causal explanations for emergent processes. *Cognitive Science*.

Chomsky, N. (1980). *Rules and representations*. New York, US: Columbia University Press.

Chomsky, N. (2005). *Rules and representations*. New York, US: Columbia Univ Pr.

Christensen, R. H. B. (2011). *Ordinal—regression models for ordinal data.* (R package version 2010.12-15 http://www.cran.r-project.org/package=ordinal/)

Christiansen, M. H., Conway, C. M., & Onnis, L. (in press). Similar neural correlates for language and sequential learning: Evidence from event-related brain potentials. *Language and Cognitive Processes*.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist.*, *16*(1), 22–29.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ, US: Lawrence Erlbaum.

Collins, A., & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407.

Colombo, L., Pasini, M., & Balota, D. A. (2006). Dissociating the influence of familiarity and meaningfulness from word frequency in naming and lexical decision performance. *Memory & cognition*, *34*(6), 1312.

Columbus, G., Bolger, P., & Baayen, R. H. (2010). *Processing Multiword Units: Degrees of Idiomaticity Seen Through Eye Movement Data.* (Paper presented at the Seventh Conference of the Mental Lexicon. University of Windsor, Ontario, Canada, June 30th -.July 3rd, 2010.)

Columbus, G., Bolger, P., & Baayen, R. H. (2011). *Implications for language models: fixation and dwell times reveal important predictors for processing multiword units.* (Paper presented at the European Conference on Eye Movement. Marseille, France, August 21st-25th, 2011)

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*(1), 72.

Connine, C. M., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *16*(6), 1084–1096.

Conway, C. M., Bauernschmidt, A., Huang, S., & Pisoni, D. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*(3), 356–371.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in brain research*, *169*, 323–338.

Criss, A., Aue, W., & Smith, L. (2010). The effects of word frequency and context variability in cued recall. *Journal of Memory and Language*.

Crocker, M., Knoeferle, P., & Mayberry, M. (2010). Situated sentence processing: The coordinated interplay account and a neurobehavioral model. *Brain and language*, *112*(3), 189–201.

Crowe, S. (1998). Decrease in performance on the verbal fluency test as a function of time: Evaluation in a young healthy sample. *Journal of Clinical and Experimental Neuropsychology*, *20*(3), 391–401.

Danks, D. (2003). Equilibria of the rescorla-wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17.

DeLong, K., Urbach, T., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117.

Dennett, D. (1991). *Consciousness explained*. Boston, MA, USA: Little, Brown and Co.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*(2), 104–123.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2010a). Are there mental lexicons? The role of semantics in lexical decision. *Brain Research*, *1365*, 66–81.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2010b). Are there mental lexicons? The role of semantics in lexical decision. *Brain Research*, *1365*, 66–81.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2010c, December). Are there mental lexicons? The role of semantics in lexical decision. *Brain Research*, *1365*, 66–81.

Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, *5*(1), 61–78.

Ellis, W. (1999). *A source book of gestalt psychology* (Vol. 2). London, UK: Psychology Press.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 211, 179.

Elman, J. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, *33*, 547–582.

Elman, J. (2011). Lexical knowledge without a lexicon? *The Mental Lexicon*, *6:1*, 1-33.

Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). Swift: a dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice

principle. *Text - Interdisciplinary Journal for the Study of Discourse*, *20*(1), 29–62.

Fano, R. M., & Hawkins, D. (1961). Transmission of information: A statistical theory of communications. *American Journal of Physics*, *29*, 793.

Fillenbaum, S., Jones, L., & Rapoport, A. (1963). The predictability of words and their grammatical classes as a function of rate of deletion from a speech transcript1. *Journal of Verbal Learning and Verbal Behavior*, *2*(2), 186–194.

Finch, W. H., Chang, M., Davis, A. S., Holden, J. E., Rothlisberg, B. A., & McIntosh, D. E. (2011). The prediction of intelligence in preschool children using alternative models to regression. *Behavior Research Methods*.

Finn, P. (1977). Word frequency, information theory, and cloze performance: A transfer feature theory of processing in reading. *Reading Research Quarterly*, 508–537.

Fodor, J. (1983). *The modularity of mind* (Vol. 341). Cambridge, MA., USA, USA: MIT press.

Forster, K. I. (1979). Levels of processing and the structure of the language processor. In *Sentence processing: Psycholinguistic studies presented to merrill garrett* (pp. 27–85).

Forster, K. I., & Hector, J. (2002). Cascaded versus noncascaded models of lexical and semantic processing: Theturple effect. *Memory & cognition*, *30*(7), 1106–1117.

Francis, W., & Kucera, H. (1982). *Frequency analysis of english usage*. Boston, MA, USA: Houghton Mifflin Company.

Frank, S. L., & Bod, R. (2011). Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychological Science*, *22*(6), 829 –834.

Frank, S. L., & Vigliocco, G. (in press). Sentence Comprehension as Mental Simulation: An Information-Theoretic Perspective. *Information*.

Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, *59*(2), 127–136.

Gagné, C. L., & Spalding, T. L. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, *60*(1), 20–35.

Geer, D. (2005). Statistical machine translation gains respect. *Computer*, *38*, 18–21.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256–281.

Glenberg, A. (1997). What memory is for. *Behavioral and brain sciences*, *20*(01), 1–19.

Goldberg, A. (2006). *Constructions at work : the nature of generalization in language*. Oxford ;;New York: Oxford University Press.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability

in lexical production. In *Proceedings of cls35* (Vol. 35, pp. 151–166). Chicago, IL, USA: Chicago Linguistic Society.

Griffin, Z., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production* 1,* 2,* 3. *Journal of Memory and Language*, *38*(3), 313–338.

Hahn, L. W., & Sivley, R. M. (2011). Entropy, semantic relatedness and proximity. *Behavior Research Methods*.

Harris, Z. (1951). *Methods in structural linguistics.* Chicago, IL, USA: University of Chicago Press.

Hay, J., Pelucchi, B., Estes, K., & Saffran, J. (2011). Linking sounds to meanings: infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution.* New York: Oxford University Press.

Jackendoff, R. (2007). A parallel architecture perspective on language processing. *Brain Research*, *1146*, 2–22.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37.

Juhasz, B. J., & Berkowitz, R. N. (2011). Effects of morphological families on English compound word recognition: A multitask investigation. *Language and Cognitive Processes*, *26*(4), 653.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguisitics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics.* Cambridge, MA., USA: MIT Press. (Series: Bradford book Bibliography note: Includes bibliographical references (p. [389]-436) and indexes Series: (Bradford book))

Kamide, Y. (2008). Anticipatory processes in sentence processing. *Language and Linguistics Compass*, *2*(4), 647.

Kilgarriff, A. (2005, August). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, *1*(2), 263–276.

Kilgarriff, A., & Grefenstette, G. (2011). Introduction to the special issue on the web as corpus. *Computational Linguistics*, *29*(3), 333–347.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*(1), 12.

Kliegl, R., Risse, S., & Laubrock, J. (2007). Preview benefit and parafoveal-on-foveal effects from word n+2. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(5), 1250.

Kohonen, T., & Somervuo, P. (1998). Self-organizing maps of symbol strings. *Neurocomputing*, *21*(1-3), 19–30.

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day american english*. Dartmouth, NH, USA: Dartmouth Publishing Group.

Kukona, A., Fang, S., Aicher, K., Chen, H., & Magnuson, J. (2011). The time course of anticipatory constraint integration. *Cognition*.

Kuperberg, G. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49.

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, *23*(7), 1089.

Kuperman, V., Bertram, R., & Baayen, R. H. (2010). Processing trade-offs in the reading of dutch derived words. *Journal of Memory and Language*, *62*(2), 83–97.

Kuperman, V., Dambacher, M., Nuthmann, A., & Kliegl, R. (2010). The effect of word position on eye-movements in sentence and paragraph reading. *Quarterly Journal of Experimental Psychology*, 1–20.

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading polymorphemic Dutch compounds: toward a multiple route model of lexical processing. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(3), 876–895.

Kutas, M., & Hillyard, S. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Kwisthout, J., Wareham, T., & Rooij, I. van. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Legge, G., Klitz, T., & Tjan, B. (1997). Mr. chips: an ideal-observer model of reading. *Psychological review*, *104*(3), 524.

Lemke, S., Tremblay, A., & Tucker, B. (2009). Function words of lexical bundles: The relation of frequency and reduction. *The Journal of the Acoustical Society of America*, *125*, 2656.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090.

Li, P. (2009). Lexical organization and competition in first and second languages: Computational and neural mechanisms. *Cognitive science*, *33*(4), 629–664.

Loewenstein, M., Tabor, W., & Tanenhaus, M. K. (1999). Dynamical models of sentence processing - a strongly interactive model of natural language interpretation. *Cognitive Science*, *23*, 491–515.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces

from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, *28*, 203–208.

Macdonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, *32*(5), 692–715.

Matthews, D., & Bannard, C. (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of Child-Directed speech. *Cognitive Science*, *34*(3), 465–488.

McDonald, S., & Shillcock, R. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*(3), 295.

McDonald, S., & Shillcock, R. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*(6), 648.

McEvoy, C., Nelson, D. L., & Komatsu, T. (1999). What is the connection between true and false memories? the differential roles of interitem associations in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1177.

McKenna, M. C. (1986). Cloze procedure as a memory-search process. *Journal of Educational Psychology*, *78*, 433 - 440.

Mirman, D., Graf Estes, K., & Magnuson, J. (2010). Computational modeling of statistical learning: Effects of transitional probability versus frequency and links to word learning. *Infancy*, *15*(5), 471–486.

Misyak, J., Christiansen, M., & Tomblin, B. J. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, *2*(1), 138–153.

Mitchell, D., Cuetos, F., Corley, M., & Brysbaert, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research*, *24*(6), 469–488.

Moyer, R. S., & Dumais, S. T. (1978). Mental comparison. *The Psychology of Learning & Motivation: Advances in Research & Theory*, *12*, 117.

Nelson, D. L., & McEvoy, C. (2007). Entangled associative structures and context. In *Proceedings of the aaai spring symposium on quantum interaction.* Palo Alto, CA, USA: AAAI Press.

Nelson, D. L., McEvoy, C., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition*, *28*(6), 887–899.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The university of south florida word association, rhyme, and word fragment norms.* (http://www.usf.edu/FreeAssociation/)

Nelson, D. L., McKinney, V., Gee, N., & Janczura, G. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review*, *105*(2), 299.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA., USA: Harvard University Press.

Newmeyer, F. (1996). *Generative linguistics : a historical perspective*. London ;;New York: Routledge.

Norris, D., & Kinoshita, S. (2008). Perception as evidence accumulation and bayesian inference: Insights from masked priming. *Journal of Experimental Psychology: General*, *137*(3), 434–455.

Osgood, C. E., Sebeok, T. A., Gardner, J., Carroll, J., Newmark, L., Ervin, S., et al. (1954). Psycholinguistics: a survey of theory and research problems. [References]. *Journal of Abnormal and Social Psychology*.

Owens, M., O'Boyle, P., McMahon, J., Ming, J., & Smith, F. (1997). A comparison of human and statistical language model performance using missing-word tests. *Language and Speech*, *40*(4), 377.

Perfetti, C. A. (1992). *The representation problem in reading acquisition.* Mahwah, NJ, US: Lawrence Erlbaum Associates, Inc.

Perfetti, C. A., Hart, L., Verhoeven, L., Elbro, C., & Reitsma, P. (2002). The lexical quality hypothesis. *Precursors of functional literacy*, *11*, 67–86.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011, Mar). Word lengths are optimized for efficient communication. *Proc Natl Acad Sci U S A*, *108*(9), 3526-9.

Pickering, M., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110.

Pinheiro, J. C., & Bates, D. M. (2009). *Mixed-effects models in S and S-PLUS*. New York, NY, USA: Springer Verlag.

Pinker, S., & Ullman, M. T. (2002, November). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456–463.

Pluymaekers, M., Ernestus, M., & Baayen, R. (2005b). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*(2-4), 146–159.

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005a). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*(2-4), 146–159.

Prior, A., & Bentin, S. (2003). Incidental formation of episodic associations: The importance of sentential context. *Memory & Cognition*, *31*(2), 306–316.

Prior, A., & Bentin, S. (2008). Word associations are formed incidentally during sentential semantic integration. *Acta Psychologica*, *127*(1), 57–71.

R Development Core Team. (2009). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raffone, A., & Leeuwen, C. van. (2003). Dynamic synchronization and chaos in an associative neural network with multiple active memories. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *13*, 1090.

Rayner, K. (2009). Eye Movements in Reading: Models and Data. *Journal of eye movement research*, *2*(5), 1–10.

Rayner, K., Inhoff, A. W., Morrison, R. E., Slowiaczek, M. L., & Bertera, J. H. (1981). Masking of foveal and parafoveal vision during eye fixations in reading. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(1), 167–179.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading.* Mahwah, NJ, US: Lawrence Erlbaum.

Recchia, G., & Jones, M. (2009). More data trumps smarter algorithms: comparing pointwise mutual information with latent semantic analysis. *Behavior research methods*, *41*(3), 647–656.

Remillard, G. (2010). Implicit learning of fifth- and sixth-order sequential probabilities. *Memory & Cognition*, *38*(7), 905–915.

Roberts, M. A. J., & Chater, N. (2008). Using statistical smoothing to estimate the psycholinguistic acceptability of novel phrases. *Behavior Research Methods*, *40*(1), 84–93.

Rodriguez, P. (2003). Comparing simple recurrent networks and n-Grams in a large corpus. *Applied Intelligence*, *19*(1), 39–50.

Roediger, H., & McDermott, K. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803.

Rosen, V., & Engle, R. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General*, *126*(3), 211.

Ruff, R., Light, R., Parker, S., & Levin, H. (1997). The psychological construct of word fluency. *Brain and Language*, *57*(3), 394–405.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926 – 1928.

Schwanenflugel, P., & LaCount, K. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 344.

Seidenberg, M., & McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell System Technical Journal*, *30*(1), 50–64.

Shaoul, C., & Westbury, C. (2008). *HiDEx: the high dimensional explorer.* Edmonton, AB. (Published: Downloaded from http://www.psych.ualberta.ca/˜westburylab/downloads.html)

Shaoul, C., & Westbury, C. (2011). Formulaic sequences: Do they exist and do they matter? *Methodological and Analytic Frontiers in Lexical Research (Part II). Special Issue of The Mental Lexicon*, *6*(1), 171-196.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Sibley, D. E., Kello, C. T., Plaut, D. C., & Elman, J. L. (2008). Large-Scale modeling of wordform learning and representation. *Cognitive Science*, *32*(4), 741–754.

Siyanova-Chanturia, A., Conklin, K., & Heuven, W. van. (2011). Seeing a Phrase "Time and Again" Matters: The Role of Phrasal Frequency in the Processing of Multiword Sequences. *Journal of Experimental Psychology: Learning Memory and Cognition*, *37*(3), 776–784.

Smith, N., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the 33rd annual meeting of the cognitive science conference.*

Speelman, C. (2005). *Beyond the learning curve : skill acquisition and the construction of mind*. Oxford: Oxford University Press.

Squire, L., & Kandel, E. (2000). *Memory: From mind to molecules*. New York, NY, USA: Holt.

Stanovich, K. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York, NY, USA: The Guilford Press.

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition*, *7*(4), 263–272.

Taylor, W. (1953). " cloze procedure": a new tool for measuring readability. *Journalism quarterly*.

Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279.

Thompson, G. L., & Desrochers, A. (2009). Corroborating biased indicators: Global and local agreement among objective and subjective estimates of printed word frequency. *Behavior Research Methods*, *41*(2), 452–471.

Toglia, M. P. (2009). Withstanding the test of time: The 1978 semantic word norms. *Behavior Research Methods*, *41*(2), 531–533.

Tomasello, M. (2003). *Constructing a language : a usage-based theory of language acquisition*. Cambridge Mass.: Harvard University Press.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and erp study of the effects of structure, frequency, and probability on immediate free recall. *Perspectives on formulaic language: Acquisition and communication*, 151–173.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*(2), 569–613.

Tremblay, A., & Tucker, B. V. (2011). The effects of N-gram probabilistic

measures on the recognition and production of four-word sequences. *The Mental Lexicon*, *6*(2), 302–324.

Troyer, A. (2000). Normative data for clustering and switching on verbal fluency tasks. *Journal of Clinical and Experimental Neuropsychology*, *22*, 370–378.

Tulving, E. (1985). How many memory systems are there?. *American Psychologist*, *40*(4), 385.

Ullman, M. T. (2001). The Declarative-Procedural model of lexicon and grammar. *Journal of Psycholinguistic Research*, *30*(1), 37–69.

Ullman, M. T., Miranda, R., & Travers, M. (2008). Sex differences in the neurocognition of language. In J. B. Becker, K. J. Berkley, & N. Gearyet (Eds.), *Sex on the brain: From genes to behavior* (p. 291-309). NY, NY, USA: Oxford University Press.

Unsworth, N., Spillers, G., & Brewer, G. (2010). Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. *The Quarterly Journal of Experimental Psychology*, *64*(3), 447–466.

Van Berkum, J. (2008). Understanding sentences in context. *Current Directions in Psychological Science*, *17*(6), 376.

Vitu, F., & McConkie, G. W. (2000). Regressive saccades and word perception in adult reading. *Reading as a perceptual process*, 301–326.

Westbury, C. (2007). *ACTUATE: Assessing Cases, The University of Alberta Testing Environment.* (Downloaded from http://www.psych.ualberta.ca/˜westburylab/)

Willems, R., & Hagoort, P. (2007). Neural evidence for the interplay between language, gesture, and action: A review. *Brain and Language*, *101*(3), 278–289.

Wood, S. (2006). *Generalized additive models: an introduction with r* (Vol. 66). New York, NY, USA: CRC Press.

Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language & Communication*, *18*(1), 47–67.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*(4), 502–529.

Zwaan, R. (2008). Experiential traces and mental simulations in language comprehension. *Symbols, embodiment, and meaning*, 165–180.