Assessing the impact of co-occurrence frequency and diversity in statistical learning accounts of language processing

RUSSELL TURK

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

September 2019

Abstract

Language is heralded as one of the foremost human achievements and is vital in scaffolding the successful development of many other skills. Yet, the mechanism by which language is acquired is still poorly understood. One possible account is Statistical Learning Theory, an explanation of language acquisition that has grown in popularity over the past three decades. The central tenet of Statistical Learning Theory is that learners are guided by statistical regularities in their environment and can utilise these to develop an implicit understanding of their natural language. Current theory holds that transitional probabilities are the best predictor of learner performance in statistical learning tasks. However, little has been done to investigate alternative statistical measures. This thesis presents two such metrics: Bigram frequency and bigram diversity and contrasts them with transitional probability in predicting task performance. Through the repurposing of primed lexical decision and sequence learning tasks, I present a novel approach to examining the impact of statistical priming on task performance in a naturalistic dataset. Model comparison using Bayesian multilevel modelling suggests that transitional probability is not as reliable a predictor as was previously believed. Moreover, I demonstrate that bigram frequency may represent a better metric for predicting task performance in these tasks. The current work highlights the importance of considering alternative metrics of statistical regularity when describing the underlying mechanisms of language acquisition and showcases alternative methods of examining statistical learning performance.

Contents

1 INTRODUCTION

Language is an essential aspect of human culture and interaction. It allows for the efficient exchange of knowledge and scaffolds the acquisition of many key skills. The ability to acquire language is therefore one of the most important that humans develop; yet, the precise mechanisms of language acquisition have not been identified. One interesting hypothesis is that humans are attuned to the statistical distributions of their environment and that these allow them to make sense of the constant sensory barrage they are exposed to daily.

In Chapter 2 I give an overview of the statistical learning literature as it applies to language before discussing the major criticisms that can be levelled at the paradigm. Following this, I introduce transitional probability - the dominant metric of statistical regularity within the statistical learning literature - and consider the strengths and weaknesses of using this statistic. I then suggest two new frequency-based measures of statistical distribution and contrast these with transitional probability, making a case for a less cognitively effortful mechanism of statistical learning. Finally, I highlight some of the challenges of working with naturalistic stimulus-sets and propose methods of overcoming these obstacles.

Building on these ideas, Chapters 3 and 4 detail four lexical decision experiments as a proof of concept for assessing previously learnt statistical associations in natural language. By working with adults, and utilising an existing corpus of language, I manage to maintain the statistical properties of language whilst foregoing the need for extensive familiarisation phases. These are further developed in Chapters 5 and 6 where I address some of the potential

methodological limitations in the experiments and build on the arguments for a frequency-based mechanism of language acquisition. Each set of experiments is presented as a pair in which the first examines the impact of bigram frequency and the second bigram diversity. Although information for both metrics is calculated in all the experiments, the initial design choice was to examine and test them individually and, as such, I have retained this format throughout. In Chapter 7, however, I aggregate data from all the experiments and investigate the impact of both bigram frequency and diversity in a meta-analysis. Finally, I present two novel sequence learning tasks to test the acquisition of new information before discussing the implications for statistical learning research as it applies to language acquisition.

1.1 REPRODUCIBILITY AND CODE

This thesis was written to be entirely reproducible. All the analyses in this work have been conducted in R (R Core Team, 2019) and all the code needed to reproduce the analyses and results have been included in blocks like the one below:

```
rstan_options(auto_write = TRUE)

options(mc.cores = parallel::detectCores())
```

All code can be copied from this document into the main R console and, in doing so, it is possible to recreate/verify the findings presented herein. All data used in the analyses can be downloaded from GitHub using the following URL:

https://github.com/russellturk/Thesis_Data; simply set your working directory (in R) as the folder containing the data and you should be able to run the code with no problems.

Although the code is included there is no requirement to reproduce any of the analyses and doing so is not necessary in order to understand this work. Finally, though it is possible to replicate all the findings - and I certainly invite you to do so, if you wish - some of the larger models can have quite long runtimes, so some discretion is advised. Running the code presented above can help the models compile more quickly if you have multiple cores available but the overall runtime will still depend on the size of the model.

2 LITERATURE REVIEW

2.1 INTRODUCTION

Language has been described as a hallmark of the human species (Christiansen & Kirby, 2003) and a defining part of our social identity (Nowak, Komarova, & Niyogi, 2002). Children begin acquiring, and reproducing, language sounds from a very young age and do so with a regularity that transcends cultural boundaries (Kuhl, 2004); furthermore, they demonstrate an implicit knowledge of language structure long before they can express a formal understanding of syntactical and grammatical rules - language, it seems, is ubiquitous, universal, and quintessentially human. As such, mastery of their natural language is a central part of every child's development. How then do humans acquire this mastery at such a young age? Language is a complex, multifaceted construct which, formally at least, is poorly understood by many speakers - even those born into a language may struggle to articulate its myriad rules with anything approaching clarity. Despite this lack of formal understanding however, humans regularly produce utterances (mostly) in accordance with the rules of their natural language and can recognise even minor violations to these rules, for example, I goed to work or can you explain me it would be easily identified as incorrect by the average English speaker even if they were unable to explain which of the formal rules had been violated. However, a similar sense of wrongness would also be elicited by the phrases costs a leg and an arm and when you come to it, cross that bridge despite them being perfectly acceptable grammatical constructs (Widdowson, 1989). This suggests that learners may

not rely on formal rule-based systems of language - though these may be acquired through explicit instruction - since the latter two examples are only violations of the commonly accepted word-order rather than of grammaticality.

It is therefore more likely that language is acquired implicitly; the naive linguist being exposed to their native language(s) can derive meaning and structure from the endless streams of speech and/or text. This is, I believe, effectively illustrated by the following quotes:

> *Language is my mother, my father, my husband, my brother, my sister, my whore, my mistress, my checkout girl. Language is a complimentary moist lemon-scented cleansing square or handy freshen-up wipette. Language is the breath of God; language is the dew on a fresh apple, it's the soft rain of dust that falls into a shaft of morning light as you clutch from an old bookshelf a half-forgotten book of erotic memoires; language is a creak on the stair, a spluttering match held to a frosted pane; it's a half-remembered childhood birthday party, the warm wet, trusting touch of a leaking nappy, the hulk of a charred Panzer, the underside of a granite boulder, the first downy growth on the upper lip of a Mediterranean girl, its cobwebs long since overrun by an old Wellington boot. (Stephen Fry, A bit of Fry & Laurie)*

> *We open our mouths and out flow words whose ancestries we do not even know. We are walking lexicons. In a single sentence of idle chatter, we preserve Latin, Anglo-Saxon, Norse; we carry a museum inside our heads, each day we commemorate people of whom we have never heard. More than that, we speak volumes - our language is the language of everything we have read. Shakespeare and the Authorised Version surface in supermarkets, on buses, chatter on radio and television. (Penelope Lively, Moon Tiger)*

Language is present in the majority of human interaction; it provides the framework within which we interact and the means by which we express those interactions. It has been suggested that language shapes the way we think and how we interpret our environment (Reines & Prinz, 2009; Whorf, 1956). It is a fundamental part of our experiences and, as such, provides a wealth of exposure through which it is possible to gauge the underlying patterns and structures required for effective communication.

Yet, despite the universality of language, the underlying mechanisms are still poorly understood. Take the parsing of complex speech streams into individual lexical items - an important aspect of vocabulary acquisition that can be performed by infants as young as five and a half months (Johnson & Tyler, 2010). Unlike written language, infants experience speech streams with no uniform pauses or 'white spaces' to indicate word boundaries (Cole & Jakimik, 1980). This is most apparent when listening to an unfamiliar language where, rather than words, we hear a continuous stream of sound.

It is therefore essential to develop a strategy that enables the identification of lexical boundaries within larger linguistic structures. One possibility is that humans adopt a strategy of learning words in isolation before applying them to longer speech streams (Nemko, 1984). Such a strategy is useful for identifying object-labelling words - where a concrete target exists, and can be referred to independently of a wider context - but fails to account for how very young children can rapidly learn to recognise novel words when no referent is available or how they can extract this meaning from within sentences (Saffran,

Aslin, & Newport, 1996). Additionally, Aslin, Woodward, LaMendola, and Bever, (1996) demonstrated that when asked to teach specific words to their children many mothers did not present them in isolation, with the majority opting to place the target word at the end of longer utterances - such as *dog* in the phrase "Look at the dog.". The same mothers also added emphatic stress in order to draw attention to the target words suggesting that prosodic cues may play a part in infant word segmentation. However, prosodic preferences have been demonstrated to be language specific and are therefore unlikely to constitute a universal explanation of word segmentation (Höhle, Bijeljac-Babic, Herold, Weissenborn, & Nazzi, 2009; Kooijman, Hagoort, & Cutler, 2009).

In fact, any explicit strategy of word-learning must contend with what Quine (1960) described as the indeterminacy problem. The learner has no way of inferring the meaning of a vocalisation without access to shared contextual information - which cannot be assumed to be possessed by infants. Take, for instance, the example of a parent vocalising the word dog whilst pointing to the self-same canine; to the proficient English speaker the referent is obvious due to a prepossession of the concept of 'dog-ness'. To the naive observer the inference is not so straightforward; perhaps the speaker is referring to some part or feature of the dog, this particular dog, four-legged mammals more generally, or even some function fulfilled by the dog (e.g., pet or friend). Behaviourally there is no way of identifying which, if any, of these interpretations is correct, and multiple encounters with the same referent do little to disambiguate word and meaning. It is therefore unlikely that language is learnt by pairing isolated words with their physical counterparts, particularly since many words do not

directly relate to concrete examples in the environment. How then do infants successfully acquire language?

An explanation that has gained traction in recent years is Statistical learning theory. The concept of Statistical learning can be traced back to Miller and Selfridge (1950) who identified that the statistical relationships between words in common usage correlated with participants' memory for wordlists. In the latter part of the 20th century Statistical learning theory experienced something of a resurgence when Saffran et al. (1996) published their seminal study suggesting infants can track the transitional probabilities of syllables in an artificial language; since then these findings have been replicated with both children and adults using a number of different paradigms (Aslin, Saffran, & Newport, 1998; Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Johnson & Tyler, 2010; Saffran et al., 1996; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997).

Given the evidence, it is uncontroversial to describe humans as being adept at recognising patterns within the environment; in fact, pattern recognition is one of the few remaining domains where humans outperform computers in terms of accuracy (Jain, Duin, & Mao, 2000; Schur & Tappert, 2016). It has been proposed that humans become attuned to, and can track, the statistical patterns in their natural language(s) and use this information to build up a lexical and grammatical repertoire to aid in the production and comprehension of novel linguistic structures. This process eliminates the need for target-referent tracking in language acquisition, since the target does not necessarily have to be present for learning to occur (though referential information may still provide

semantic benefits) and allows for the acquisition of abstract concepts such as love, or hate, for which there may be no immediate environmental reference. The underlying premise is that, by tracking statistical regularities within the environment, information can be extracted and implicitly applied to the generation and recognition of novel data. This is the central tenet of statistical learning - that learning occurs with no conscious effort - specifically, that learners can become attuned to the statistical regularities in their environment. With regards to language, this information can be the relationship between symbols and sounds, the ordering of individual speech sounds into units of meaning, or words into sentences that can be used to build up a lexical and grammatical repertoire to aid in the production and comprehension of novel linguistic structures. Furthermore, acceptable grammatical structures can be iteratively modelled through interaction with, and imitation of, expert language users.

As such statistical learning is, at its core, a powerful mechanism for the acquisition of patterns from external data. This ability has been the subject of extensive research over the last couple of decades and has been demonstrated across a number of different modalities including shape (e.g., Kirkham, Slemmer, & Johnson, 2002), music (Daikoku, Yatomi, & Yumoto, 2014; Koelsch, Busch, Jentschke, & Rohrmeier, 2016; Liu & Kager, 2011; Hay, Pelucchi, Estes, & Saffran, 2011; Saffran, Johnson, Aslin, & Newport, 1999), tactile stimuli (Conway & Christiansen, 2005) and, most pertinently, psycholinguistics where studies have demonstrated that learners are capable of using distributional statistics for a number of complex language-related tasks including word segmentation and

sentence parsing (Saffran et al., 1996; Thiessen & Erickson, 2013; Toro, Sinnett, & Soto-Faraco, 2005; Vouloumanos, 2008); the acquisition of vocabulary and lexical information (Goodman, Dale, & Li, 2008; Harris, Barrett, Jones, & Brookes, 1988; Schwartz & Terrell, 1983); and the discrimination of grammatical structures (Reeder, Newport, & Aslin, 2017; Theakston, Lieven, Pine, & Rowland, 2004).

2.2 LANGUAGE

Statistical learning theory has been used extensively in the study of language, particularly regarding word segmentation. This is an essential early task whereby infants need to extract meaningful units from continuous speech - a process made more difficult by the lack of pauses between words. Probably the most prominent example of statistical learning in word segmentation is provided by Saffran, Aslin, and Newport (1996). This study was instrumental in kickstarting statistical learning theory and defining the common methodologies employed in its investigation. Saffran and colleagues generated a mini-language of four tri-syllabic nonsense words (bidaku, padoti, golabu, and tupiro) made up of twelve consonant-vowel pairs (syllables). These were then pseudo-randomly concatenated into speech streams lasting two-minutes and consisting of six hundred tokens (The randomisation was constrained so that no word immediately followed itself in any speech stream). The stimuli were formulated using a text-to-speech synthesiser to remove all boundary information except for distributional information. Thus, the only cue available for the segmentation of the speech stream was the statistical disparity in intra- and inter-word transitions. For instance, the intra-word transition from bi to da or from la to bu

are characterised by probabilities of 1.0 - that is, the two syllables only ever occur together in the specified order - whereas inter-word transitions varied between .25 and .33. Using the head-turn preference procedure (Fernald, 1985) infants were assessed on their ability to differentiate between the familiar stimuli and part-word stimuli generated by combining two syllables of the familiar words with one syllable of an adjacent word in the speech stream (e.g., dotigo or labutu); these part-words therefore violate the statistical structure of the language by having internal transitions that are not equal to 1.0. Preferential listening was then measured for stimuli presented to either side of the infant, with longer listening times taken as an indication of preference. They found that infants showed a preference for the part-words which was interpreted as preference for novelty – implying that the infants had learnt something about the language and were now 'familiar' with it. They therefore concluded that infants must be capable of extracting words from longer utterances based on statistical cues.

This ability must be learned since it cannot be assumed that individuals are born with an innate knowledge of the statistical regularities of their natural language. To borrow an example from Saffran (2003), pretty and baby are both words which exist in English, but ttyba (which spans the boundary between pretty and baby) is not. Saffran suggests that infants utilise the statistical structure of language in their environment to inform their discovery of word boundaries in fluent speech. In English, the syllable pre can only be followed by a relatively small set of syllables, including tty, tend, and cedes; in natural, infant directed, speech (to infants) pre is succeeded by tty roughly 80% of the time.

However, since tty occurs at the end of a word, it can potentially be followed by any syllable that can be used to initiate an English word. The chance that tty is followed by ba, as in pretty baby, is therefore much lower (roughly 0.03%). This disparity is considered indicative that pretty is an English word, and ttyba is not. Individuals can therefore use these cues to discern the likelihood of two or more syllables constituting a word in their natural language. This is echoed by Perruchet and Pacton (2006) who suggest that learning may involve preferentially selecting chunks of sound that occur with high probability and recognising them as individual word-units. These word-units can form standalone words or be combined to produce structures that are more complex. Since pre and tty occur with a relatively high frequency they will be chunked together as a single unit which can then be used as a reference for parsing novel speech; allowing infants to build up a lexicon of statistically related word-units.

This mechanism has proven to be fairly robust and, as previously noted, it has been suggested that children as young as five and a half months possess the ability to extract individual words from continuous speech using little more than the statistical regularities of the language (Johnson & Tyler, 2010); furthermore, infants begin to demonstrate an awareness of these regularities after very short exposure times (Saffran, Aslin & Newport, 1996). Johnson & Tyler (2010) replicated the findings of Saffran and colleagues by testing infants' segmentation ability after two and a half minutes of exposure to a four-word artificial language and found that infants showed a preference for words compared to cross-boundary part-words. It can be argued that this demonstrates an implicit use of the statistical structure of language which cannot be explained by rule-

based learning. In fact, statistical learning paradigms have been used to demonstrate that learning can take place in isolation from both context and grammar (Jusczyk & Aslin, 1995; Saffran, Aslin & Newport, 1996; Saffran, Newport & Aslin, 1996; Saffran, Newport, Aslin, Tunick & Barrueco, 1997; Aslin, Saffran, Newport, 1998; Pelucchi, Hay, & Saffran, 2009) and that participants ranging from very young children (e.g., Johnson & Tyler, 2010; Pelucchi, Hay, & Saffran, 2009) to adults (e.g., Koelsh, Busch, Jentschke, & Rohrmeier; 2016; Saffran, Johnson, & Aslin, 1999) are capable of tracking the distributional properties of a language even when encoding and testing are temporally separated (e.g., Durrant, Taylor, Cairney, & Lewis, 2011; Kim, Seitz, Feenstra, & Shams, 2009). In fact, children and adults show remarkably similar statistical learning ability; Saffran, Johnson, Aslin, & Newport (1999) tested adults and eight-month-old infants using sequences of tones (based on the original language in Saffran, Aslin, & Newport, 1996) and found that both groups were capable of discriminating between familiar and novel sequences – It should be noted, however, that adults are generally tested on their familiarity with the stimuli through the use of alternative-forced-choice tasks whereas infant studies focus on preferential looking (or listening) times.

However, linguistic development is more than just the extraction of words from speech. Gómez and Gerken (1999) exposed infants to two artificial grammars consisting of five CVC words (JIC, PEL, RUD, TAM, and VOT). Both grammars produced utterances beginning and ending with the same word but differing on the order of internal word-pairs. Infants were trained on one of the grammars and then tested on their ability to discriminate between unfamiliar utterances drawn from the training grammar and utterances from the alternate grammar

and a marked preference was observed for the familiar grammar. This is also demonstrated by Chambers, Onishi, & Fisher (2003; see also Chambers, Onishi, & Fisher, 2010; Dell, Reed, Adams, & Meyer, 2000; Goldrick, 2004; Goldrick & Larson, 2008; Onishi, Chambers, & Fisher, 2002; Seidl, Cristià, Bernard, & Onishi, 2009; Warker, Dell, Whalen, & Gereg, 2008; Warker, Xu, Dell, & Fisher, 2009) who used a similar procedure to demonstrate that participants can learn phonotactic regularities during familiarisation and can apply these to novel stimuli during testing. As with the aforementioned studies, infants are assumed to be attending to the distributional statistics of the grammar in the absence of other cues. Since there was no overlap in utterances between familiarisation and testing, it can be inferred that the preference for the trained grammar cannot be attributed to memory for the previously encountered utterances. This suggests that infants are capable not only of extracting words from speech but can also begin to build-up rules relating to word order and higher-level grammatical structures. To further illustrate, both adults and children have been shown to adopt familiar patterns in their own utterance production (Bock, 1986; Pickering & Ferreira, 2008), matching the distributional patterns of experienced language to their own speech. This structural priming has been demonstrated to be independent of both vocabulary and context (Bock, 1989; Bock & Loebell, 1990) and thus cannot be attributed to mimicry of existing utterances. Impressively, this ability is robust enough that infants less than twelve-months old can make grammatical generalisations when only 83% of the familiarisation strings conform to underlying statistical structure of the language (Gómez & Lakusta, 2004).

Furthermore, evidence from statistical learning paradigms has identified a potential link between performance on statistical learning tasks and the processing and comprehension of natural language (Conway, Bauernschmidt, Huang, & Pisoni, 2010; Misyak & Christiansen, 2012). This is particularly true of participants who display atypical language development; these participants tend to perform poorly on statistical learning tasks and struggle to generalise between the familiarisation and testing phases (Plante, Gómez, & Gerken, 2002; Grunow, Spalding, Gómez, & Plante, 2006; Richardson, Harris, Plante, & Gerken, 2006; Tomblin, Mainela-Arnold, & Zhang, 2007). This deficit is characterised by the need for longer exposure times; for example, Evans, Saffran, and Robe-Torres (2009) investigated whether children with Specific Language Impairment varied in their ability to discriminate between familiar and unfamiliar pseudowords using a two-alternative forced-choice paradigm. They demonstrated that after twenty-one minutes of exposure typically developing children perform significantly better than chance whereas children with specific language impairment do not. However, after forty-two minutes of exposure both the typical and SLI children were able to perform the task at better than chance. Additionally, Riches, Tomasello, and Conti-Ramsden (2005) investigated the effect of increased frequency of presentation and demonstrated that children with SLI performed better on a verb comprehension test when the number of exposures was increased. The same pattern of results was not present in typically developing matches however, suggesting that children with SLI may possess a less efficient statistical learning mechanism which requires a greater number of presentations in order to achieve comparable levels of learning. Interestingly, this deficit appears to transcend modalities; Tomblin et al. (2007)

were able to demonstrate that reduced performance on a pattern learning task was strongly associated with grammatical difficulties whereas Conway, Pisoni, Anaya, Karpicke, and Henning (2011) observed that visual sequence learning correlates with language outcomes for children with cochlear implants. It has therefore been suggested that certain language impairments may arise from a general deficit in statistical learning (Hsu & Bishop, 2010) and that this leads to slower learning of statistical regularities. It has therefore been suggested that certain language impairments may arise from a general deficit in statistical learning (Hsu & Bishop, 2010).

2.3 **CRITICISMS**

There have been several criticisms relating to the validity of the early statistical learning literature. Endress and Mehler (2009b) demonstrated that learners could not reliably segment word-units from continuous speech, and that they were just as likely to identify novel word-units as familiar providing they had the same statistical structure as the target items. This suggests that participants were able to learn the statistical nature of the language but were unable to effectively match this to the phonemic properties of the word. They claim that co-occurrence statistics alone are insufficient for the segmentation of words in spoken language - though they do note that may not be the case for written stimuli - and that prosodic cues may be necessary to delineate word boundaries. Moreover, they claim that early statistical learning experiments (e.g. Fiser & Aslin, 2002; Hauser, Newport, & Aslin, 2001; Saffran, Johnson, Aslin, & Newport,1999; Toro & Trobalón, 2005; Turk-Browne, Jungé, & Scholl, 2005) did

not adequately demonstrate word segmentation; rather, they indicate an ability to discriminate between familiar and unfamiliar statistical structures across word sub-units and only with the inclusion of non-statistical cues are learners able to extract complete word-units (see also, Endress & Mehler, 2009a). Since natural language contains a wealth of information - including, but not limited to, prosodic cues, onset stress, and phonotactic regularity - that is not present in artificial languages, it is unsurprising that these cues would contribute to word segmentation.

This is somewhat echoed by Johnson and Tyler (2010) who suggest that distributional statistics may simply act as a stepping-stone to learning language-specific segmentation cues (see also, Saffran, Werker, & Werner, 2006). This would allow infants to start building a representation of their natural language prior to gaining an understanding of the phonotactic properties of the language. Additionally, they claim that the languages used in these early studies lack the complexity of naturalistic speech and that this may aid in word-segmentation through the introduction of additional regularities; specifically, the majority of studies utilise fixed word-lengths of two or three syllables (e.g., Johnson & Jusczyk, 2001; Saffran, Aslin, & Newport, 1996; Thiessen & Saffran, 2003). Upon varying the length of targets within the stimulus-set, it was demonstrated that infants were less successful at the segmentation task when word-length is held constant. This throws into question the ability of statistical learning theory to scale-up to naturalistic settings (van Heugten & Johnson, 2010). To address the discrepancy between natural and artificial languages Pelucchi, Hay, and Saffran (2009) introduced

(non-Italian) infants to a subset of words from Italian to capture "virtually all of the complexity of natural language" (p. 3). This subset comprised four target words (fuga, melo, pane, and tema) embedded in grammatically correct Italian sentences (e.g., La zia Carola si è esibita in una fuga colla bici verde). They demonstrated that infants were still able to discriminate between familiar and novel Italian words despite the added complexity of the stimulus. However, the target words in this study were characterised by internal transitional probabilities of 1.00 - so, for example, fu and ga only ever occur together in the familiarisation phase. This is an extreme example of co-occurrence that rarely appears in natural language and, in contrast to the more realistic transitional probabilities found in the non-target stimuli, may have provided additional cues to learning.

The second major criticism of early statistical learning studies is that they fail to accurately represent the statistical distributions found in natural languages. This is, broadly speaking, an artefact of the time constraints inherent to studies of learning. More complex languages necessarily require longer familiarisation periods than may be practical in the majority of experimental research Erickson and Thiessen (2015). This is a particular problem for studies involving infants and/or young children whose attentional capabilities are limited (e.g., McCall & Kagan, 1970). This often leads to mini-languages comprising four to six words with perfect within-stimulus transitions and unrealistic cross-boundary statistics.

Frank, Goldwater, Griffiths, & Tenenbaum (2010) noted that the artificial languages used in previous statistical learning tasks have been relatively limited

and that this may contribute to learning by inflating the statistical relationships between syllables. It has been suggested that increasing the number of unique syllables (and thus the number of possible words) will increase the difficulty of the word segmentation task. By using three, four, five, six, and nine-word languages they demonstrated a negative relationship between language complexity and segmentation efficiency after two and a half minutes of familiarisation. To highlight this disparity, in their seminal study, Saffran, Newport, & Aslin, 1996) report inter-syllable and cross-boundary transitional probabilities of 1.0 and less than .33 respectively (Transitional probabilities are discussed in more detail below). In comparison, naturally occurring transitional probabilities are often considerably lower. If we consider the common English bigrams dog food or bank holiday, we see transitional probabilities of .02 and .03, several orders of magnitude smaller than those reported by Saffran and colleagues. Figure 2.1 shows the distribution of transitional probabilities for all bigrams in the British National Corpus with a frequency of greater than ten; it is apparent that the transitional probabilities in these studies do not adequately reflect those found in a large, naturalistic dataset.

```
library(ggplot2)


df <- read.csv("trans_prob_illustration.csv")


ggplot(df, aes(trans_prob)) + geom_density() + xlim(0, 1) +  ylim(0,
    40) + theme_minimal()
```

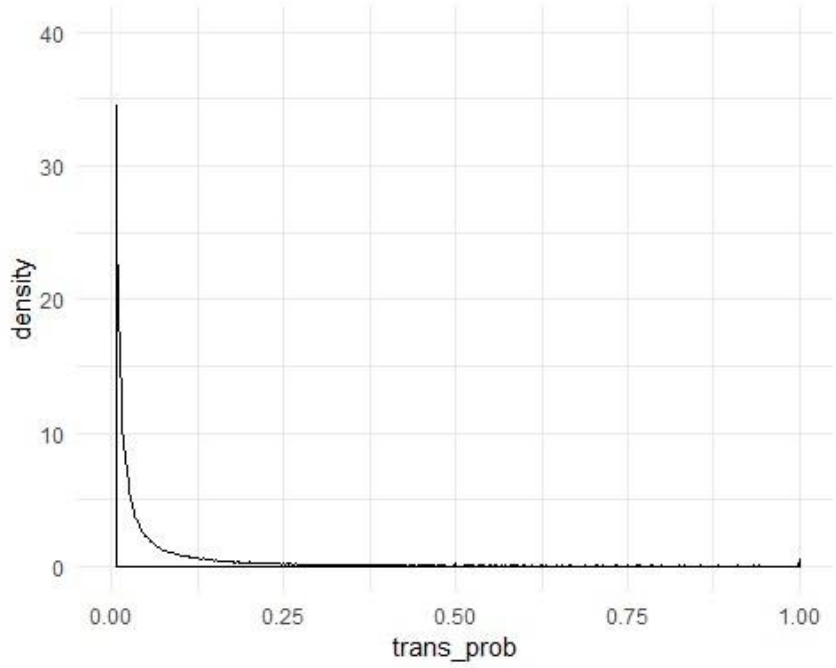

Figure 2.1: Density plot showing the distribution of transitional probabilities for all bigrams in the British National Corpus with a co-occurrence frequency greater than ten. Density, shown on the y-axis, indicates the proportion of bigrams with a given transitional probability, as shown on the x-axis. Most bigrams have a transitional probability of less than .10 illustrating that the distributional statistics used in existing statistical learning paradigms do not accurately represent those found in natural language.

The intention of this work is not to undervalue the contribution of these early studies to the understanding of how infants might begin to parse words from continuous speech streams. Indeed, the simplification of natural language is undeniably necessary if we are to make causal claims as to role of distributional statistics. However, it could be argued that by sanitising the input learners are exposed to it no more represents the language experience than do non-linguistic sequences (such as shapes or tones). Therefore, if we are to build upon the foundations laid by these early studies, it is necessary to investigate the

identified phenomena in more linguistically rich stimulus-sets. This has been somewhat addressed in subsequent investigations (e.g., Frank et al., 2010) but not to the extent of using natural language corpora as experimental stimulus-sets (something that I shall discuss in more detail below).

Words in real languages have a more flexible statistical structure than those seen in experimental languages; this leads to richer and more varied word composition in which elements (e.g., phonemes, graphemes, or syllables) can be repeated. This is not the case with most of the mini languages developed for statistical learning paradigms. For example, in Saffran, Aslin, and Newport (1996) words are generated by concatenating three of twelve unique syllables which may have led to more predictable word boundaries (since the onset of a repeated syllable necessarily indicates a new word). Furthermore, the increased statistical flexibility exhibited by natural languages means that the difference between within- and between-word transitional probabilities is likely to be less pronounced than those seen in experimental languages. Indeed, some words may even include internal transitions with a lower probability than those seen at word-boundaries. It has, in fact, been suggested that although a reductionist approach to statistical learning is the norm in experimental paradigms it may, paradoxically, prove detrimental to language learning more generally. For example, Kidd, Piantadosi, & Aslin (2012) describe how infants prefer stimuli that are neither too complex nor too simple suggesting that a certain amount of complexity may aid in statistical learning. Similarly, Gerken, Wilson, & Lewis (2005) demonstrated that children can use distributional statistics to identify grammatical gender only when there are additional statistical cues to category

membership. This implies that the presence of redundant distributional information may scaffold the learning of grammatical rules. This phenomenon is also seen in infant learning of musical structures (Thiessen & Saffran, 2009) where children under eight-months old learnt either the lyrics or the melodies of a musical piece more efficiently when they were presented together rather than as distinct components. This may be due to overlapping sources of information serving to reinforce otherwise ambiguous associations (Thiessen & Erikson, 2015) which provides a benefit that outweighs the increased cost of processing the additional information (Teinonen, Aslin, Alku, & Csibra, 2008). Furthermore, infants have the ability to learn non-adjacent dependencies - a-X-b relationships where b is predicted by a, but X is an unrelated element that takes a number of forms (e.g., Creel, Newport, & Aslin, 2004; Frost & Monaghan, 2016; Gebhart, Newport, & Aslin, 2009; Gómez & Maye, 2005; Newport & Aslin, 2004; Van Heughen & Shi, 2010). Gómez (2002) claims that greater variability for element X aids in the learning of the a-b relationship, possibly by reducing the likelihood of developing strong a-X or X-b representations. This reinforces the idea that additional complexity in the stimulus-set can aid in the learning of statistical structures if it does not introduce conflicting information. It is possible then that by sanitising naturalistic stimulus-sets we remove some of the statistical information necessary to facilitate learning and, it could be argued, that to truly ascertain the efficacy of the statistical learning mechanism it must be studied in complex, naturalistic, stimulus-sets.

Finally, there is some discrepancy as to whether longer listening times (in the head-turn preference paradigm) represent a preference for novelty (e.g., Chambers, Onishi, & Fisher, 2003; Saffran, Aslin, & Newport, 1996) or familiarity

(e.g., Gómez & Gerken, 1999; Seidl et al., 2009). This is an unfortunate artefact of research with infant participants which could be construed as allowing the preferential interpretation of data to support a favourable conclusion. The research presented over the coming chapters circumvents this issue by using adult participants; this confers the advantage of allowing more precise measures of familiarity than could realistically be observed in an infant population.

2.4 TRANSITIONAL PROBABILITY

Despite the growing body of research, clear evidence is yet to be provided as to exactly what distributional information is being attended to. The most prevalent theory is that individuals are accessing transitional probabilities (Fiser, 2009) - the probability of an item occurring given that another item has already occurred - and there is a wealth of evidence suggesting this may be the case. Further to the seminal study by Saffran, Aslin, and Newport, several studies have used transitional probabilities to describe the statistical learning mechanism. Saffran, Newport, Aslin, Tunick et al. (1997) exposed both children and adults to Saffran, Aslin, and Newport's (1996) artificial language by playing it in the background whilst they engaged in a computer-based illustration task. They demonstrated that, when adults were asked to indicate which of two novel stimuli sounded more like the familiarisation language, participants performed significantly better than chance; furthermore, infants showed a marked preference for the unfamiliar stimuli. Similarly, Thiessen and Erickson (2013) showed the same pattern of learning with infants as young as five-months.

Furthermore, Thompson and Newport (2007) used a language of eighteen CVC nonsense-words to show that participants are also sensitive to the transitional probabilities across phrase boundaries. Over the past two decades researchers have continued to find transitional probabilities to be a robust indicator of performance across several different tasks and languages (e.g. Aslin et al., 1998; Conway & Christiansen, 2005; Daikoku et al., 2014; Frank et al., 2010; Goodman et al., 2008; Hay et al., 2011; Johnson & Tyler, 2010; Kirkham et al., 2002; Koelsh et al., 2016; Liu & Kager, 2011; Newport & Aslin, 2004; Reeder et al., 2017; Saffran, Johnson et al., 1999; Saffran, Newport, & Aslin, 1996; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997; Theakston et al., 2004; Thiessen & Erickson, 2013; Toro et al., 2005; Vouloumanos, 2008). Table 2.1 shows a cross-section of studies chosen at random from the statistical learning literature. This includes several studies of statistical learning as well as the chosen paradigm, participant sample, type of stimuli used, and the distributional statistics investigated. A marked preference for transitional probability and related probabilistic measures of statistical distribution can be seen, with very few studies examining alternate measures. This represents only a small proportion of the statistical learning literature but, due to the sampling procedure chosen, should provide a fair assessment of the distribution of metrics reported in previous research.[1]

[1] 33, 102 studies were identified using Nottingham Trent University's Library OneSearch function using the search term Statistical learning and the filters: Psychology, Years: 1996-2019, Peer-reviewed journals. These were then exported to Excel and allocated a random number using the =RAND() function and sorted from low to high. The first 24 items were then selected as being representative of the literature.

Table 2.1: Summary of a selection of statistical learning studies including the study paradigm, participants, stimuli, and the distributional statistic used

Study	Paradigm	Participants	Metric	Stimuli
Anderson, Morgan, & White. (2003, Experiment 1)	Headturn Preference Procedure	Infants aged ~8.5 months	Likelihood Criterion	Three syllable-pairs (2 English, 2 Hindi, 2 Neutral)
Anderson, Morgan, & White. (2003, Experiment 2)	Headturn Preference Procedure	Infants aged 6-7 months	Likelihood Criterion	Three syllable-pairs (2 English, 2 Hindi, 2 Neutral)
Anderson, Morgan, & White. (2003, Experiment 3)	2-AFC	Adults (Age unspecified)	Likelihood Criterion	Three syllable-pairs (2 English, 2 Hindi, 2 Neutral)
Aslin, Saffran, & Newport (1998)	Headturn Preference Procedure	Infants aged 8-months	Transitional Probability	Four trisyllabic words comprising twelve unique syllables
Conway, Bauernschmidt, Huang, & Pisoni (2010, Experiment 2)	Non-word Repetition	Undergraduate students aged 20-25	Transitional Probability	Four CVC non-words
Conway & Christiansen (2005)	Gramaticality judgements	Undergraduate students (Age unspecified)	Associative chunk strength	Tactile sequences (Experiment A)
				Visual sequences (Experiment B)
				Tonal Sequences (Experiment C)
Daikoku, Yatomi, & Yumoto (2015)	Familiarity judgement	Adults aged 24-36	Transitional Probability	Tonal sequences
Dell, Reed, Adams, & Meyer (2000)	Non-word reading	Undergraduate students (Age unspecified)	Phonotactic Regularity	Thirty-two CVC syllables
Durrant, Taylor, Cairney, & Lewis (2011)	2-AFC	Adults aged 19-45	Transitional Probability	Tonal sequences
Evans, Saffran, & Robe-Torres (2009)	2-AFC	Children aged 6 to 14 years with SLI and matched controls	Transitional Probability	Four trisyllabic words comprising twelve unique syllables
Fiser & Aslin (2002)	2-IFC	Undergraduate students (Age unspecified)	Joint probability	Twelve simple black shapes
Frank, Goldwater, Griffiths, & Tenenbaum (2010)	2-AFC	Adults (Age unspecified)	Sentence length and number of tokens	Six-word language comprising words between two and four syllables.
			Transitional probability	
			Mutual information	
			Dirichlet distribution	

Study	Paradigm	Participants	Metric	Stimuli
Gerken, Wilson, & Lewis (2005)	Headturn Preference Procedure	Infants aged ~1.5 years	Morphological markers	Six masculine and six feminine Russian words
Hauser, Newport, & Aslin (2001)	Orientation response	Cotton-top tamarins	Transitional probability	Four trisyllabic words comprising twelve unique syllables
			Co-occurrence frequency	
Hay, Pelucchi, Estes, & Saffran (2011)	Headturn Preference Procedure	17-month old infants	Transitional probability	Four four-letter Italian words
Isbilen, McCauley, Kidd, & Christiansen (2107)	2-AFC; Recall task	Undergraduate students (Mean age 19.78)	Transitional probability	Six trisyllabic words comprising eighteen unique syllables
Johnson & Jusczyk (2001)	Headturn Preference Procedure	8-month old infants	Transitional probability	Four trisyllabic words comprising twelve unique syllables
Johnson & Tyler (2010)	Headturn Preference Procedure	Infants aged 5.5 & 8-months	Transitional probability	Four pseudorandomised disyllabic and trisyllabic words
Kim, Seitz, Feenstra, & Shams (2009)	Rapid serial visual detection; recognition questionnaire	Undergraduate students aged 18-35	Joint probability	Twelve simple black shapes
Kirkham, Slemmer, & Johnson (2002)	Preferential looking procedure	Infants aged 2-8 months	Transitional probability	Six coloured shapes
Koelsh, Busch, Jentschke, & Rohrmeier (2016)	2-AFC	Adults (Mean age 25.6)	Probability	Six timbre sounds combined into triplets
Lew-Williams & Saffran (2012)	Modified headturn Preference Procedure	Infants aged 9-11 months	Transitional probability	Thirty bisyllabic and trisyllabic words
Milne, Petkov, & Wilson (2017)	Preferential looking procedure with eye-tracking	Adult Rhesus macaques	Transitional probability	Artificial grammar comprising five unique stimuli
Monroy, Geron, & Hunnius (2017)	Preferential looking procedure with eye-tracking	Toddlers aged ~19 months	Transitional probabilities	Six observed actions
Saffran, Aslin, & Newport (1996)	Headturn Preference Procedure	8-month old infants	Transitional probability	Four trisyllabic words comprising twelve unique syllables
Thiessen & Erickson (2013)	Headturn Preference Procedure		Transitional probability	

1) Note that, although not identical to transitional probability, joint probability, associative chunk strength, and likelihood criteria are all probabilistic functions of the stimulu-set. 2) In the stimuli column, the term 'words' refers to letter strings concatenated to form pronouncable word-like items

This wealth of evidence would suggest that transitional probability does in fact constitute a robust predictor of learning performance and has contributed to it becoming the preferred metric of statistical distribution. In addition, it has been claimed that transitional probabilities insulate the learner against the under-segmentation of high frequency pairs (Aslin et al., 1998) whilst still incorporating the raw frequency of co-occurrence. For example, if both the and dog are high frequency items, a learner utilising a frequency-based mechanism may struggle to disambiguate the two, rendering them as a single item in the lexicon. Since transitional probabilities also account for the presence of other items, high frequency pairs are still represented but a learner is less likely to suffer under-segmentation errors. For this reason, raw co-occurrence frequency has largely been overlooked in statistical learning paradigms.

However, transitional probabilities cannot reasonably account for several effects highlighted in the existing literature. Saffran, Newport, and Aslin (1997) exposed both children and adults to either twenty-one or forty-two minutes of their artificial language (Saffran, Aslin, & Newport, 1996) and found that both groups performed better on a two-alternative fixed-choice test after the longer exposure time. Crucially, the forty-two-minute condition was achieved by repeating the twenty-one-minute sequence. This means that the longer condition maintained the same transitional probabilities as the shorter sequence. However, the frequency of the items was doubled in the forty-two-minute sequence. Therefore, the improvement in statistical learning performance cannot convincingly be attributed to differences in transitional probability.

Furthermore, Dell et al. (2000) tested participants on their ability to read sequences of four CVC words (e.g. sef-gem-mek-heg) in one of four conditions. These conditions varied on the legal onsets and codas (the sounds at the beginning and end of a syllable) of words within each sequence and whether the participants were informed of these rules. Over the course of the experiment, participants demonstrated an adherence to the phonotactic structure of over 97%, regardless of condition. This shows that participants were able to align themselves to the underlying structure of the language even when not explicitly aware that such structure existed. Most interestingly, however, is that fact that the presentation of items within the sequence was randomly generated within frequency constraints. That is, individual words were restricted to only appear either eight, twelve, or twenty-four times within a ninety-six-sequence set but different concatenations were generated for each participant. This means that, although the transitional probabilities remain the same within items, they cannot be reliably tracked across items – that is, each participant encountered marginally different inter-item transitional probabilities – making a transitional probability hypothesis less tenable. Crucially, the identification of the onset-coda relationship could also be explained using a frequency hypothesis. Unfortunately, data on whether participants were more error-prone on the lower frequency items is not available as this would allow some measure of discrimination between the two hypotheses.

2.5 THE CASE FOR BIGRAM FREQUENCY

However, Erickson and Thiessen (2015) argue that the explicit computation of transitional probabilities is less psychologically plausible than a frequency based chunking mechanism since the latter is more flexible when switching between different units within the language (e.g. phonemes, syllables, or words). This is consistent with evidence from computational models such as PARSER (Perruchet & Vinter, 1998) and MOSAIC (Model of Syntactic Acquisition in Children; Freudenthal et al., 2015). Freudenthal and colleagues used a modified version of MOSAIC to model errors in children's speech based on a frequency driven chunking mechanism. By assigning a co-occurrence threshold to individual words MOSAIC creates lexical entries for common phrases (e.g. go here, make that) leading to a reduction in output errors. Through this they successfully demonstrate that co-occurrence frequencies contribute to the development of formalised grammar. It is not implausible then to suggest that existing research, which describes the effects of transitional probability, may be tapping into a simpler, frequency-based mechanism of learning which is being masked by transitional probabilities. Unfortunately, current paradigms are incapable of differentiating between the two effects. There has been some attempt to contrast the relative contributions of transitional probability and frequency to statistical learning; for example, Endress and Langus (2017) examined participants' ability to learn sequences of shapes and pictures of everyday objects and concluded that transitional probabilities were weighed higher than frequency in French, Italian, and

Spanish/Catalan speakers. However, the transitional probabilities used in their study are .5 and .33. As discussed above, these are much higher than those seen in naturalistic stimulus-sets and may result in a biased estimate in favour of transitional probability; a fact that endures throughout much of the statistical learning literature.

The acceptance of transitional probabilities has arguably led to a state-of-play in which research has neglected to examine other potential variables in favour of transitional probabilities - though there have been calls to reconsider this position (e.g., Slone & Johnson, 2018). This is surprising given that frequency has been described as ubiquitous in language acquisition (Ambridge, Kidd, Rowland, & Theakston, 2014), yet comparatively little has been done to investigate the effects of frequency in statistical learning (e.g., Oganian, Conrad, Aryani, Heekeren, & Spalek; 2015; Schuler, Reeder, Newport, & Aslin, 2017) despite claims by Erikson and Thiessen (2015) that this may be a more psychologically plausible mechanism than a probabilistic account. The lack of plausibility attributed to transitional probabilities may be due to the potential computational effort required to track and calculate them - as noted by Saffran et al. (1996).

The transitional probability for any given pair of stimuli can be expressed as:

$$P(w_t|w_{t-1}) = \frac{P(w_t, w_{t-1})}{P(w_{t-1})}$$

Where w_t represents the second item in a given two-item sequence and w_{t-1} represents the initial item. The formula therefore shows that the probability P of the second item, given the occurrence of the first item, is equal to the probability of the two-item sequence divided by the probability of the initial item.

It is therefore necessary to know the probability of the first stimulus as well as the probability of the two-stimulus combination. These in turn require calculations based on frequency of the stimulus and total size of the stimulus set. In isolation, these do not represent particularly effortful calculations; however, each new interaction between learner and stimulus-set changes the probabilistic representation of the entire set. Consider the following example, in which the transitional probability of the bigram AB is calculated for the binary sequence:

$$BBAABABA$$

Using the previously presented formula it is apparent that $P(B|A) = .5^2$ since A is followed by itself once, by B twice, and by the end of the sequence. If we increase the sequence, as would happen with unfolding sentences or conversations:

$$BBAABABAB$$

[2] Here, P(B|A) is shorthand for the probability of B occurring if the previous item in the sequence is A.

The transitional probability of AB now becomes $P(B|A) = .75$, since A is still followed by itself once but is now followed by B three times (and is no longer followed by the end of the sequence). However, note that this changes the distribution for other associated stimuli within the sequence. For example, $P(A|B)$ changes from .75 to .60 despite there being no change in the frequency of the bigram BA. It then becomes apparent that in a large, constantly evolving stimulus set - such as that represented by language - transitional probabilities must be constantly maintained in order to provide a meaningful metric to judge inter-stimuli associations.

Given the additional complexity of calculating transitional probabilities, this raises two questions: (1) If a simpler (frequency-based) mechanism can facilitate effective learning, what benefit (if any) arises from the use of a more complex one? and (2) do learners require an accurate probabilistic representation of the stimulus-set to learn its inherent properties? There has been little attempt within the statistical learning literature to address these questions. However, decision-making in other domains (e.g., medicine) shows that both domain experts and naive participants consistently perform better when problems are framed in terms of frequency rather than probability (McDowell, Galesic, and Gigerenzer, 2018). Moreover, work by Tversky and Kahneman (1973) on classic reasoning tasks suggests that individuals prefer to make decisions based on heuristics rather than probability, even when probabilistic information is made available and that presenting problems in terms of frequency reduces cognitive bias (Kahneman, Slovic, & Tversky, 1982) and errors arising from the conjunction of two related events (Hertwig &

Gigerenzer, 1999; Tversky & Kahneman, 1983). While a representation of the stimulus-set based on transitional probability is more accurate, therefore, it may not provide a learning benefit commensurate to the increased computational complexity.

Contrast this with a frequency-driven account of statistical learning in which learners make decisions based on the frequency of items within the set. In such an account, the addition of more items requires only that the learner update the frequency of that item rather than their probabilistic representation of the entire stimulus set. To revisit the previous example, extending the sequence increases the frequency of AB from two occurrences to three and has no effect on the frequency of the other bigrams within the sequence. That is not to say that a frequency-based representation is the best (or even an accurate) representation of the stimulus-set but that it presents the less cognitively effortful of the two mechanisms and therefore, potentially, a more plausible and parsimonious account of statistical learning.

Frequency-based accounts of learning are not a new concept; there is a wealth of evidence documenting frequency-based effects across several diverse areas. Frequency has been shown to have a faciliatory effect on both serial- and free recall tasks (Balota & Neely, 1980; MacLeod & Kampe, 1996; Hulme, Roodenrys, Schweickert, Brown, Martin, & Stuart, 1997; Stretch & Wixted, 1998). Moreover, data from reading research has shown that higher frequency words and phrases result in increased fluency, shorter fixation periods and better parafoveal preview effects (Dahan, Magnuson, & Tanenhaus, 2001; Gerhand, & Barry, 1998; Inhoff & Rayner, 1986; Raynor & Duffy, 1986) as well as better sentence

comprehension and production (Arnon & Snider, 2010; Diessel, 2007). Word frequency is also considered to be a major predictor of word naming and lexical decision performance (Grainger, 1990; Perea & Carreiras, 1998; Schilling, Rayner, & Chumbley, 1998). These effects are often accounted for by experiential models of learning, in which frequency of occurrence is considered an indicator of prior experience. Descriptions of how this experience manifests generally fall into three broad categories (though their exact nature varies across individual models); stronger representations of more frequent items (e.g., Bybee, 1998; Tomasello, 2000) stronger connections between frequently co-occurring items (e.g., Rumelhart, Hinton, & McClelland, 1986), or larger/more enhanced representations for frequently occurring items (Jones, 2016, Jones & Macken, 2018). It is easy to imagine a cognitive architecture in which repeated exposure to words and associations across words could create new associative knowledge or increase the strength of the associated representations, and/or the links between those representations. It is less clear how (or why) probabilistic information would be represented in such a system since this would require not only the individual representations but also an overarching representation of all previously experienced language from which to calculate transitional probabilities. Though it is possible that developing these probabilistic representations is, in fact, useful for scaffolding learning the question remains as to whether the utility of doing so is commensurate to the extra effort involved in building and maintaining such a system.

Furthermore, it has been demonstrated that children who display atypical language development, such as those with SLI, can learn the implicit statistical

structure of a language after longer exposure periods but not shorter ones (Evans et al., 2009). Since the transitional probabilities of the language do not change based on length of exposure there is no reason to assume that they are responsible for the improvement in performance. Frequency, on the other hand, does increase in relation to the length of the exposure - participants are exposed to twice as many instances of each stimulus in a forty-two-minute sample than in a twenty-one-minute sample of the language - it is therefore more plausible to suggest an effect of frequency in learning rather than one of transitional probability.

However, it is undeniable that transitional probability provides additional information beyond that which can be explained by a frequency-based model of learning. The transitional probability of any given bigram stems from an interaction between the frequency of the bigram AB and the number of potential candidates for what can follow A. For example, to calculate P(B|A) one needs to know how often the bigram AB occurs as well as how often A is followed by other items. It therefore becomes necessary to introduce a second distributional metric - which we will term bigram diversity - to examine the key components of transitional probability.

2.6 THE CASE FOR BIGRAM DIVERSITY

It is recognised that predictability is an important facet of language processing which draws heavily on the statistical regularities of the text (Bates & MacWhinney, 1987; Glenberg & Gallese, 2012; Goldberg, Casenhiser, &

Sethuraman, 2005; Pickering & Garrod, 2004; 2007; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005) to aid reading speed and comprehension (Conway, Bauernschmidt, Huang, & Pisoni, 2010). Since the predictability of a language is directly related to the probability of Y following X in the sequence XY it follows therefore that a larger number of potential competitors for stimulus Y would serve to reduce predictability and thereby prove detrimental to response fluency. However, despite the demonstrable impact of SL mechanisms they have been mostly ignored in the wider literature.

Bigram diversity is defined here as the number of items that potentially follow a word in a two-word sequence (e.g., the number of candidates for *X* that follow the word *A* in the sequence *AX*). A more concrete example can be seen in the bigram credit card which occurs a total of 508 times throughout the British National Corpus (2007) giving it a bigram frequency of 508. The word credit however is followed by 109 different words including account, agreement, and note; it therefore has a bigram diversity of 109. Like bigram frequency, this also has the benefit of requiring less computational effort than transitional probability since learners are only required to keep track of the number of contexts in which a word appears rather than the relative frequencies of those contexts. This can be likened to the concept of contextual diversity; which can be derived by counting the number of contexts – for example, the number of documents within a given corpus - in which the item occurs (Adelman, Brown & Quesada, 2006). Furthermore, since the number of words that co-occur with A is likely correlated with the number of contexts in which it appears, it follows

that the observed effects of contextual diversity may be similar to those of bigram diversity.

Adelman et al. (2006; also, Adelman & Brown, 2008) demonstrated that both lexical decision times and word-naming performance improve for more contextually diverse items independent of individual word frequency, suggesting that participants develop a stronger lexical representation for items that occur in multiple linguistic contexts. Likewise, increased diversity in caregiver speech improves vocabulary acquisition in children (Hurtado, Marchman, & Fernald, 2008; Jones & Rowland, 2017; Rowe, 2008); Yu and Smith (2007) suggest that having access to multi-context cues may help learners solve the indeterminacy problem (Quine, 1960) - possibly through the development of context-independent lexical representations. Being able to disambiguate lexical representations from their observed context(s) may facilitate response fluency
(as in Adelman et al., 2006), particularly if the paradigm is context independent. Given these trends, we would expect higher diversity bigrams to provide a faciliatory effect to learning.

However, it is also claimed that predictability is an important facet of language processing (Bates & MacWhinney, 1987; Glenberg & Gallese, 2012; Goldberg et al., 2004; Pickering & Garrod, 2004, 2007; Van Berkumet al., 2005). Since bigram diversity is essentially an indicator of predictability it follows that a larger number of potential competitors for stimulus *X* in the bigram *AX* would serve to reduce predictability and thereby prove detrimental to response

fluency - a trend we would also expect if learning is guided by transitional probability. Given the competing nature of these predictions the role (if any) of bigram diversity remains unclear.

2.7 RESEARCH IN NATURAL LANGUAGE

The case for using natural language corpora to study statistical learning is one of ecological validity (Erickson & Thiessen, 2015; Romberg & Saffran, 2010). Much of the statistical learning literature, particularly pertaining to linguistic stimuli, is concerned with the ability of learners to detect distributional patterns in relatively small artificial grammars. It has been argued that these languages lack the complexity required to allow for valid conclusions as to how learners are able to process distributional statistics within natural language (Frank et al., 2010; Johnson & Tyler, 2010), something that may be particularly true in studies that utilise very short utterance lengths. It is not unreasonable therefore to suggest that learning under the simplified conditions of artificial grammars cannot adequately represent performance in more naturalistic arenas. This is less of a problem for frequency-based accounts since the frequency of the item is only affected by the size of the stimulus-set to the extent that the number of occurrences is likely to increase as a function of the overall exposure to the language whereas transitional probabilities cannot discriminate based on the length of exposure - a two-minute sample of Saffran, Aslin, and Newport's mini-language retains the same transitional probabilities as a ten-minute sample.

It has also been claimed that performance in statistical learning tasks may be influenced by existing statistical biases arising from an overlap in speech sounds between artificial and natural languages (Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018); making it impossible to dissociate learning from existing statistical preconceptions. This leads to poor internal consistency since performance for individual stimuli can be predicted by their similarity to real world examples. Given that the limited stimulus-sets in artificial grammars are characterised by inflated statistical associations and potentially vulnerable to existing linguistic bias, any conclusions regarding the efficacy of statistical learning can only be tentative until the effects are replicated with naturalistic stimulus-sets. Natural language corpora represent an opportunity to extend the contribution of artificial grammar research by enabling the design of naturalistic, yet quantifiable stimulus-sets. Databases of real-world language allow for the extraction of distributional statistics that resemble participants' existing representations. Using stimuli from these corpora, rather than an artificial grammar, retains the complexity and diversity of natural language, whilst allowing for the accurate tracking of distributional cues.

However, examining statistical learning in a natural language corpus requires an unconventional approach to testing. Traditionally, statistical learning paradigms consist of a familiarisation phase - where participants are exposed to an unfamiliar stimulus-set - and a testing phase. For example, the Headturn preference procedure (Fernald, 1985) is commonly used to assess statistical learning in infants (e.g., Anderson, Morgan, & White, 2003; Evans et al., 2009; Johnson & Jusczyk, 2001; Johnson & Tyler, 2010; Lew-Williams & Saffran, 2012;

Maye, Werker, & Gerken, 2002; Saffran, 2001; Saffran, Aslin, & Newport, 1996; Saffran & Wilson, 2003; Thiessen & Saffran, 2003) where the stimuli are presented aurally. During this procedure the infant is usually seated with a caregiver in the centre of a sound-attenuated cubicle with a fixation light to either side and another directly in front of them. Following the familiarisation phase test items are presented from either the left or right side of the cubicle; infant looking behaviour is then taken as a measure of learning. Alternatively, older participants can be presented with discrimination tasks (e.g., Fiser & Aslin, 2002; Saffran, Johnson & Aslin, 1999; Toro, Sinnett, Soto-Faraco, 2005; Turk-Browne, Jung, & Scholl, 2005) where they are presented with several (usually two) options and asked to indicate which they find most familiar, based on the previous familiarisation phase. Variations on this task include asking participants to indicate whether a novel sequence follows the same 'rules' as those presented during familiarisation (e.g., Conway & Christiansen, 2005; Milne, Petkov, & Wilson, 2017) or to predict some outcome or continuation of the sequence (e.g., Monroy, Gerson, & Hunnius, 2017; Romberg & Saffran, 2013). More recently, novel approaches to assessing statistical learning have been developed (e.g., Isbilen, McCauley, Kidd, & Christiansen, 2017) but these too are vulnerable to the limitations of artificial grammars.

The foremost concern with natural language stimuli is that it they are unsuitable for use with any of the aforementioned methodologies. While it is possible to use an abstracted domain to manipulate familiarity based on exposure within (for example) an artificial grammar, it is not possible to do the same when using natural language datasets where participants already have

considerable prior knowledge. Therefore, when dealing with natural language stimuli, one solution may be to use non-native language stimulus-sets since these are comparable to in both size and complexity whilst circumventing the problem of familiarity. However, compared to native languages exposure to non-native stimuli is necessarily pauce and does not provide comparable opportunity for the encoding of their statistical associations without the need for prohibitive familiarisation periods. As such, it may be preferable to find new ways of assessing learning whilst still retaining the complexity of the native language and avoiding a lengthy familiarisation process.

One solution is to reframe existing language tasks to examine the effects of statistical learning. One such task, which has been used extensively within the word-recognition literature, is the primed lexical decision task. This task involves asking participants to discriminate between word and non-word stimuli and has been shown to be sensitive to a broad range of variables (e.g., Perea, Marcet, Vergara-Martínez, & Gómez, 2016) including structural- (e.g. Dijkstra, Hilberink-Schulpen, & van Heuven, 2010) and associative-priming effects (e.g., Perea & Gómez, 2010). There is ample evidence that individual trial performance can be affected by a previously shown prime. Examples can be seen in work by Lester, Feldman, and del Prado Martin (2017), who used data from the Semantic Priming Project (Hutchison et al, 2013) to show that responses to a target word vary as a function of syntactic similarity; or Yap, Hutchison, and Tan (2016) who showed semantic priming to be a reliable predictor of lexical decision performance.

It is theoretically possible that a statistical priming effect could be elicited by manipulating the prime-target relationship based on the natural distributional statistics of a language. This should allow for the examination of the previously learned statistical associations inherent in natural language whilst avoiding the oversimplification of artificial grammars or the lengthy familiarisation periods necessary with more complex languages.

2.8 **THIS THESIS**

Over the course of the next five chapters the current work attempts to address issues of complexity and ecological validity in statistical learning research by taking a novel approach to stimuli generation. The experiments presented herein draw on the British National Corpus as a source of naturalistic stimuli and assess the influence of distributional statistics on task performance in a simple, primed lexical decision task. Following this, I will present two novel sequence learning tasks

Furthermore, I will compare the relative merits of transitional probability, bigram frequency, and bigram diversity in predicting task performance.

3 PROOF OF CONCEPT

CHAPTER OVERVIEW

Over the course of this chapter I shall:

- Assess the viability of using lexical decision tasks to investigate statistical learning performance in naturalistic stimulus-sets
- Use lexical decision data to inform Bayesian multi-level models of word recognition performance
- Compare different statistical models of task performance using both leave-one-out and Bayesian methods
- Detail the most accurate model for both bigram frequency and bigram diversity and briefly discuss the theoretical implications

3.1 PREPARATION

The following code excerpt initialises the packages necessary to run the analyses in this chapter and introduces some global settings in the interest of reproducibility.

```
library(formatR)
library(readr) library(brms)
library(GGally)
set.seed(100)
```

The studies presented in this chapter are intended to act as a proof of concept and highlight the sensitivity of the task to the inherent associations within natural language. In a departure from traditional statistical learning paradigms participants will not be required to learn any new information, thus eliminating the need for a lengthy familiarisation period. Instead, the task attempts to access previously learnt associations and demonstrate their influence on response times. Experiment 1 examines these associations by manipulating bigram frequency whereas Experiment 2 assesses the impact of bigram diversity.

3.2 EXPERIMENT 1: BIGRAM FREQUENCY.

Experiment one used a lexical decision task to assess the extent to which bigram frequency affects word recognition. The aim of the experiment was to show any statistical priming effect that may result from high frequency word pairs within natural language.

3.2.1 Participants.

Thirty participants (24 females) aged between 18 and 60 years (M= 34, SD= 11.56) were recruited from within Nottingham, UK; all participants reported English as their first language and reported having no language difficulties. Participants took part in both experiments; research participation credits were offered for participation where applicable. Participants who responded correctly to fewer than 80% of trials on the lexical decision task (N=3) were excluded from the analysis.

3.2.2 **Materials.**

The experimental stimuli consisted of ninety bigrams and ninety non-word stimuli (paired with real-word primes) between three and eight letters long. Non-word stimuli were created by transposing letters from the target items (e.g., SIHGT, PTAH, WHSOE). Each non-word was paired with a unique real word prime chosen pseudo-randomly from the BNC - primes were constrained to not appear more than once across the two experiments. Bigrams were extracted from the BNC by using a python script to parse the .xml version of the corpus into word pairs before writing them to a database and tallying the number of occurrences. This resulted in a list of 12,293,349 unique bigrams. A further script was used to remove any bigrams with a frequency of less than .1 per million. The remaining corpus was then filtered to exclude any bigrams containing acronyms, initialisations, contractions, hyphenations, non-standard or non-English words, names, numerals, or words with fewer than three letters.

Data was also obtained for frequency (Leech, Rayson, & Wilson, 2001), concreteness (Brysbaert, Warriner, & Kuperman, 2014), and number of letters for the target words in each bigram; bigram diversity was also calculated but was free to vary across stimuli and not used in the initial analysis. The bigrams used in the experiment were selected to include an equal number of high, low, and zero frequency items; examples of each are given in Table 3.2. For illustrative purposes, mean values are also provided for bigram frequency, transitional probability, and individual word frequency for both prime and target as they appear per million words in the BNC, as well as target length and concreteness; values are expressed as logarithms where this was used in the analyses. Stimuli from each level of bigram frequency were balanced

so as to not differ significantly on any of the aforementioned characteristics using independent-samples t-tests (each $p > .05$) with the exception that, when compared with high frequency bigrams, low frequency bigrams differed significantly on the number of letters in the target word ($p=0.04$); full stimuli lists are available in the appendices, descriptive statistics for each level are presented in table 3.1 and example bigrams are shown in table 3.2.

Table 3.1: Group means and standard deviation (in parenthesis) for High, Low, and Zero frequency bigrams.

Level	Bigram_frequency	Bigram_diversity	Target_frequency	Letters	Concreteness	Transitional_probability
High Frequency	742.31 (21.77)	248.38 (647.78)	136.31 (9.81)	4.83 (1.12)	3.67(.96)	.13 (.13)
Low Frequency	10.94 (3.15)	293.06 (582.76)	108.38 (5.90)	5.46 (1.16)	3.29 (.85)	.08 (.09)
Zero frequency	.00 (.00)	205.87 (488.84)	32.12 (10.94)	5.40 (1.22)	3.40 (1.06)	.00 (.00)
Log-transformed values						
High frequency	6.63 (3.09)	5.51 (6.47)	4.93 (2.29)	-	-	-2.04 (2.04)
Low Frequency	2.4 (1.15)	5.68 (6.37)	4.7 (1.78)	-	-	-2.53 (2.41)
Zero frequency	-13.82 (.00)	5.33 (6.19)	3.48 (2.40)	-	-	-13.82 (.00)

Table 3.2: Example stimuli, including descriptive statistics, for Experiment 1

Prime	Target	Group	Bigram_Diversity	Bigram_Frequency	Transitional_Probability	Letters	Concreteness	Word_Frequency
eighth	army	High Frequency	21	187	0.146322379	4	4.70	114.41
bile	acid	High Frequency	17	234	0.197802198	4	4.25	49.68
fleet	street	High Frequency	25	305	0.135736538	6	4.75	196.14
rugby	union	High Frequency	39	428	0.124238026	5	3.38	176.07
good	practice	High Frequency	850	461	0.005705799	8	2.52	171.14
daily	post	High Frequency	74	512	0.067112335	4	4.30	93.39
cash	flow	High Frequency	134	573	0.066721006	4	3.72	52.44
always	accept	Low Frequency	567	10	0.000216319	6	3.03	98.07
craggy	face	Low Frequency	1	10	0.082644628	4	4.87	349.78
interest	account	Low Frequency	149	10	0.000362174	7	3.08	158.91
local	access	Low Frequency	568	10	0.000215689	6	2.71	109.40
people	achieve	Low Frequency	679	10	0.000080500	7	2.29	67.68
rustic	style	Low Frequency	1	10	0.041152263	5	2.67	107.25
time	across	Low Frequency	569	10	0.000064500	6	3.07	252.03
abase	number	Zero Frequency	0	0	0.000000000	6	3.30	493.85
building	food	Zero Frequency	177	0	0.000000000	4	4.80	189.92
drubs	nudge	Zero Frequency	0	0	0.000000000	5	4.47	1.53
geese	wits	Zero Frequency	4	0	0.000000000	4	1.76	4.00
lifer	hugs	Zero Frequency	0	0	0.000000000	4	4.14	1.03
oval	hipster	Zero Frequency	9	0	0.000000000	7	2.50	190.60
rethinks	scaly	Zero Frequency	0	0	0.000000000	5	4.22	0.75

3.2.3 **Procedure.**

Participants were presented with letter strings and were asked to indicate whether the string constituted a real English word by pressing either ‘z’ or ‘m’ on a standard QWERTY keyboard; key mapping was systematically varied so that half of all participants used ‘z’ to indicate a word and ‘m’ to indicate a non-word whilst half responded with ‘m’ for words and ‘z’ for non-words. Strings were presented for a maximum of 3000ms or until the participant responded and were preceded by a 250ms prime. These times are slightly longer than those traditionally used in lexical decision but were chosen to give participants the best possible chance of encoding the prime since it was unclear whether any statistical priming effect might exist. All prime-target pairs mapped exactly onto bigrams from the stimuli lists whereby the

first word of the bigram acted as a prime for the second word. A fixation point was presented in the centre of the screen for 500ms prior to both the prime and target words. The prime was presented for 250ms and the target for a maximum of 3000ms or until the participant responded. A blank white screen was presented for 0ms between each aspect of the trial. Prime-Target pairs were presented in two counterbalanced blocks and the order of presentation for trials was randomised for each participant. A graphical representation of the experiment can be seen in Figure 3.1.

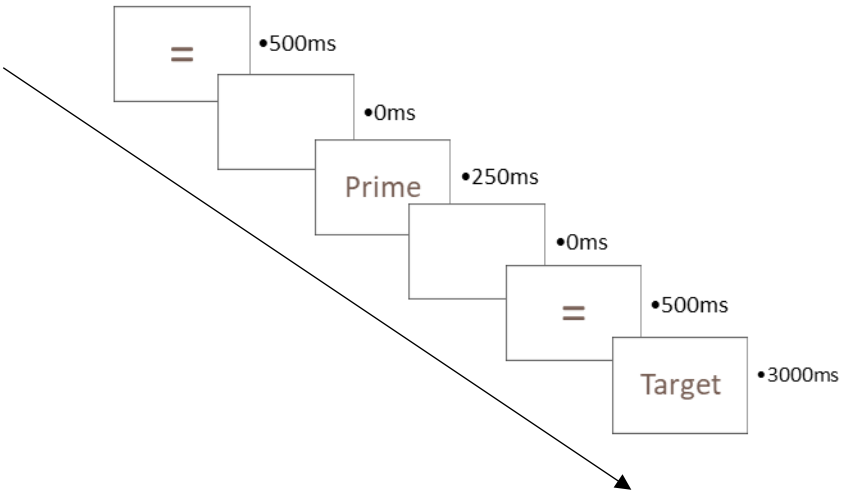

Figure 3.1: Diagram of the experimental procedure in Experiment 1.

3.2.4 **Choice of analysis**

The experiment was originally designed with the intention of comparing performance across frequency groups (high, low, and zero) using a one-way ANOVA since this considered the most appropriate analysis given my existing knowledge. However,

there is little theoretical justification for the use of arbitrarily defined levels for bigram frequency or diversity since they exist on a continuous scale in natural language. As such, I made the decision to conduct the analyses using Bayesian multi-level models rather than ANOVA. This enabled me to account for participant and item effects and to draw conclusions based on model fit rather than point estimates provided by p values. It is not my intention to address the arguments surrounding Bayesian vs. Frequentist approaches here since that would require a tome of its own and is beyond the scope of the current work. Suffice to say, the use of Bayesian modelling allows the examination of evidence for the null hypothesis rather than only the experimental hypothesis (see figure 3.2) and for the integration of priors derived from these first experiments to increase the efficiency of later models.

3.3 **RESULTS**

Data was first trimmed to exclude incorrect responses, then those more extreme than three standard deviations from the participant's mean (Madan, Shafer, Chan, & Singhal, 2016), finally responses faster than 200ms or slower than 1500ms were removed (Perea, Marcet, Vegara-Martínez, & Gomez, 2016). Following this procedure 5.78% of the remaining correct trials were removed across participants. Individual trial data (N=1828) was then analysed with Bayesian multi-level modelling using the brms package in R, full details of which are documented below. In addition to Bigram frequency and transitional probability, target-word frequency, concreteness, and number of letters as well as participant age were included as covariates. Unless otherwise stated, the following applies to all models: Monte Carlo Markov Chain (MCMC) sampling was achieved using the No-U-Turn Sampler (NUTS, Hoffman &

Gelman, 2014) implemented in Stan (Carpenter et al., 2017) using the RStan package (Stan Development Team, 2017); each model had four chains of 2000 iterations with a burn-in of 1000 iterations; and, all models used half Student-t priors with three degrees of freedom. Where specified, priors are expressed using the notation N(μ, σ) where μ is the mean and σ is the standard deviation of a normal distribution (N).

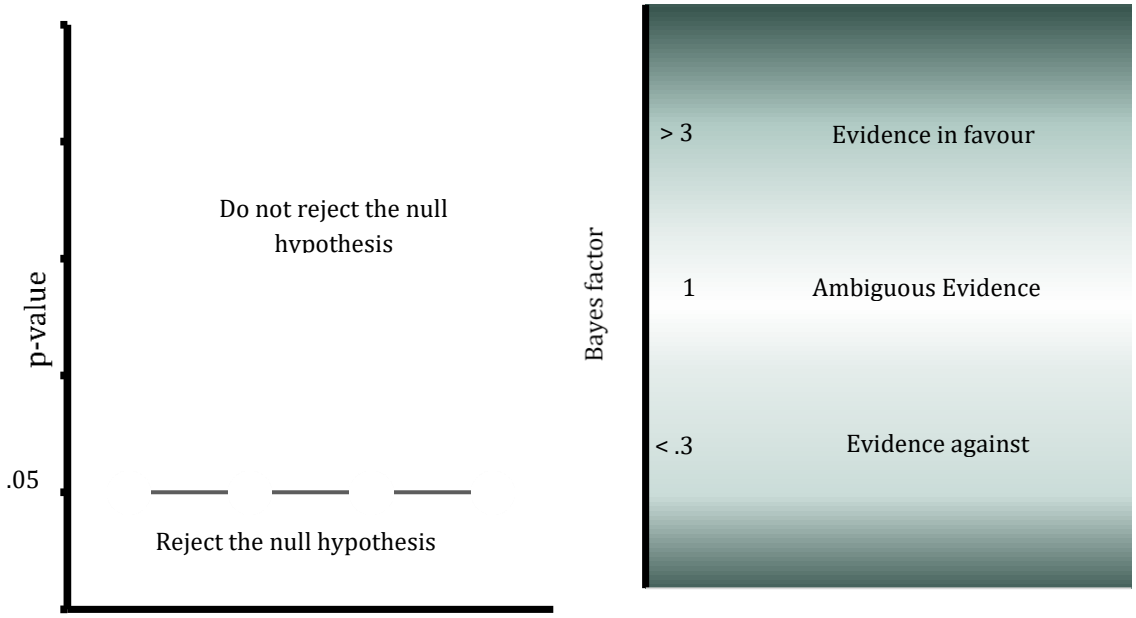

Figure 3.2: Valid statements based on p-values and Bayes factors. The p-value and the Bayes factor allow fundamentally different statements concerning the null hypothesis. The p-value can be used to make a discrete decision: reject or retain the null hypothesis. The Bayes factor grades the evidence that the data provide for and against the null hypothesis. Adapted from Hoekstra, Monden, von Ravenzwaaij, & Wagenmakers (2018)

The future predictive power of the models was assessed using Leave-One-Out Cross Validation (LOO-CV). LOO-CV is calculated by removing one observation from the data and training the model on the remaining n-1 observations; this process is repeated n times (where n is the total number of observations). The LOO-CV statistic is obtained by averaging across all iterations to obtain the expected log predictive density (elpd), this value is then converted to the deviance scale by multiplying the elpd by -2 allowing it to be interpreted in the same manner as Akaike Information Criteria (AIC) or equivalent (see Gelman, Huang, & Vehtari, 2014 for a discussion of information criteria in Bayesian model selection). Additionally, Bayes Factors were also computed using the built-in function in brms to show the likelihood of each model when compared to the others (see Rouder, Haaf, & Vandekerckhove, 2018 for an overview of Bayes Factors). The analyses resulted in some extreme Bayes factor values; since the aim is to show the likelihood of one model over another it was judged sufficient to express these values as being > 999 or $< .001$ as applicable.

3.3.1 **Data preparation.**

Data was read into R and assessed for normality and multicollinearity (See figure 3.3). Bigram frequency, transitional probability, and response time were log transformed prior to the analysis to achieve an approximation of a normal distribution; a small constant was added to all the values to avoid errors resulting from trying to calculate log(0). Descriptive statistics were also calculated for each variable and are shown in table 3.3.

```
df <- read_csv("Exp1_data.csv") ggpairs(data = df, columns = c(4:5, 7:8,
    12:13)) +  theme(panel.grid = element_blank())
df$log_word_freq <- log(df$word_freq + .000001)
df$log_bigram_freq <- log(df$bigram_freq + .000001)
df$log_trans_prob <- log(df$trans_prob + .000001)
df$log_response_time <- log(df$response_time + .000001)
```

Table 3.3: Means, standard deviations (SD), range, and inter-quartile range (IQR) for variables in Experiment 1

Variable	Mean	SD	Min	Max	Range	IQR
age	34.30	11.70	18.00	60.00	42.00	22.00
bigram_freq	484.90	1279.00	0.00	8465.00	8465.00	311.00
concreteness	3.50	0.95	1.68	4.97	3.29	1.71
diversity	211.78	497.00	0.00	3442.00	3442.00	153.00
letters	5.25	1.20	4.00	8.00	4.00	2.00
response_time	705.21	237.00	214.00	1497.00	1283.00	305.00
trans_prob	0.07	0.12	0.00	0.71	0.71	0.08
word_freq	14988.74	11477.00	51.00	49385.00	49334.00	13363.00

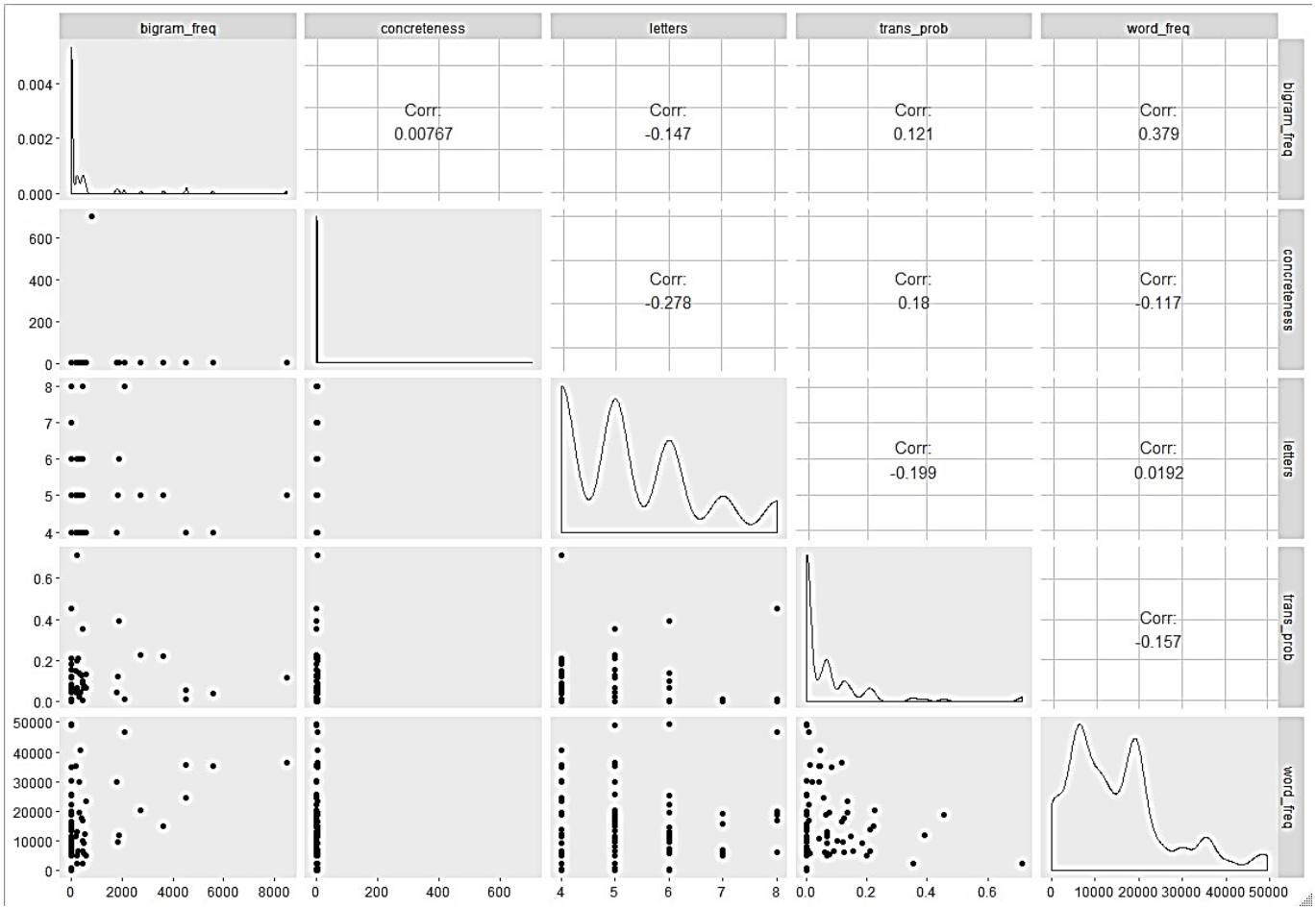

Figure 3.3: Matrix showing the correlations between predictors in Experiment 1. Also shown are the scatterplot showing the correlations and the distribution of values for each predictor.

It is interesting to note the lack of correlation between bigram frequency and transitional probability. Given that bigram frequency is a key component of the transitional probability calculation, one might expect the two to be highly correlated. However, transitional probability is weighted by the relative frequency of the bigram compared to all other bigrams starting with the same first word – for example, the transitional probability for the bigram chocolate mousse is weighted according to the relative frequencies of other bigrams that include chocolate in the first position, including chocolate fountain, chocolate covered, and chocolate lover. This weighting means that bigrams with the same frequency can have wildly different transitional probabilities. Equally, bigrams with the same transitional probability such as premier league and instances of - which have transitional probabilities of .35 – can have vastly different frequencies (879 and 270, respectively). Thus, although we might imagine some correlation between the two metrics no such relationship exists, as shown in figure 3.3.

3.3.2 **Specifying the models**

Firstly, a baseline model was run for the purpose of comparison. This model includes none of the predictors or anticipated covariates; if the baseline model fits the data better than the experimental models then we can conclude that there is either no effect of bigram frequency or transitional probability or that the task is not sensitive enough to detect any effects that may exist. As well as the baseline model, four additional models were also run. A covariate only model was used for comparative purposes – if this model is found to be the best predictor of the

data then it can be inferred that neither bigram frequency nor transitional probability are influencing response time in the lexical decision task. Additionally, three experimental models were used to assess the effects of A) bigram frequency, B) transitional probability, and C) both bigram frequency and transitional probability; these models are set out below.

```
base_model_1 <- brm(log_response_time ~ 1, data = df,  save_all_pars
    = TRUE)
cov_model_1 <- brm(log_response_time ~ age + concreteness + letters
    + word_freq, data = df, save_all_pars = TRUE, silent = TRUE,
    refresh = 0)
model_1a <- brm(log_response_time ~ log_bigram_freq + age +
    concreteness + letters + word_freq + (1 | subject) + (1 | item),
    data = df, save_all_pars = TRUE, silent = TRUE, refresh = 0)
model_1b <- brm(log response_time ~ log_trans_prob + age +
    concreteness + letters + word_freq + (1|subject) + (1|item),
    data = df, save_all_pars = TRUE, silent = TRUE, refresh = 0)
model_1c <- brm(log_response_time ~ log_bigram_freq +
    log_trans_prob + age + concreteness + letters + word_freq +  (1
    | subject) + (1 | item), data = df, save_all_pars = TRUE,
    silent = TRUE, refresh = 0)
```

Model B, the transitional probability model, failed to converge after 1000 iterations and was rerun using a maximum treedepth of 15; this allows for more efficient evaluation of the model parameters. A very accessible description of treedepth in Monte Carlo models can be found at: https://www.weirdfishes.blog/blog/fitting-bayesian-models-with-stan-andr/#a-note-on-divergences.

```
model_1b <- brm(log response_time ~ log_trans_prob + age +  concreteness
    + letters + word_freq + (1|subject) + (1|item),  data = df,
    save_all_pars = TRUE, silent = TRUE, refresh = 0,      control =
    list(max_treedepth = 15))
```

Note: It is possible to view a summary of any of the models by using summary(model_name) but I have not done that at this point because the model comparisons are more interesting at this stage of the analysis. I revisit the individual models after cross-validation and Bayes factor comparison.

3.3.3 **Cross-validation.**

Model comparison was performed using leave-one-out cross-validation with the loo() function in R (Vehtari, Gelman, & Gabry, 2017); smaller LOOIC values indicating less variance from the observed values and therefore represent a better description of the data than higher values. Information criteria for all the models are displayed in Table 3.4.

```
cv_base1 <- loo(base_model_1)
cv_cov1 <- loo(cov_model_1)
cv_m1a <- loo(model_1a)
```

```
cv_m1b <- loo(model_1b)

cv_mlc <- loo(model_1c)
```

Table 3.4: Leave-one-out information criteria comparing the statistical models of word recognition performance for Experiment 1. The table shows the population- and group-level predictors for each model as well as the information criteria and standard deviation (in parenthesis).

Model	Population-level	Group-level	LOOIC (SD)
Base	None	None	1018.5 (55.4)
Covariate	Age, letters, word frequency, concreteness	None	874.3 (60.3)
Model A	Age, letters, word frequency, concreteness, bigram frequency	participant, item	-229.2 (73.7)
Model B	Age, letters, word frequency, concreteness, transitional probability	participant, item	-228.6 (73.6)
Model C	Age, letters, word frequency, concreteness, bigram frequency, transitional probability	participant, item	-228.6 (73.6)

Cross-validation shows that the baseline model fits the data least well of the five models presented here whereas the covariate only model provides a reasonable improvement in predictive value compared to the baseline. All three experimental models perform better than both the baseline and covariate models suggesting that the lexical decision task is in fact sensitive enough to pick up on improvements to task performance stemming from participants' use of statistical information. Of the experimental models, Model A (Bigram Frequency) performs marginally better at predicting the data than the other models; however, since the models differ by less than 1.96 times the leave-one-out criterions standard deviation (as a heuristic for 95% confidence) it cannot be concluded that there is any meaningful difference in the predictive accuracy and as such, an alternative method is necessary to distinguish amongst them. High variance (as shown by the large standard deviations around each LOOIC) is not unusual in leave-one-out cross-validation; since each training set comprises

n-1 samples there is necessarily a large amount of overlap between iterations – since each training set differs from another by only one datum – which leads to highly correlated estimates and therefore higher variance (Hastie, Tibishirani, & Friedman, 2009).

3.3.4 **Bayes factors.**

Bayes factors were also computed using the bayes_factor() function and allow for direct comparison of the models in terms of a likelihood ratio.

```
bf_covbase1 <- bayes_factor(cov_model_1, base_model_1, silent = TRUE)
bf_1abase <- bayes_factor(model_1a, base_model_1, silent = TRUE)
bf_1bbase <- bayes_factor(model_1b, base_model_1, silent = TRUE)
bf_1cbase <- bayes_factor(model_1c, base_model_1, silent = TRUE)
bf_acov <- bayes_factor(model_1a, cov_model_1, silent = TRUE)
bf_bcov <- bayes_factor(model_1b, cov_model_1, silent = TRUE)
bf_ccov <- bayes_factor(model_1c, cov_model_1, silent = TRUE)
bf_1ba <- bayes_factor(model_1b, model_1a, silent = TRUE) bf_1ca <-
bayes_factor(model_1c, model_1a, silent = TRUE) bf_1cb <-
bayes_factor(model_1c, model_1b, silent = TRUE)
```

The resultant Bayes factor represents the strength of evidence for one hypothesis over another – assuming both hypotheses are equally likely - which can be interpreted as a ratio of BF:1, with possible values for Bayes factors ranging from zero to ∞. For two competing hypotheses a Bayes factor of 20 would therefore suggest that the data are 20 times more likely under the first hypothesis than the second. Conversely, a Bayes factor of .05 would indicate that

the data are 20 times more likely under the second hypothesis whereas a Bayes factor of one would indicate equal support for both hypotheses. These ratios are used here to directly compare the likelihood of each model and can be seen in Table 3.5. More extreme Bayes factors indicate a greater likelihood of one model over another given the observed data; the strength of evidence for a given Bayes factor is somewhat subjective but guidance on their interpretation is provided by Raftery (1995) who describes four categories of evidence: weak (BF = 1-3), positive (BF = 3-20), strong (BF = 20-150), and very strong (BF > 150); it is these criteria that I shall be subscribing to in my analyses. Furthermore, where the analysis results in a Bayes factor of less than one, the strength of evidence for the null hypothesis can be ascertained using 1/BF; for example, if BF = .169 then 1/BF = 5.92 which can be regarded as positive evidence for the null hypothesis using Raftery's guidelines. Note that, in the current analyses, I am using Bayes factors to compare two experimental models rather than contrasting an experimental and null hypothesis. In this case, a Bayes factor of less than one represents support for the first model and those greater than one for the second model. It is also worth noting that when BF = 3 this is roughly equivalent to p = .05 in a frequentist framework; therefore, any Bayes factor of greater than three would be regarded as significant using this interpretation (Dienes, 2014).

Table 3.5: Between model comparisons for Experiment 1 using Bayes factors. Comparisons relate to the strength of evidence for models in the left most column over those listed at the top of the table.

Model	Base	Covariate	A (Bigram frequency)	B (Transitional probability)
Covariate	>999			
A (Bigram frequency)	>999	>999		
B (Transitional probability)	>999	>999	<.001	
C (Combined)	>999	>999	<.001	0.169

The Bayes factors displayed in Table 3.5 show the strength of evidence for one model over another in the form of a ratio. For example, the data support Model A (Bigram Frequency) over the Base and Covariate models by a ratio of over 999:1 – a pattern we see repeated for all the experimental models - confirming that the experimental models are markedly better than both the baseline and covariate models given the observed data. Furthermore, there is a difference in the likelihood of Model A (bigram frequency) over Models B and C since the Bayes Factor in both cases is less than .001. This suggests that, although the models display the same predictive power (as measured using leave-one-out cross-validation), bigram frequency is a more plausible predictor of response time than transitional probability. It is also worth noting that Model C, which includes both bigram frequency and transitional probability, is less likely than the transitional probability model but only by a factor of around five, which would be considered positive but not strong evidence in favour of single experimental variable. This could be because transitional probability and bigram frequency are both capturing an element of frequency information and suggests that it is unlikely that participants are attending to both sets of statistical regularity.

3.3.5 **Model summary.**

A full summary of the model can be obtained using:

```
summary(model_1a)
```

Based on cross-validation statistics, all the models display similar predictive performance (i.e. how well they can generalize to new data). However, interpretation of the Bayes Factors using the thresholds set out by Raftery (1995) suggests the strongest evidence for Model A (bigram frequency) when compared to all other models. Given these results, we conclude that bigram frequency and not transitional probability has the most value in predicting response times for lexical decision in a statistical priming paradigm. Full details of Model A (Bigram Frequency) are shown in Table 3.6.

Table 3.6: Summary statistics for Model A (Bigram Frequency) expressed on a natural logarithmic scale

-	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Group-level effects						
Item	0.06	0.01	0.05	0.08	1,688.00	1.00
Participant	0.23	0.03	0.17	0.31	969.00	1.01
Population-level effects:						
Intercept	6.58	0.16	6.26	6.9	1,126.00	1.00
Bigram frequency	[-.01, .00]	[.00, .01]	[-.01, .00]	[-.01, .00]	4,000.00	1.00
Age	0.01	[.00, .01]	[-.01, .00]	0.01	1,054.00	1.00
Concreteness	-0.02	0.01	-0.04	[-.01, .00]	3,052.00	1.00
Letters	0.03	0.01	0.02	0.05	2,607.00	1.00
Word frequency	-0.03	0.01	-0.05	-0.02	3,371.00	1.00
Family specific parameters:						
Sigma	0.22	[.00, .01]	0.21	0.23	4,000.00	1.00

Bayesian multi-level models are interpreted in much the same way as frequentist multi-level models with Estimate and Est.Error replacing the unstandardized coefficients ß and Std.Error respectively; Rhat is a measure of model convergence where values much greater than one indicate a failure to converge; Eff.Sample is the effective number of independent samples drawn by the Monte Carlo Markov Chain (MCMC) after adjusting for autocorrelation (See Bürkner, 2017 more complete overview).

In the table, Estimate and Est.Error are functionally equivalent to the unstandardized coefficients and standard error seen in non-Bayesian multiple regression. The upper and lower credible intervals presented are analogous to frequentist confidence intervals but are based on different assumptions of the data. A 95% confidence interval can be summarised by the statement "in 100 experiments, it can be reasonably expected that 95 of the 100 confidence intervals will include the true value of a given parameter Y" thus, the confidence intervals are random intervals within which the true (fixed) value of Y falls. This is different to a credible interval which considers there to be no one true value of Y and can be summarised as "a probability distribution centred on the Estimate in which 95% of values fall within the credible interval". Based on this definition, any credible interval which does not include zero can be interpreted as being a meaningful predictor in the model – since 95% of values drawn from the posterior distribution fall exclusively above or below zero. Also shown in the table are Rhat, which represents a comparison of the within- and between-chain parameter estimates to assess model convergence and Eff.Sample which shows the effective number of independent samples drawn by the Monte Carlo Markov Chain after adjusting for autocorrelation. In a perfect, uncorrelated model these values would be Rhat = 1 and Eff.Sample = 4000, respectively.

Also, of note is the notation [0, .01] which is used in the table to indicate that a value falls within a given range – for example, [1, 2] indicates that the value falls between zero and .01 (not inclusive). In the model summary tables, it is not appropriate to use the notation < .01 since the value of any given parameter is not bounded at zero. That is, a notation of < .01 would include all values ranging

from .0099 to -∞. Furthermore, brms output returns values truncated to two decimal places which can result in values of -.00 or .00 when such values are impossible given that they represent parameters in a probability distribution. Therefore, the notation [0, .01] more accurately represents the value as being less than .01 but greater than zero.

3.4 **DISCUSSION**

These findings represent an important development in understanding how learners interact with the distributional information in language. Firstly, the data show that participants demonstrated a sensitivity to the statistical associations of the bigrams as evinced by the difference in likelihood between the experimental and covariate models. This, in turn, suggests that learners can track these distributions within natural language and that the strength of the associations are retained and can be retrieved at a later time. The results also suggest that there is some validity in the use of existing language tasks to assess pre-learned associations. Secondly, the data suggests that there is a case for bigram frequency as a metric of distributional information with participants responding to targets from higher frequency bigrams more quickly. Finally, I was surprised to see that although the transitional probability model does represent an improvement over the baseline and covariate only models, it performs less well than the bigram frequency model. This is particularly noteworthy given the weight of evidence within the literature suggesting that this should not be the case. Moreover, including transitional probability in the model with bigram frequency also results in poorer model performance; this

implies that there is little to be gained from the diversity component of transitional probability, something we now investigate in Experiment 2.

3.5 EXPERIMENT 2: BIGRAM DIVERSITY

The design and procedure were identical to the first experiment with the exception that bigram diversity was manipulated rather than bigram frequency. The nature of bigram diversity is such that the manipulation in this experiment focuses on the prime rather than the target word of the bigram.

3.5.1 **Participants.**

The same thirty participants (24 females) as the previous experiment took part in another lexical decision task. Participants were aged between 18 and 60 years (M= 34, SD= 11.56), were recruited from within Nottingham, UK. English was the first language for all participants, with no language difficulties reported. Participants were offered research participation credits where applicable and those who scored lower than 80% on the lexical decision task (N=3) were excluded from the analysis.

3.5.2 **Materials.**

Measuring bigram diversity required examining the number of words that follow a prime word ('followers') in the BNC. For example, armed is followed by forty unique words in the BNC and therefore has forty followers and a bigram diversity of 40. The stimulus-list for experiment two comprised of ninety

bigrams and ninety non-word pairs (non-words paired with a real-word prime) which were selected from the same stimulus pool as in the first experiment and organised into high, low, and zero diversity items (defined as words with no followers in the BNC). Since I am not aware of any studies previously using bigram diversity as a measure, these levels were based on similar values used in studies of word frequency effects. Levels of diversity were compared using independent-samples t-tests and balanced to not differ significantly on word frequency, concreteness, number of letters, and phonemes, all *p*s $> .05$ with the following exceptions: The high diversity list differed significantly from both the low and zero diversity list on both concreteness (low: $p < .001$, zero: $p < .001$) and number of letters (low: $p = .03$, zero: $p < .01$); this is due to the unusual nature of the words in the low and no diversity condition as well as the theoretical decision to prioritise controlling individual word frequency as the largest predictor of word recognition performance (Brysbaert & New, 2009; ; Ferrand et al., 2010; Keuleers, Diependaele, & Brysbaert, 2010; Keuleers, Lacey, Rastle, & Brysbaert, 2012; Yap & Balota, 2009). Non-words were generated by transposing the middle letters of the target items from the bigram list. Both word and non-word targets were between three and eight letters long. Though these categories were not used in the analysis, descriptive metrics are included here for illustrative purposes (See table 3.7). Bigram frequency was not controlled across stimuli since attempting to do so resulted in a prohibitively small stimulus-pool, as such bigram frequency was free to vary across items. Example stimuli are displayed in table 3.8.

Table 3.7: Group means and standard deviation (in parenthesis) for High, Low, and Zero diversity bigrams.

Level	Bigram_diversity	Bigram_frequency	Target_frequency	Letters	Concreteness	Transitional_probability
High Diversity	99.08 (3.02)	153.64 (2.66)	1484.81 (1472.42)	5.54 (1.14)	2.92 (.95)	0.002 (1.56)
Low Diversity	1.31 (2.01)	18.93 (2.05)	1525.79 (1553.49)	4.97 (.87)	3.91 (1.06)	0.13 (1.07)
Zero Diversity	.00 (.00)	.00 (.00)	834.00 (129.87)	4.57 (.97)	3.90 (0.71)	.00 (.00)
Log-transformed values						
High Diversity	4.61 (1.11)	5.05 (.98)	7.30 (7.29)	-	-	-6.45 (.44)
Low Diversity	0.26 (.70)	2.95 (.72)	7.33 (7.35)	-	-	-2.03 (.07)
Zero Diversity	-13.82 (.00)	-13.82 (.00)	6.73 (4.87)	-	-	-13.82 (.00)

A constant of .000001 was added to the zero values before transformation

Table 3.8: Example low diversity stimuli for Experiment 2 including descriptive statistics for bigrams and target words.

Prime	Target	Group	Bigram_Diversity	Bigram_Frequency	Transitional_Probability	Letters	Concreteness	Word_Frequency
that	place	High Diversity	5,217.00	367.00	< .001	5.00	3.48	48,651.00
this	ancient	High Diversity	2,909.00	84.00	< .001	7.00	2.04	5,083.00
they	beat	High Diversity	1,616.00	226.00	< .001	4.00	3.97	5,675.00
will	appeal	High Diversity	1,128.00	117.00	< .001	6.00	1.73	11,002.00
very	deep	High Diversity	987.00	146.00	< .001	4.00	3.38	10,700.00
could	happen	High Diversity	790.00	350.00	< .001	6.00	1.78	8,760.00
must	keep	High Diversity	448.00	230.00	< .001	4.00	2.37	27,813.00
common	object	High Diversity	269.00	47.00	< .001	6.00	3.66	6,325.00
support	machine	High Diversity	208.00	48.00	> .001	7.00	4.25	8,938.00
heart	failure	High Diversity	120.00	107.00	0.01	7.00	2.08	7,763.00
beady	eyes	Low Diversity	2.00	21.00	0.38	4.00	4.85	29,706.00
downright	rude	Low Diversity	2.00	12.00	0.04	4.00	2.52	985.00
volt	meter	Low Diversity	1.00	11.00	0.09	5.00	4.70	487.00
snare	drum	Low Diversity	1.00	10.00	0.10	4.00	4.96	985.00
polyp	group	Low Diversity	1.00	10.00	0.11	5.00	4.12	41,547.00
hallowed	ground	Low Diversity	1.00	11.00	0.08	6.00	4.77	16,200.00
worldly	goods	Low Diversity	1.00	19.00	0.08	5.00	4.26	10,142.00
carat	gold	Low Diversity	1.00	71.00	0.50	4.00	4.81	7,792.00
cadence	design	Low Diversity	1.00	10.00	0.19	6.00	3.27	12,939.00
marbled	effect	Low Diversity	1.00	10.00	0.09	6.00	1.80	23,361.00
abaya	digest	Zero Diversity	0.00	0.00	0.00	6.00	3.07	475.00
canorous	pony	Zero Diversity	0.00	0.00	0.00	4.00	4.90	710.00
inunct	gall	Zero Diversity	0.00	0.00	0.00	4.00	2.60	1,150.00
panics	yards	Zero Diversity	0.00	0.00	0.00	5.00	4.82	3,678.00
effable	heat	Zero Diversity	0.00	0.00	0.00	4.00	3.79	5,957.00
panurgic	stab	Zero Diversity	0.00	0.00	0.00	4.00	4.07	428.00
uniped	dash	Zero Diversity	0.00	0.00	0.00	4.00	3.39	758.00
logomachy	soot	Zero Diversity	0.00	0.00	0.00	4.00	4.61	196.00
omophagy	sigh	Zero Diversity	0.00	0.00	0.00	4.00	3.89	1,171.00
wanker	away	Zero Diversity	0.00	0.00	0.00	4.00	2.23	38,747.00

3.5.3 **Procedure.**

The procedure was identical to that of Experiment 1. Participants were once again presented with letter strings and asked to decide whether they were observing a real English word or a non-word. They were then instructed to press either 'z' or 'm' on a standard QWERTY keyboard to indicate their decision. Key mapping was systematically varied so that odd numbered participants used 'z' to indicate a word and 'm' to indicate a non-word whilst even numbered participants were required to press 'm' for words and 'z' for non-words. Strings were presented until a response was made or for 3000ms if no response was made. Prime-target pairs mapped exactly onto bigrams from the stimuli lists whereby the first word of the bigram acted as a prime for the second word. A fixation point was presented in the centre of the screen for 500ms prior to both the prime – which remained on screen for 250ms - and target words. Stimuli were organised into two counterbalanced blocks and trial order was randomised within each block.

3.5.4 **Results**

Accuracy was comparable for both word and non-word trials, with all participants scoring over 80% on both. Data from experiment two was trimmed and analysed using the same procedure as the first experiment, a total of 2.04% of correct trials were removed (this did not change the pattern of results). All response time data were log-transformed; response times for each participant were then analysed using Bayesian multi-level regression. Individual trial data

(N = 2170) was used to predict log-transformed response times in a lexical decision task using random-intercept models. Individual participants and items were included as group-level effects. Bigram diversity and transitional probability were included as population-level effects, both individually (Models A & B) and in conjunction (Model C). Target-word frequency, concreteness, target-word length, and participant age were also included as covariates. Leave-one-out cross-validation statistics were used to compare model fit, with smaller values considered indicators of goodness-of-fit. Log-transformed values were used for bigram diversity, word frequency, transitional probability and response time; a constant of one was added to all values to avoid errors resulting from items with values equal to zero.

3.5.5 **Data preparation.**

Data was read into R and analysed in the same manner as Experiment 1, figure 3.4 shows the correlations between predictors. The Bigram diversity, transitional probability, and response time variables were log-transformed prior to the analysis; a small constant was added to all the values to avoid errors resulting from trying to calculate log(0). Descriptive statistics for each of the variables are shown in table 3.9.

```
df2 <- read_csv("Exp2_data.csv) ggpairs(data = df2, columns =
    c(5:6, 8:9, 14)) +  theme(panel.grid = element_blank())
df2$word_freq <- log(df2$word_freq + 1)
df2$diversity <- log(df2$diversity + 1)
df2$trans_prob <- log(df2$trans_prob + 1)
df2$response_time <- log(df2$response_time + 1)
```

Table 3.9: Descriptive statistics for Experiment 2 including means standard deviations (SD), and inter-quartile range (IQR)

Variable	Mean	SD	Min	Max	Range	IQR
age	33.780	11.250	18.000	60.000	42.000	22.000
bigram_freq	87.000	172.000	0.000	1071.000	1071.000	80.000
concreteness	3.600	1.010	1.660	5.000	3.340	1.660
word_freq	146.080	149.790	1.820	493.850	492.030	191.860
letters	5.030	1.050	3.000	8.000	5.000	2.000
response_time	684.980	227.150	234.000	1488.000	1254.000	261.000
trans_prob	0.070	0.140	0.000	0.570	0.570	0.080

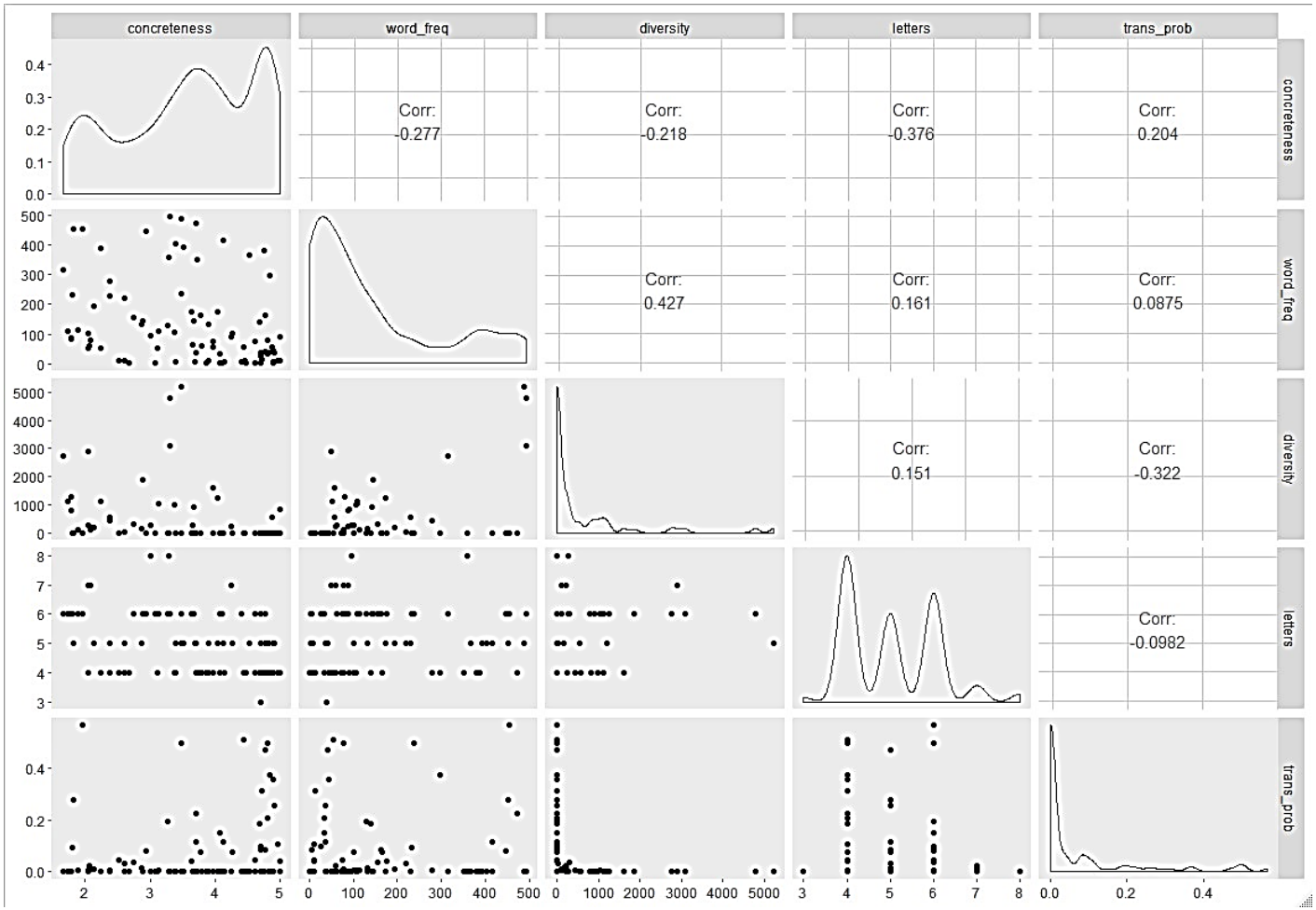

Figure 3.4: Correlation matrix for Experiment 2. Correlation coefficients show that there is no multicollinearity between the predictors.

3.5.6 **Specifying the models**

Models[3] were run in the same way as Experiment 1 and consist of a baseline, covariate and three experimental models. Models A, B, and C examined bigram diversity, transitional probability, and both variables respectively; all models included participant age, target word frequency, concreteness, and number of letters as population-level effects and participant and item as group-level effects.

```
base_model_2 <- brm(response_time ~ 1, data = df2, save_all_pars =
    TRUE, silent = TRUE, refresh = 0)
cov_model_2 <- brm(response_time ~ age + concreteness + letters +
    word_freq, data = df2, save_all_pars = TRUE, silent = TRUE,
    refresh = 0)
model_2a <- brm(response_time ~ diversity + age + concreteness +
    letters +  word_freq + (1 | subject) + (1 | item), data = df2,
    save_all_pars = TRUE, silent = TRUE, refresh = 0)
model_2b <- brm(response_time ~ trans_prob + age + concreteness +
    letters +  word_freq + (1 | subject) + (1 | item), data = df2,
    save_all_pars = TRUE, silent = TRUE, refresh = 0)
model_2c <- brm(response_time ~ diversity + trans_prob + age +
    concreteness  + letters + word_freq + (1 | subject) + (1 | item),
    data = df2,  save_all_pars = TRUE, silent = TRUE, refresh = 0)
```

[3] Throughout this document I will be referring to specific statistical models using the term Model (e.g., Model A) and models more generally using the uncapitalized model (e.g., model comparison).

3.5.7 Cross-validation

Model comparison was performed using leave-one-out cross-validation with the loo() function in R. Information criteria for all the models are displayed in Table 3.10.

```
cv_base2 <- loo(base_model_2)
cv_cov2 <- loo(cov_model_2)
cv_m2a <- loo(model_2a)
cv_m2b <- loo(model_2b)
cv_m2c <- loo(model_2c)
```

Table 3.10: Leave-one-out cross-validation statistics for Experiment 2. Also shown are the population- and group-level predictors for each statistical model

Model	Population-level	Group-level	LOOIC (SD)
Base	None	None	957 (65.1)
Covariate	Age, letters, word frequency, concreteness	participant, item	-367.1 (75.3)
Model A	Age, letters, word frequency, concreteness, bigram diversity	participant, item	-354.9 (75.3)
Model B	Age, letters, word frequency, concreteness, transitional probability	participant, item	-258.2 (64.4)
Model C	Age, letters, word frequency, concreteness, bigram diversity, transitional probability	participant, item	-247.7 (64.3)

Interestingly, cross-validation shows that the covariate only model demonstrates better predictive accuracy than both the experimental and baseline models suggesting that the inclusion of bigram diversity and/or transitional probability in these models is detrimental to predictive accuracy. However, all three experimental models fall within +/- 1.96 times the standard error of the covariate model, making any conclusions unreliable for the observed data. Therefore, as in Experiment 1, Bayes factors were used to help further differentiate between the models.

3.5.8 **Bayes factors**

Bayes factors were used for model comparison and can be seen in Table 3.11.

```
bf_covbase2 <- bayes_factor(cov_model_2, base_model_2, silent = TRUE)
bf_2abase <- bayes_factor(model_2a, base_model_2, silent = TRUE)
bf_2bbase <- bayes_factor(model_2b, base_model_2, silent = TRUE)
bf_2cbase <- bayes_factor(model_2c, base_model_2, silent = TRUE)
bf_acov2 <- bayes_factor(model_2a, cov_model_2, silent = TRUE)
bf_bcov2 <- bayes_factor(model_2b, cov_model_2, silent = TRUE)
bf_ccov2 <- bayes_factor(model_2c, cov_model_2, silent = TRUE)
bf_2ba <- bayes_factor(model_2b, model_2a, silent = TRUE)
bf_2ca <- bayes_factor(model_2c, model_2a, silent = TRUE)
bf_2cb <- bayes_factor(model_2c, model_2b, silent = TRUE)
```

Table 3.11: Bayes factors showing comparisons between statistical models for Experiment 2

Model	Base	Covariate	A (Bigram diversity)	B (Transitional probability)
Covariate	>999			
A (Bigram diversity)	>999	<.001		
B (Transitional probability)	>999	<.001	>.001	
C (Combined)	>999	<.001	>.001	>.001

Surprisingly, the covariate only model is more likely than all the experimental models. This suggests that bigram diversity does not influence response times in a lexical decision paradigm. However, the inclusion of transitional probability further reduces the likelihood of the model, this is unexpected given the wealth of evidence

suggesting that transitional probability is associated with statistical learning performance.

3.5.9 **Model summary**

Experiment 2 showed that both bigram diversity and transitional probability were ineffective at predicting response times. Looking at the covariate model (Table 3.12) however, concreteness and word frequency are negatively associated with response times - i.e., as concreteness and word frequency increase, response time decreases - whereas age and number of letters are positively associated. It is well documented in the lexical decision literature that these three covariates have a reliable effect on speed of word recognition (Murray & Forster, 2004; New, Ferrand, Pallier, & Brysbaert, 2006; Yap & Balota, 2009). Given that we see no further benefit of the experimental variables it can be suggested that, based on these data, bigram diversity and transitional probability provide no benefit in facilitating word recognition speed.

```
summary(cov_model_2)
```

Table 3.12: Summary of the covariate only model for Experiment 2

-	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Population-level effects:						
Intercept	6.35	0.06	6.23	6.46	3,175.00	1.00
Age	0.01	[.00, .01]	[.00, .01]	0.01	4,000.00	1.00
Concreteness	-0.01	0.01	-0.03	[-.01, .00]	3,329.00	1.00
Letters	0.02	0.01	0.01	0.03	3,350.00	1.00
Word frequency	-0.03	0	-0.04	-0.02	3,928.00	1.00
Family specific parameters:						
Sigma	0.29	[.00, .01]	0.28	0.3	2,814.00	1.00

Bayesian multi-level models are interpreted in much the same way as frequentist multi-level models with Estimate and Est.Error replacing the unstandardized coefficients ß and Std.Error respectively; Rhat is a measure of model convergence where values much greater than one indicate a failure to converge; Eff.Sample is the effective number of independent samples drawn by the Monte Carlo Markov Chain (MCMC) after adjusting for autocorrelation (See Bürkner, 2017 more complete overview).

3.6 **DISCUSSION**

Results from Experiment 2 suggest that there was no effect of either bigram diversity or transitional probability on participant response times. This was unexpected since there is a well-documented effect of transitional probability in statistical learning tasks; it is possible however, that the effect of transitional probability is too small to be detected by the lexical decision task, though this seems unlikely given that Experiment 1 was able to effectively identify a statistical priming effect. Similarly, there is a convincing amount of evidence that contextual diversity influences performance in these types of task. If, in fact, bigram diversity was acting as a measure of predictability we would expect to see a positive relationship with response time since more predictable transitions should elicit quicker responses. Conversely, if bigram diversity is more akin to contextual diversity in its relationship with response time then we would predict faster

response times for more diverse items. Given that we see neither effect here it is tempting to conclude that bigram diversity is not a meaningful metric of distribution within a natural language stimulus-set.

Finally, since it was not controlled across the different levels of bigram diversity it could be argued that bigram frequency constitutes a potential confound in this experiment since the high diversity stimuli also have a higher bigram frequency than the low and zero diversity stimuli. However, if this were the case then we would expect to see an effect of bigram diversity like that seen for bigram frequency in Experiment 1. Given that we see no effect of bigram diversity, we can reasonably rule out any confounding effect of bigram frequency.

3.7 GENERAL DISCUSSION

Throughout this chapter I have presented two experiments designed to assess the plausibility of using lexical decision to examine statistical learning in a large natural language stimulus set. Experiment 1 demonstrates that it is possible to use statistical priming to reduce response time, implying that participants can tap into previously learnt associations within their natural language. Moreover, Experiment 1 highlighted that bigram frequency constitutes a better predictor of task performance that the more commonly utilised transitional probability. This is a surprising result which questions how learners are developing these statistical relationships during language acquisition.

As we can see, the most effective model at predicting word recognition speed suggests that there is a small, but non-trivial, contribution to task performance for bigram frequency. Furthermore, as we would expect, individual word frequency and concreteness also facilitate lexical identification. That is, more common items, with easily identifiable referents, are more quickly identified as real English words than more nebulous, and less frequent ones. This is reflected in both the word recognition and memory literature which suggests that participants perform better with both concrete words (e.g., de Groot, 1989; de Groot & Keijzer, 2008; Kanske & Kotz, 2007; Zhang, Guo, Ding, & Wang, 2006) and those with a higher frequency (e.g., Brysbaert, Mandera, & Keuleers, 2017; Morrison & Ellis, 1995; Rayner & Duffy, 1986). Also of note is the fact that word length (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; New, Ferrand, Pallier, & Brysbaert, 2006; O'Reagan & Jacobs, 1992) and participant age (e.g., Houx, Jolles, & Vreeling, 1993; Wingfield, Lindfield, & Goodglass, 2000) were demonstrated to be inhibiting factors to word recognition in the wider literature.

As previously discussed, bigram frequency has been dismissed due to its potential correlation with frequency - the claim being that higher frequency words will co-occur more frequently by virtue of being more common. However, when word frequency (and other common covariates) is partialled out, as in the current study, there is still an identifiable effect of bigram frequency that can be presumed to be independent of the more widely recognised word frequency effect. This, if nothing else, demonstrates that more consideration needs to be given to alternative measures of statistical regularity.

These experiments not only provide proof of concept but also go some way towards addressing several the criticisms I laid out in the previous chapter. Firstly, they demonstrate that statistical learning theory may be applicable to natural languages whereas previously studies have focused on small-scale artificial grammars with unrealistic distributional statistics - transitional probability is particularly vulnerable to inflation in smaller stimulus sets, which may explain its lack of impact in these experiments. Moreover, Siegelman et al. (2018) demonstrated that it is impossible to examine statistical learning without introducing own-language biases into the task. By working within a natural language stimulus-set it is possible to eliminate this bias as a confound by examining it directly.

Experiment 1, particularly, demonstrates the lack of impact from transitional probability when bigram frequency is considered. It is suggested that any effect of transitional probability in previous studies may be the result of over inflation or could potentially be masking a frequency effect. That being the case, it can be suggested that since bigram frequency and transitional probability differ primarily on their predictive weighting there is little to be gained from the diversity component of the latter. This is something we see again in Experiment 2 which shows no effect of either bigram diversity or transitional probability. That said, it is surprising to see no effect of transitional probability since we would expect the frequency component of transitional probability to facilitate recognition speed even in the absence of a diversity effect. This prompted a review of the experimental procedures to isolate any potential confounds that may explain this

seeming lack of sensitivity to the statistical priming effect. As such, a few potential methodological issues were identified.

Firstly, some of the zero-diversity stimuli in the second experiment were proving problematic. Since the primes, by necessity, did not appear within the BNC (e.g. abaya, canorous, glabrous, hodiernal) there was some question as to whether participants might be distracted or confused by their unusual nature, this is consistent with work by Diependale, Brysbaert, and Neri (2012) who found that responses in lexical decision are more reliable when the stimuli are known to the participants. It was therefore decided to re-run experiment two using a slightly modified stimulus list which replaced these words with more familiar ones.

Additionally, in Experiment 1 there was a large amount of variance stemming from the individual target items; in an attempt to address this, a further experiment was conducted in which the target words were held constant across the high, low, and zero frequency bigrams in such a way that participants see three separate trials where they are asked to identify the same target word after viewing a different prime. For example, one set of high, low, and zero frequency items might be: steam engine, port engine, and mouse engine. Although this has the potential to introduce practice effects - since participants will already have been exposed to the target word in previous trials - it should go some way towards reducing the inter-item variance seen in the first two experiments.

Holding the words constant across all three levels of either bigram frequency or bigram diversity was deemed necessary since, in the experiments detailed in this chapter, I was unable to completely control for variation in key metrics such as concreteness, number of letters, and individual word frequency. Although steps

were taken to minimize group differences on these dimensions significant variation was observed for the number of letters in the target word (Experiment 1 & 2), and concreteness ratings (Experiment 2). On top of this, bigram frequency and diversity were free to vary across stimuli due to the concern that restricting them further would result in a prohibitively small stimulus-list and therefore an unreasonably small number of experimental trials. As previously noted, the original experiment was designed with an eye to comparing bigram frequency and diversity across different categorical groups; in that case, an ideal scenario would be to match all stimuli perfectly on all other dimensions in order to isolate the effects of bigram frequency and diversity from those of the covariates. This turned out to be impossible even in the relatively large sample of the BNC and so, as a compromise, the target items were held constant across the three levels but varied amongst themselves within each level; this would have allowed a direct comparison of individual items from each level and represented a purer test of differences between the groups. However, the change of statistical analysis from one-way ANOVA to multi-level model allowed for a better account of the covariates and the isolation of individual effects within the model.

In Chapter 4 I present an attempt to replicate the results from the first two experiments whilst also attempting to address the aforementioned limitations.

CHAPTER SUMMARY

Over the course of this chapter I have shown that:

- Statistical learning effects are still present in large-scale naturalistic stimulus-sets
- Lexical decision tasks may be an appropriate method for evaluating statistical learning in naturalistic language stimuli
- Bigram frequency may represent a better frequency-based metric of statistical learning than transitional probability in word recognition performance
- Predictability, as represented by bigram diversity and transitional probability, does not appear to influence response times in lexical decision
- There are several potential methodological issues with the current experiments that need to be addressed to improve the reliability of the data

4 ADDRESSING METHODOLOGICAL LIMITATIONS

CHAPTER OVERVIEW

Over the course of the coming chapter I aim to:

- Build on the findings of Chapter 3 by addressing the methodological limitations highlighted therein
- Assess the replicability of the statistical learning effects shown in the previous experiments using Bayesian multi-level modelling
- Test the fit of statistical models of task performance using leave-one-out cross-validation and Bayes factor comparison
- Detail the most likely model for both Experiments

4.1 PREPARATION

The following code excerpt initialises the packages necessary to run the analyses in this chapter and introduces some global settings in the interest of reproducibility.

```
library(formatR)
library(readr)
library(brms)
library(GGally)
Set.seed(100)
```

4.2 EXPERIMENTS

In the previous chapter I presented two experiments with the aim of establishing whether bigram frequency and bigram diversity constitute a useable distributional statistic for predicting learning in a natural language dataset. Experiment 1 successfully identified a priming effect for bigram frequency that surpasses that demonstrated by the more commonly used transitional probability. This demonstrates that participants were able to access representations of the statistical associations within the bigrams - though, as discussed previously, the exact form these representations take is yet unknown - and use them to improve task performance.

However, there was a large amount of variation in performance across both individual items and between participants. This is a recurrent problem in psycholinguistics that cannot be addressed by simply balancing the words on any number of dimensions (e.g., word frequency or concreteness). The language-as-fixed-effects fallacy (Clark, 1973) suggests that, even when perfectly balanced, two words may differ qualitatively and experientially by participant. In fact, the only way we can be certain that each trial is qualitatively identical to another is to use exactly the same word as the stimulus for all trials in a given experiment; this, however, is not possible in the current work since asking participants to make a lexical judgement on the same word for every trial invalidates the task somewhat. I therefore attempt to find a compromise in these experiments by holding the stimulus constant across different levels of bigram frequency and diversity whilst still allowing them to vary within levels.

Experiment 3 is a conceptual replication of the first experiment with an identical procedure but with the target stimuli repeated with a high- , low-, and zero-frequency prime so that each stimulus is seen three times during the experiment, with a different prime each time.

Experiment 2 suffered from an additional complication in that there may have been some confusion over the more abstruse stimuli (e.g., canorous, zoolatry, jumentous) which may have led to a reduction in performance. It was therefore decided that a replication of the experiment should be completed using more familiar words for the zero-diversity items to improve the reliability of the data (Diependale et al., 2012).

4.3 **EXPERIMENT 3**

Experiment three is a conceptual replication of Experiment 1 in which each target appears three times with different primes over the course of the experiment.

4.3.1 **Participants**

Fifty participants (6 Male) aged between 18 and 60 years (M= 21.49, SD= 7.96) were recruited from within Nottingham, UK; all participants reported English as their first language and reported having no language difficulties. All participants responded correctly to at least 80% of lexical decision trials; research participation credits were offered for participation where applicable.

4.3.2 **Materials**

The experimental stimuli consisted of ninety bigrams and ninety non-word stimuli between three and eight letters long. Non-word stimuli were created using entries from the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002) selected so that only non-words with legal orthographic structures (in English) were used (e.g., THRIFF, DRANNS, SNARFED). Non-words were then paired with a real word prime chosen pseudo-randomly from the BNC - primes could not be chosen completely randomly since they were constrained so as not to appear more than once across the two experiments.

For each item, descriptive metrics comprising frequency (Leech et al., 2001), concreteness (Brysbaert et al., 2014), and number of letters for the target words in each bigram; bigram diversity was also calculated but was free to vary across stimuli and not used in the initial analysis. The bigrams used in the experiment were selected to include an equal number of high, low, and zero frequency items; group descriptive statistics are detailed in table 4.1 and examples from each are given in Table 4.2. Note that although the individual words in the zero frequency bigrams do not appear together in the British National Corpus the first word of the bigram still occurs with other items in the corpus. Although these bigrams have a frequency of zero, the bigram diversity is derived solely from the initial word of the bigram and is therefore included in the table of group descriptive statistics.

Table 4.1: Group descriptive statistics for Experiment 3, values are given on a natural logarithmic scale where such was used in the analysis

Level	Bigram_frequency	Bigram_diversity	Target_frequency	Letters	Concreteness	Transitional_probability
High Frequency	420.53 (3.44)	191.31 (4.20)	690.30 (1977.90)	5.21 (1.40)	3.11 (1.05)	0.02 (2.16)
Low Frequency	10.41 (1.15)	59.59 (4.74)	690.30 (1977.90)	5.21 (1.40)	3.11 (1.05)	0.001 (1.99)
Zero Frequency	.00 (.00)	19.90 (4.46)	690.30 (1977.90)	5.21 (1.40)	3.11 (1.05)	.00 (.00)
Log-transformed values						
High Frequency	6.06 (1.24)	5.27 (1.44)	6.54 (7.59)	-	-	-3.82 (.77)
Low Frequency	2.35 (0.14)	4.10 (1.56)	6.54 (7.59)	-	-	-6.62 (.69)
Zero Frequency	-13.82 (.00)	3.00 (1.50)	6.54 (7.59)	-	-	-13.82 (.00)

A constant of .000001 was added to the zero values before transformation

Table 4.2: Example stimuli for Experiment 3 including descriptive statistics

Prime	Target	Group	Bigram_Frequency	Bigram_Diversity	Transitional_Probability	Letters	Concreteness	Word_Frequency
talking	about	High Frequency	5,376.00	56.00	0.40	5.00	1.77	197,116.00
legal	aid	High Frequency	981.00	89.00	0.07	3.00	3.10	8,607.00
can	assure	High Frequency	259.00	768.00	0.00	6.00	2.43	971.00
most	basic	High Frequency	268.00	371.00	0.00	5.00	2.26	11,227.00
then	becomes	High Frequency	143.00	690.00	0.00	7.00	1.61	7,718.00
down	beside	High Frequency	262.00	235.00	0.00	6.00	2.59	5,793.00
small	bowel	High Frequency	198.00	384.00	0.00	5.00	4.53	1,253.00
due	course	High Frequency	708.00	48.00	0.10	6.00	3.82	19,694.00
basic	data	High Frequency	261.00	93.00	0.02	4.00	3.93	18,217.00
notion	that	High Frequency	720.00	10.00	0.20	4.00	1.54	1,115,382.00
second	about	Low Frequency	10.00	355.00	0.00	5.00	1.77	197,116.00
sex	aid	Low Frequency	10.00	76.00	0.00	3.00	3.10	8,607.00
will	assure	Low Frequency	11.00	870.00	0.00	6.00	2.43	971.00
full	basic	Low Frequency	10.00	212.00	0.00	5.00	2.26	11,227.00
area	becomes	Low Frequency	10.00	138.00	0.00	7.00	1.61	7,718.00
displayed	beside	Low Frequency	10.00	22.00	0.00	6.00	2.59	5,793.00
normal	bowel	Low Frequency	11.00	128.00	0.00	5.00	4.53	1,253.00
arts	course	Low Frequency	10.00	39.00	0.00	6.00	3.82	19,694.00
cost	data	Low Frequency	10.00	101.00	0.00	4.00	3.93	18,217.00
collar	that	Low Frequency	10.00	16.00	0.01	4.00	1.54	1,115,382.00
mud	about	Zero Frequency	0.00	3.00	0.04	5.00	1.77	197,116.00
theory	aid	Zero Frequency	0.00	67.00	0.01	3.00	3.10	8,607.00
outfit	assure	Zero Frequency	0.00	13.00	0.05	6.00	2.43	971.00
pop	basic	Zero Frequency	0.00	138.00	0.00	5.00	2.26	11,227.00
dialect	becomes	Zero Frequency	0.00	3.00	0.04	7.00	1.61	7,718.00
example	beside	Zero Frequency	0.00	12.00	0.03	6.00	2.59	5,793.00
partner	bowel	Zero Frequency	0.00	42.00	0.01	5.00	4.53	1,253.00
bucket	course	Zero Frequency	0.00	12.00	0.09	6.00	3.82	19,694.00
project	data	Zero Frequency	0.00	96.00	0.00	4.00	3.93	18,217.00
unity	that	Zero Frequency	0.00	79.00	0.00	4.00	1.54	1,115,382.00

4.3.3 **Procedure**

The procedure was identical to that of Experiment 1, participants were presented with letter strings and were asked to indicate whether the string constituted a real English word by pressing either ‘z’ or ‘m’ on a standard QWERTY keyboard. Key mapping was systematically varied so that half of all participants used ‘z’ to indicate a word and ‘m’ to indicate a non-word whilst half responded with ‘m’ for words and ‘z’ for non-words. Strings were presented for a maximum of 3000ms

and were preceded by a 250ms prime. All prime-target pairs mapped exactly onto bigrams from the stimuli lists whereby the first word of the bigram acted as a prime for the second word. A fixation point was displayed in the centre of the screen prior to both the prime and target words. Prime-Target pairs were presented in two counterbalanced blocks and the order of presentation for trials was randomised for each participant.

4.4 **RESULTS**

Data was trimmed to exclude incorrect responses as well as those made faster than 200ms, slower than 1500ms (Perea et al., 2016), or more extreme than three standard deviations from the participant's mean (Madan et al., 2016), following this procedure 2.29% of correct trials were removed across participants. Individual trial data (N=1828) was then analysed with Bayesian multi-level modelling using the brms package in R[4].

In addition to Bigram frequency and transitional probability, target-word frequency, concreteness, number of letters and participant age were included as covariates.

[4] MCMC sampling was achieved using the No-U-Turn Sampler (NUTS, Hoffman & Gelman, 2014) implemented in Stan (Carpenter et al., 2017) using the RStan package (Stan Development Team, 2017); each model had four chains of 2000 iterations with a burn-in of 1000 iterations; all models used half Student-t priors with three degrees of freedom. Where specified priors are expressed using the notation N(μ, σ) where μ is the mean and σ is the standard deviation of a normal distribution.

The models were compared using Leave-One-Out Cross-Validation (LOO-CV). Where this does not provide enough discrimination between the models Bayes Factors were also computed using the bayes_factor function. The analyses resulted in some extreme Bayes factor values; since the aim is to show the likelihood of one model over another it was judged enough to express these values as being > 999 or $< .001$ as applicable.

4.4.1 **Data preparation**

Data was read into R and assessed for normality; bigram frequency, transitional probability, and response time were log-transformed prior to the analysis to achieve an approximation of a normal distribution; a small constant was added to all the values to avoid errors resulting from trying to calculate log(0). Figure 4.1 shows that there are no strong correlations between the predictors.

```
df3 <- read_csv("Exp3_data.csv")
ggpairs(data = df3, columns = c(1:3, 5, 14)) + theme(panel.grid =
   element_blank())
df3$log_word_freq <- log(df3$word_freq + 1e-06)
df3$log_bigram_freq <- log(df3$bigram_freq + 1e-06)
df3$log_trans_prob <- log(df3$bigram_freq + 1e-06)
df3$log_response_time <- log(df3$response_time + 1e-06)
```

Descriptive statistics were also calculated for each of the variables in Experiment 3 and are shown in table 4.3. Included are the means, standard deviations , upper and lower values, range and inter-quartile range.

Table 4.3: Means, standard deviations (SD), range and inter-quartile range (IQR) for each of the variables in Experiment 3

Variable	Mean	SD	Min	Max	Range	IQR
bigram_freq	548.480	3495.480	0.000	40008.000	40008.000	164.000
concreteness	3.090	1.040	1.220	4.920	3.700	1.780
word_freq	69033.300	197713.930	698.000	1115382.000	1114684.000	16145.000
letters	5.260	1.430	3.000	8.000	5.000	2.000
response_time	529.840	175.890	200.000	1498.000	1298.000	185.000
age	22.290	9.440	18.000	53.000	35.000	1.000
diversity	157.290	247.720	1.000	1550.000	1549.000	147.000
trans_prob	0.050	0.130	0.000	0.860	0.860	0.040

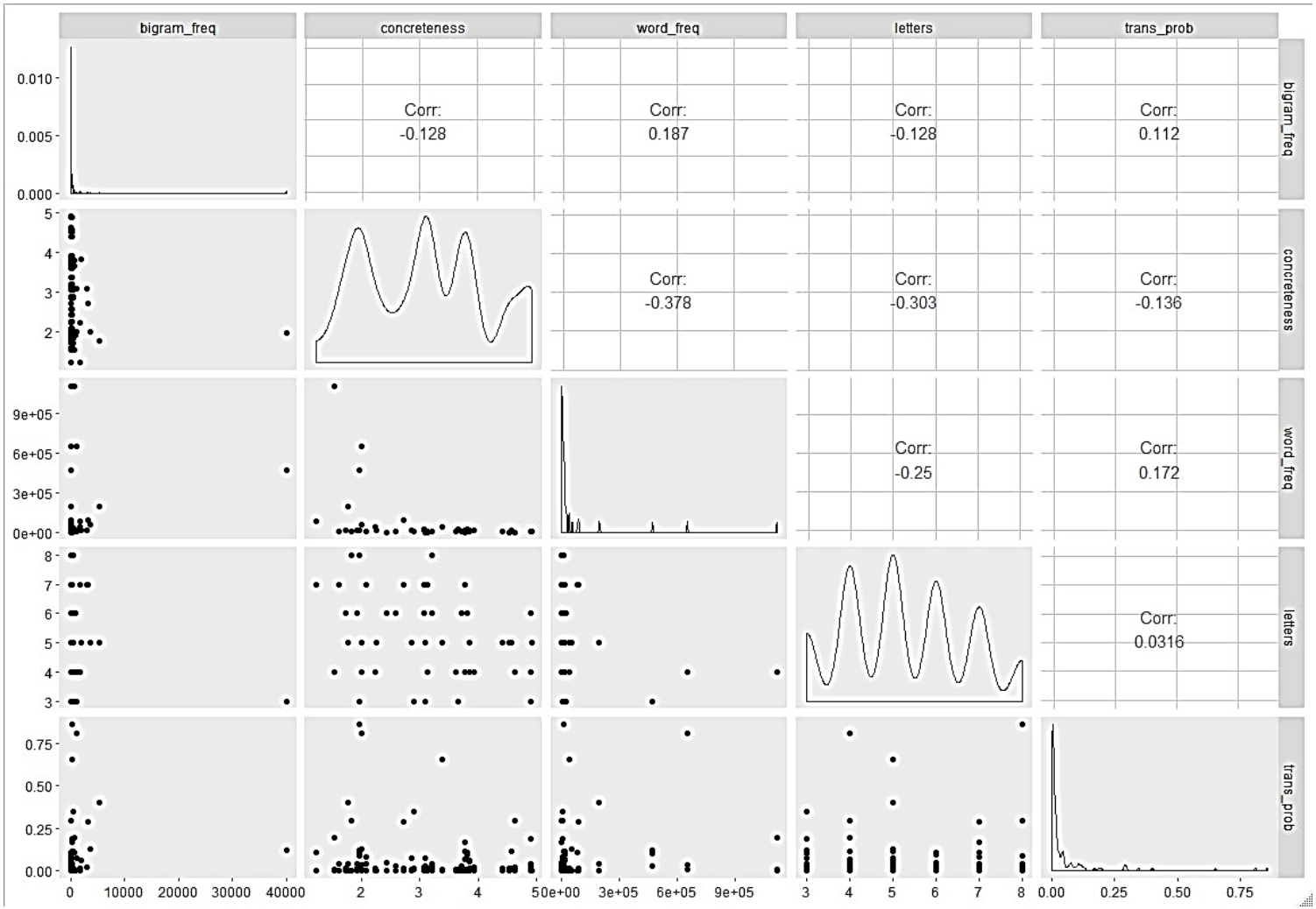

Figure 4.1: Matrix showing correlations between predictors in Experiment 3

4.4.2 **Specifying the models**

A total of five Bayesian multi-level models were used to analyse the data from Experiment 3. As in the previous chapter, baseline and covariate only models were included for comparison purposes whereas Models A and B look at the individual contribution of bigram frequency and transitional probability, respectively. Model C combines both transitional probability and bigram frequency in order to examine their relationship with response time in the lexical decision task.

```
base_model_3 <- brm(log_response_time ~ 1, data = df3,  save_all_pars =
     TRUE, silent = TRUE, refresh = 0)
cov_model_3 <- brm(log_response_time ~ age + concreteness + letters  +
   log_word_freq, data = df3, save_all_pars = TRUE, silent = TRUE,
   refresh = 0)
model_3a <- brm(log_response_time ~ bigram_freq + age + concreteness  +
   letters + log_word_freq + (1 | subject) + (1 | item), data = df3,
   save_all_pars = TRUE, silent = TRUE, refresh = 0)
model_3b <- brm(log_response_time ~ trans_prob + age + concreteness  +
   letters + log_word_freq + (1 | subject) + (1 | item), data = df3,
   save_all_pars = TRUE, silent = TRUE, refresh = 0)
model_3c <- brm(log_response_time ~ log_bigram_freq + log_trans_prob +
  age + concreteness + letters + log_word_freq + (1 | subject) +
   (1 | item), data = df3, save_all_pars = TRUE, silent = TRUE,
   refresh = 0)
```

4.4.3 **Cross-validation**

Leave-one-out cross-validation was used to compare the models in the first instance.

This resulted in a LOOIC statistic for each model; these are set out in Table 4.4.

```
cv_base3 <- loo(base_model_3)
cv_cov3 <- loo(cov_model_3)
cv_m3a <- loo(model_3a)
cv_m3b <- loo(model_3b)
cv_m3c <- loo(model_3c)
```

Table 4.4: Leave-one-out statistics for the base, covariate, and experimental models for Experiment 3

Model	Population-level	Group-level	LOOIC (SD)
Base	None	None	2471.3 (124.6)
Covariate	Age, letters, word frequency, concreteness	None	2423.2 (125.7)
Model A	Age, letters, word frequency, concreteness, bigram frequency	participant, item	275.4 (131.3)
Model B	Age, letters, word frequency, concreteness, transitional probability	participant, item	283.9 (106.8)
Model C	Age, letters, word frequency, concreteness, bigram frequency, transitional probability	participant, item	285.4 (106.7)

As we saw in Experiment 1, Model A (Bigram Frequency) once again provides the best description of task performance based on the data collected in the current Experiment. However, we also see a large amount of deviation around the mean LOOIC for all models which makes it impossible to discriminate between them using LOOIC alone. It is once again necessary to use an alternative mode of comparison in order to clarify these findings. As such, models were compared using Bayes factors which will allow us to gain ascertain the strength of evidence for one model over another.

At this point, it occurs that some justification is needed as to why Bayes factors were not chosen as the initial analysis when cross-validation was unable to provide definitive conclusions in the previous chapter – thereby requiring the use of Bayes factors to disambiguate. Raftery (1998) argues that Bayes factors should constitute "the final criterion for model comparison" (p. 412) – by which he means the ultimate criterion by which models should be assessed. However, as Liu and Aitkin (2008) point out, Bayes factors are somewhat sensitive to the chosen priors and, as such, should be interpreted with caution. This is little problem if the prior distribution accurately represents that of the estimated parameters in the models (Bernardo & Smith, 1994; Raftery & Zheng, 2003) but can result in wildly different estimates if poorly chosen. However, given enough data, the posterior distribution in any given model is less susceptible to influence from the prior. This is because the information within the observed data effectively overwhelms the prior and leads to the same conclusions regardless of our prior beliefs about the 'real' distribution. As mentioned above, the experiments in this and the previous chapter made use of uninformative prior distributions which, by definition, do not accurately mirror those found in the data and should therefore be treated with caution. Therefore, I am reticent to rely on Bayes factors alone – even ignoring the philosophical implications of using Bayes factors as a substitute for p-values – when using cross-validation will allow more robust interpretations of the data in those cases where we have clear differences in information criteria. Bayes factors then, in this case, are used as a contingency to assist with model selection if cross-validation fails to provide compelling evidence.

4.4.4 **Bayes factors**

Bayes factors were also computed using the bayes_factor() function and allow for direct comparison of the models in terms of a likelihood ratio.

```
bf_covbase3 <- bayes_factor(cov_model_3, base_model_3, silent = TRUE)
bf_3abase <- bayes_factor(model_3a, base_model_3, silent = TRUE)
bf_3bbase <- bayes_factor(model_3b, base_model_3, silent = TRUE)
bf_3cbase <- bayes_factor(model_3c, base_model_3, silent = TRUE)
bf_acov <- bayes_factor(model_3a, cov_model_3, silent = TRUE)
bf_bcov <- bayes_factor(model_3b, cov_model_3, silent = TRUE)
bf_ccov <- bayes_factor(model_3c, cov_model_3, silent = TRUE)
bf_3ba <- bayes_factor(model_3b, model_3a, silent = TRUE)
bf_3ca <- bayes_factor(model_3c, model_3a, silent = TRUE)
bf_3cb <- bayes_factor(model_3c, model_3b, silent = TRUE)
```

Table 4.5 shows the Bayes factor comparisons for the base, covariate, bigram frequency (A), transitional probability (B), and combined models (C). Values are expressed as a likelihood ratio indicating the strength of evidence for one model over another.

Table 4.5: Bayes factors for statistical model comparisons for Experiment 3

Model	Base	Covariate	A (Bigram frequency)	B (transitional probability)
Covariate	562.38			
A (Bigram frequency)	>999	>999		
B (Transitional probability)	>999	>999	>999	
C (Combination)	>999	>999	>999	0.043

Looking at the comparative Bayes factors (Table 4.5), Model B, which includes transitional probability, as well as the known covariates, item, and participant-level effects, is the most likely given the observed data. With a Bayes factor of > 999 it clearly exceeds Raftery's threshold of $BF > 150$ representing very strong evidence for this model over the others. This is contrary to the results from Experiment 1, which showed bigram frequency to be the best predictor of task performance by a similar degree. It is also worth noting that, once again, the combined model (Model C) is worse than either the transitional probability or bigram frequency models.

4.4.5 **Model summary**

A summary of Model B (transitional probability) can be seen in Table 4.6. Although the analysis shows that transitional probability is a better predictor of task performance than bigram frequency (which does not feature in the most likely model) it demonstrates a positive relationship with response time. This suggests that higher transitional probabilities may be detrimental to participant performance.

```
summary(model_3b)
```

Table 4.6: Summary statistics for Model B, the transitional probability model, values are shown on a logarithmic scale where such was used in the analysis

-	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Group-level effects						
Item	0.04	0.01	0.02	0.05	1,280.00	1.00
Participant	0.17	0.02	0.14	0.21	490.00	1.00
Population-level effects:						
Intercept	6.34	0.11	6.11	6.56	686.00	1.01
Transitional probability	0.01	0.01	[-.01, .00]	0.03	4,000.00	1.00
Age	0.00	>.01	[-.01, .00]	0.01	616.00	1.00
Concreteness	-0.01	0.01	-0.02	0.01	886.00	1.01
Letters	0.01	0.01	[-.01, .00]	0.02	1,281.00	1.00
Word frequency	-0.02	0.01	-0.03	-0.01	1,177.00	1.00
Family specific parameters:						
Sigma	0.25	>.01	0.24	0.25	4,000.00	1.00

Estimate and Est.Error are equivalent to unstandardized coefficients and Std.Error respectively; Rhat is a measure of model convergence, values much greater than one indicate a failure to converge; Eff.Sample is the effective number of independent samples drawn by the MCMC after adjusting for autocorrelation.

Once again, we see positive relationships for number of letters and participant age as well as negative associations for word frequency and concreteness as would be expected in a lexical decision task. Although the Bayes factor analysis favours the transitional probability model, transitional probability is not acting as a facilitatory factor in word recognition; in fact, higher transitional probabilities seemingly result in increased response latency. The fact that transitional probability seems to be inhibiting participant responses is particularly surprising given that the overwhelming body of evidence from statistical learning paradigms suggests that greater predictability stemming from transitional probabilities is a robust indicator of learning and should result in faster recognition of words in this task.

4.5 DISCUSSION

Experiment 3 was designed to replicate the findings of Experiment 1 whilst reducing the impact of inter-item variability. Once again, we see an overwhelming strength of evidence for the experimental models over both the baseline and covariate only models. This is heartening since it suggests that participants are sensitive to the underlying statistical regularities in the stimulus-set. However, unlike in the Experiment 1 the data does not support the bigram frequency model over the transitional probability model.

Additionally, in Experiment 1, large (relative to the other variables) effects of both item and participant were found. An attempt was made to reduce this by holding the target word constant across three levels of bigram frequency (zero, low, and high); table 4.6 shows that this was successful in reducing the inter-item variance from .06 (table 3.6) to .04.

4.6 EXPERIMENT 4

In a conceptual replication of Experiment 2, Experiment 4 utilises a lexical decision task with a revised stimulus list to examine its effect, if any, on response time in a lexical decision task. Unlike Experiment 3, the same targets were not used across levels, but the stimulus-list was updated relative to Experiment 2 to remove the more obscure items from the zero-diversity condition and replace them with more recognisable items.

4.6.1 **Participants**

Fifty participants (6 Male) aged between 18 and 60 years (M= 21.49, SD= 7.96) were recruited from Nottingham, UK; all participants reported English as their first language and having no language difficulties. Research participation credits were offered for participation where applicable. These were the same participants that took part in Experiment 3.

4.6.2 **Materials**

The stimulus-list for experiment two comprised of ninety bigrams and ninety nonword pairs (non-words paired with a real-word prime) which were identical to those in Experiment 2 apart from the zero diversity items which were changed to be more recognisable to the participants. These were once again organised into three lists of thirty high (>100), low (<50), and zero diversity items (defined as words with no followers in the BNC), descriptive statistics for which can be seen in table 4.7. Both word and non-word targets were between three and eight letters long. Bigrams were selected to include an equal number of high, low, and zero diversity items examples of which can be seen in Table 4.8. Bigram frequency was not controlled across stimuli since attempting to do so resulted in fewer than ten items in each category, as such bigram frequency was free to vary across items. None of the bigrams were repeated across the experiments.

Table 4.7: Group descriptive statistics for high, low, and zero diversity items in Experiment 4

	Level	Bigram_frequency	Bigram_diversity	Target_frequency	Letters	Concreteness	Transitional_probability
	High Diversity	376.848925	483.5146123	152.09 (120.93)	5.46 (1.09)	2.93 (.95)	0.002 (.61)
	Low Diversity	20.71221971	1.989522297	150.52 (117.47)	4.92 (.17)	3.98 (.30)	0.38 (.86)
	Zero Diversity	.00 (.00)	.00 (.00)	150.72 (116.19)	5.05 (.25)	3.19 (.34)	.00 (.00)
Log-transformed values							
	High Diversity	5.95 (1.18)	6.20 (1.51)	5.02 (4.78)	-	-	-6.24 (0.49)
	Low Diversity	3.04 (.70)	0.69 (.58)	5.01 (4.77)	-	-	-0.98 (.86)
	Zero Diversity	-13.82 (.00)	-13.82 (.00)	5.02 (4.76)	-	-	-13.82 (.00)

A constant of .000001 was added to the zero values before transformation

Table 4.8: Example stimuli for Experiment 4

Prime	Target	Group	Bigram_Diversity	Bigram_Frequency	Transitional_Probability	Letters	Concreteness	Word_Frequency
that	place	High Diversity	2,074.00	367.00	0.00	5.00	3.48	48,651.00
were	almost	High Diversity	1,170.00	600.00	0.00	6.00	1.66	31,588.00
will	gain	High Diversity	870.00	151.00	0.00	4.00	2.24	5,218.00
could	happen	High Diversity	617.00	350.00	0.00	6.00	1.78	8,760.00
very	deep	High Diversity	542.00	146.00	0.00	4.00	3.38	10,700.00
even	among	High Diversity	372.00	186.00	0.00	5.00	2.38	22,864.00
common	object	High Diversity	155.00	47.00	0.00	6.00	3.66	6,325.00
total	lack	High Diversity	152.00	80.00	0.00	4.00	2.04	10,068.00
clearly	defined	High Diversity	128.00	363.00	0.02	7.00	2.07	5,898.00
living	alone	High Diversity	112.00	239.00	0.01	5.00	2.86	13,265.00
strict	sense	Low Diversity	18.00	64.00	0.03	5.00	2.61	21,935.00
musty	smell	Low Diversity	2.00	16.00	0.11	5.00	3.70	3,755.00
croquet	club	Low Diversity	1.00	11.00	0.08	4.00	3.78	16,465.00
dimmer	switch	Low Diversity	1.00	11.00	0.15	6.00	4.07	3,316.00
hallowed	ground	Low Diversity	1.00	11.00	0.08	6.00	4.77	16,200.00
loony	left	Low Diversity	1.00	34.00	0.23	4.00	3.70	47,089.00
revolve	around	Low Diversity	1.00	72.00	0.57	6.00	1.96	45,286.00
shady	corner	Low Diversity	1.00	12.00	0.04	6.00	4.61	7,500.00
snare	drum	Low Diversity	1.00	10.00	0.10	4.00	4.96	985.00
whacking	great	Low Diversity	1.00	12.00	0.28	5.00	1.81	45,217.00
orate	red	Zero Diversity	0.00	0.00	0.01	3.00	4.24	15,136.00
warren	kept	Zero Diversity	0.00	0.00	0.00	4.00	2.79	14,306.00
opulent	recent	Zero Diversity	0.00	0.00	0.01	6.00	2.50	15,858.00
conflate	art	Zero Diversity	0.00	0.00	0.01	3.00	4.17	15,587.00
comely	answer	Zero Diversity	0.00	0.00	0.04	6.00	2.89	14,421.00
taxes	lead	Zero Diversity	0.00	0.00	0.03	4.00	4.10	14,555.00
sable	field	Zero Diversity	0.00	0.00	0.02	5.00	4.26	15,298.00
hoary	despite	Zero Diversity	0.00	0.00	0.04	7.00	1.33	14,592.00
detritus	works	Zero Diversity	0.00	0.00	0.91	5.00	3.79	14,528.00
ley	poor	Zero Diversity	0.00	0.00	0.16	4.00	2.70	15,125.00

4.6.3 **Procedure**

The procedure was identical to that of Experiment 2. Participants were once again asked to view letter strings as part of a primed lexical decision task and to indicate whether the string constituted a real English word or not by pressing either 'm' or 'z' on a standard QWERTY keyboard; response key allocation was varied systematically so that even numbered participants used 'm' to indicate a word and odd numbered participants were instructed to respond with 'z' if the target was a word. Stimuli were randomly presented and organised into two counterbalanced blocks containing fifteen each of high, low, and zero diversity items plus forty-five non-words. Targets remained on the screen for 3000ms, or until a response was given; each target was preceded by a 250ms prime. A fixation point was displayed at the centre of the screen before both the prime and target.

4.6.4 **Results**

All participants completed the lexical decision task with at least 80% accuracy. The data was trimmed using the same criteria as the previous experiments. Incorrect responses, responses faster than 200ms or slower than 1500ms, and outliers which fell more than three standard deviations from the participants' mean were removed. This resulted in the omission of 3.32% of the data but did not change the pattern of results. Individual trial data (N = 1981) was then used to create five random intercept multi-level Bayesian models (see below for details). Following convergence, model comparison was conducted using leave-one-out cross-validation and Bayes factors.

4.6.5 **Data preparation**

Data was read into R and analysed in the same manner as previous experiments. The Bigram frequency, transitional probability, and response time variables were log transformed prior to the analysis; a small constant was added to all the values to avoid errors resulting from trying to calculate log(0). In addition, descriptive statistics were calculated and are displayed in table 4.9.

```
df4 <- read_csv("Exp4_data.csv")

ggpairs(data = df4, columns = c(2:3, :7, 14)) +  theme(panel.grid =

    element_blank)

df4$log_word_freq <- log(df4$word_freq + 1e-06)

df4$log_bigram_freq <- log(df4$bigram_freq + 1e-06)

df4$log_trans_prob <- log(df4$bigram_freq + 1e-06)

df4$log_response_time <- log(df4$response_time + 1e-06)
```

Table 4.9: Descriptive statistics for variables in Experiment 4

Variable	Mean	SD	Min	Max	Range	IQR
bigram_freq	91.860	181.670	0.000	1071.000	1071.000	84.000
concreteness	3.360	1.080	1.330	5.000	3.670	1.880
word_freq	150.630	115.370	4.870	493.850	488.980	83.910
diversity	398.240	966.120	0.000	5217.000	5217.000	269.000
letters	5.150	1.080	3.000	8.000	5.000	2.000
response_time	559.750	175.710	200.000	1519.000	1319.000	192.000
age	20.150	4.820	18.000	34.000	16.000	0.000
trans_prob	0.010	0.060	0.000	0.570	0.570	0.000

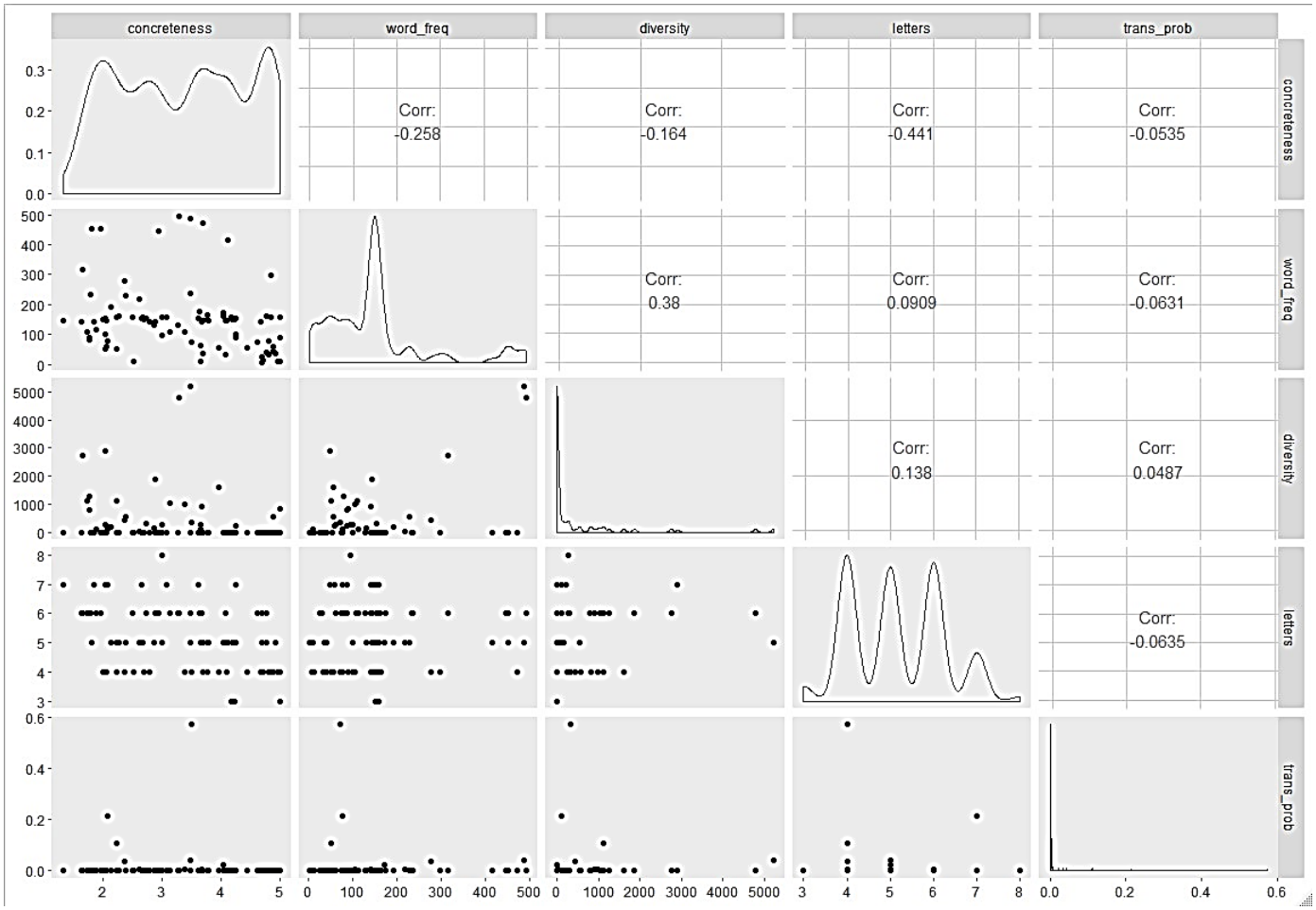

Figure 4.2: Correlation matrix for Experiment 4

4.6.6 **Specifying the models**

As with the previous experiments, models were specified and run using the default priors and settings from the brms package in R. Once again, comparisons are drawn between a baseline model, covariate only model, and three experimental models. The experimental models are parameterised with either bigram diversity (Model A), transitional probability (Model B), or both variables (Model C) alongside several widely recognised covariates. Item- and participant-level effects were also included in the experimental models but not in the baseline or covariate models.

```
base_model_4 <- brm(log_response_time ~ 1, data = df4,
    save_all_pars = TRUE, silent = TRUE, refresh = 0)
cov_model_4 <- brm(log_response_time ~ age + concreteness + letters
    + log_word_freq, data = df4, save_all_pars = TRUE, silent =
    TRUE,  refresh = 0)
model_4a <- brm(log_response_time ~ log_diversity + age +
    concreteness  + letters + log_word_freq + (1|subject) +
    (1|item), data = df4,  save_all_pars = TRUE, silent = TRUE,
    refresh = 0)
model_4b <- brm(log_response_time ~ log_trans_prob + age + concreteness
       + letters + log_word_freq + (1|subject) + (1|item), data = df4,
       save_all_pars = TRUE, silent = TRUE, refresh = 0)
model_4c <- brm(log_response_time ~ log_diversity + log_trans_prob + age
       + concreteness + letters + log_word_freq + (1|subject) + (1|item),
       data = df4, save_all_pars = TRUE, silent = TRUE, refresh = 0)
```

Model A, the bigram diversity model failed to converge with four chains of 2000 iterations (the default for brms); as such, the model was rerun with four chains of 3000 iterations resulting in full convergence.

```
model_4a <- brm(response_time ~ diversity + age + concreteness +
    letters + log_word_freq + (1|subject) + (1|item), data = df4,
    save_all_pars = TRUE, silent = TRUE, refresh = 0, iter = 3000)
```

4.6.7 **Cross-validation**

Model comparison was performed using leave-one-out cross-validation with the `loo()` function in R. Information criteria for all the models are displayed in Table 4.10.

```
cv_base4 <- loo(base_model_4)
cv_cov4 <- loo(cov_model_4)
cv_m4a <- loo(model_4a)
cv_m4b <- loo(model_4b)
cv_m4c <- loo(model_4c)
```

Table 4.10: LOOIC for the Bayesian multi-level models from Experiment 4

Model	Population-level	Group-level	LOOIC (SD)
Base	None	None	633.00 (72.7)
Covariate	Age, letters, word frequency, concreteness	participant, item	629.90 (74.0)
Model A	Age, letters, word frequency, concreteness, bigram diversity	participant, item	-32.00 (76.1)
Model B	Age, letters, word frequency, concreteness, transitional probability	participant, item	-10.20 (63.0)
Model C	Age, letters, word frequency, concreteness, bigram diversity, transitional probability	participant, item	-8.1 (62.9)

Table 4.10 shows that the bigram diversity model (Model A) is a better predictor of task performance than either the transitional probability or combined models since it

has the lowest leave-one-out information criterion. Unfortunately, it is not possible to accurately discriminate between the models based on LOOIC alone due to the large standard deviation around the criteria. To illustrate this point, if we consider a distribution centred on the LOOIC and distributed according to the standard distribution then we can infer that in 95% of cases the true value of the LOOIC for Model A – the bigram diversity model - would lie somewhere between -181.16 and 117.16 whereas the LOOIC value for Models B and C lie in the ranges of -133.68 to 113.28 and -131.38 to 115.18 respectively. Given the large amount of overlap between these ranges it would be inappropriate to base any conclusions as to which model best fits the data on LOOIC. Bayes factor comparisons were therefore chosen as an alternate method of identifying the most likely model.

4.6.8 **Bayes factors**

Bayes factors were calculated using the bayes_factor() function built into brms and used for model comparison, these can be seen in table 4.11.

```
bf_covbase4 <- bayes_factor(cov_model_4, base_model_4, silent = TRUE)
bf_4abase <- bayes_factor(model_4a, base_model_4, silent = TRUE)
bf_4bbase <- bayes_factor(model_4b, base_model_4, silent = TRUE)
bf_4cbase <- bayes_factor(model_4c, base_model_4, silent = TRUE)
bf_acov4 <- bayes_factor(model_4a, cov_model_4, silent = TRUE)
bf_bcov4 <- bayes_factor(model_4b, cov_model_4, silent = TRUE)
bf_ccov4 <- bayes_factor(model_4c, cov_model_4, silent = TRUE)
bf_4ba <- bayes_factor(model_4b, model_4a, silent = TRUE)
bf_4ca <- bayes_factor(model_4c, model_4a, silent = TRUE)
bf_4cb <- bayes_factor(model_4c, model_4b, silent = TRUE)
```

Table 4.11: Bayes factors for model comparison in Experiment 4

Model	Base	Covariate	A (Bigram diversity)	B (Transitional probability)
Covariate	0			
A (Bigram diversity)	>999	>999		
B (Transitional probability)	>999	>999	>999	
C (Combination)	>999	>999	0.01	<.001

Examination of the Bayes factors in table 4.11 shows that, given the observed data, there is convincing evidence that Model B, the transitional probability model, is better than both the other experimental models and the baseline/covariate models. A Bayes factor of more than 999 can be considered as very strong evidence for the transitional probability model over both bigram frequency and the combined model.

4.6.9 **Model summary**

Bayes factor comparisons indicate the greatest strength of evidence for Model B which includes transitional probability as a predictor as well as the concreteness, number of letters, and frequency of the target-word and participant age as covariates. Item and participant effects are also included at the group level, the full model is displayed in table 4.12.

```
summary(model_4b)
```

Table 4.12: Summary of the transitional probability model for Experiment 4.

-	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Group-level effects:						
Item	0.04	0.01	0.03	0.06	1,684.00	1.00
Participant	0.16	0.02	0.12	0.21	1,097.00	1.00
Population-level effects:						
Intercept	6.35	0.16	6.04	6.65	1,488.00	1.00
Age	0.00	0.01	-0.01	0.01	1,462.00	1.00
Concreteness	-0.01	0.01	-0.03	0.01	4,000.00	1.00
Letters	0.02	0.01	-0.03	<.01	4,000.00	1.00
Word Frequency	-0.02	0.01	-0.03	<.01	4,000.00	1.00
Transitional probability	-0.04	0.03	-0.09	0.01	4,000.00	1.00
Family specific parameters:						
Sigma	0.24	<.01	0.23	0.25	4,000.00	1.00

Estimate and Est.Error are equivalent to unstandardized coefficients and Std.Error respectively; Rhat is a measure of model convergence, values much greater than one indicate a failure to converge; Eff.Sample is the effective number of independent samples drawn by the MCMC after adjusting for autocorrelation.

4.7 DISCUSSION

Experiment 4 addressed a potential issue with some of the more uncommon zero diversity stimuli used in Experiment 2. The recondite nature of these stimuli may have had the effect of distorting participants' responses, resulting in there being no evidence of an effect for either bigram diversity or – more surprisingly – transitional probability. The new stimuli were therefore selected to be more recognisable to the average participant whilst still being absent from bigrams within the BNC. Since the data from Experiment 4 now better supports one of the experimental models over the covariate model – in this case the transitional probability model, as would be

predicted from most of the published evidence - it is not unreasonable to assume that this change resolved the issue.

I posited in Chapter 2 that the influence of bigram diversity could take one of two forms. In the first instance, one could assume that lower diversity items would result in more predictable transitions and therefore would improve response times due to greater lexical activation – a hypothesis which would be supported by the work of Conway et al. (2010) amongst others (e.g., Bates & MacWhinney, 1987; Glenberg & Gallese, 2012; Goldberg et al., 2005; Pickering & Garrod, 2004; 2007; van Berkum et al., 2005). This interpretation is akin to the explanations provided for transitional probability and, in fact, the latter better captures this facet of statistical regularity albeit at a potentially higher computational cost. Alternatively, bigram diversity could be likened to contextual diversity (Adelman et al., 2006, Adelman & Brown, 2008) where encountering items in a wider range of contexts has been shown to improve lexical decision performance. Under this hypothesis it follows that higher bigram diversity would result in improved performance.

The data from the current experiment are most effectively explained by the transitional probability model and demonstrate that higher transitional probability leads to faster response times. These findings are a direct contrast to those we would expect given a diversity hypothesis since more diverse items are necessarily less predictive than their less diverse counterparts. For example, a highly diverse item like that is followed by 2074 unique words in the BNC making it practically useless for predicting what comes next. Conversely, croquet has only one follower in the BNC and is therefore much more useful as a predictive cue.

However, the model also shows a wide credible interval for the effect of transitional probability, ranging from -.09 to .01, meaning that we cannot be acceptably confident that the true value for the effect of transitional probability is not zero (or, in fact, a positive value) and should therefore treat the findings with caution.

4.8 GENERAL DISCUSSION

The studies presented in this chapter build on those of Chapter 3 in demonstrating a statistical priming effect. However, the results are inconsistent with those of the previous experiments. Experiment 1 showed that bigram frequency was a better predictor than transitional probability whereas data from Experiment 3 suggests that the bigram frequency model is a poorer descriptor of the observed data than the transitional probability model and that higher transitional probabilities impair lexical decision performance based on Bayes factor comparisons – a surprising outcome given the overwhelming theoretical support for the metric. It should be noted, once again, that the outcome of the LOOIC and the Bayes factor comparisons are inconsistent. In this case, although Bayes factors based on non-informative priors should be interpreted with caution, cross-validation cannot provide a reasonable measure of difference in model fit. Therefore, it is necessary to base any conclusions solely on the Bayes factors with the understanding that they do not represent an ideal method of comparison. Though this leaves us with a somewhat inconclusive view of both bigram frequency and transitional probability it reinforces the need to consider the efficacy of different distributional statistics rather than accepting their pedigree at face value.

In the previous chapter I also highlighted a several potential methodological flaws with the aim of addressing them over the course of the two experiments presented here. In Experiment 1 it was noted that there was a large amount of inter-item variability. To reduce this Experiment 3 introduced target consistency across each level of bigram frequency. That is, the same target was paired with three different primes in order to form a high, low, and zero frequency trial. Examination of the final model from Experiment 3 shows that inter-item variability was reduced from .06 to .04. Although it is impossible to definitively trace this reduction to the introduction of target consistency – the experiment was conducted using different participants who may have demonstrated less bias towards particular items – it is not implausible to suggest that this is the case given that all other aspects of the design were identical.

Another issue that was addressed in the current chapter was the potential confusion arising from the zero-diversity stimuli primes in Experiment 2. In order to find primes that did not appear within the BNC it was necessary to utilise items which may not have been recognisable to the participants. It was suggested that some of the more obscure primes might have been confusing or misleading to participants and could have resulted in unreliable data. Experiment 4 was conceived as a replication of this experiment with a slightly modified stimulus-list whereby the more abstruse items were replaced by more common items which still do not appear as part of the BNC. Following this change, Experiment 4 shows a non-meaningful effect of transitional probability which tentatively supports a statistical learning strategy consistent with the predictability hypothesis discussed earlier in this work.

In summary, the methodological changes made to the experiments in this chapter were somewhat successful in addressing the aforementioned limitations. Experiment

3 demonstrated a reduction in inter-item variability whereas Experiment 4 showed a noticeable statistical priming effect with the modified stimulus-list.

However, the experiments detailed in this work so far were intended as a proof of concept study and, as such, were designed to give participants the best possible opportunity to benefit from the statistical priming effect. To this end, the experimental timings in each of the experiments were deliberately extended. Since the associations between words in a naturalistic stimulus-set are relatively weak (at least in comparison to those seen in artificial grammars) it was felt that longer presentation times may be required to ensure that the prime was consciously observed and that participants had the best possible opportunity to benefit from the statistical priming effect. This resulted in a longer display time for primes and an increased prime-target interval compared to those seen in traditional primed lexical decision tasks. However, since a statistical priming effect was successfully observed in most of the experiments, I made the decision to re-run all four experiments using more typical timings.

Ferre, Guasch, Garcia-Chico, and Sanchez-Casas (2015) used a semantic priming paradigm not dissimilar to the statistical priming paradigm used here. Participants were shown a fixation point in the middle of the screen which was replaced after 500ms by the prime-word which was displayed for 150ms rather than the 250ms used in the current experiments. Furthermore, in the four experiments covered so far, I interposed a fixation point between the prime and target words to allow participants time to fully process the prime before being exposed to the target; this is incongruent with Ferre et al.'s paradigm where the target immediately followed the prime. In retrospect, this is unrepresentative of the way in which language is encountered and may have resulted in the decay of lexical activation over time.

Similarly, in two slightly different but comparable tasks Kusonose, Hino, and Lupker (2016) and Ortells, Keifer, Castillo, Megias, and Morillas (2016) presented primes for 33 and 33.5ms, respectively. Finally, Yap, Balota, and Tan (2013) used a 150ms prime but also included a 650ms delay between prime and target. These timings demonstrate that a priming effect can be observed with a significantly lower display time than was used in the current experiment. However, 33ms was still judged to be an insufficient duration given the relatively small effect sizes in my experiments. Moreover, Adelman (2011) showed that although participants reach asymptotic lexical decision accuracy after 30ms for some prime types, comparable accuracy was not achieved for all primes until around 40ms. As such, I decided to reduce the amount of time the prime was displayed for to 75ms, this is above the threshold demonstrated by Adelman (2011) for asymptotic accuracy whilst also allowing for the relatedly small effect sizes observed thus far. Additionally, apart from Yap et al., none of the studies introduced a delay between the prime and target words. Considering this the experimental sequence was also altered so that the prime was immediately followed by the target word rather than being delayed by 500ms. These new timings should be sufficient to allow participants to consciously process the prime whilst avoiding any potential decaying of lexical activation resulting from the prime-target delay.

The experiments in this and the previous chapter have acted as proof of concept for using a statistical priming paradigm with lexical decision to investigate whether participants can use the existing statistical properties of natural language to improve task performance. Based on the data presented we can tentatively conclude that transitional probability may not be an accurate predictor of statistical learning

performance. Experiment 1 shows no effect of transitional probability whereas Experiment 3 shows an extremely surprising positive relationship between transitional probability and response time when bigram frequency is manipulated. It can also be concluded – again, rather tentatively – that transitional probability performs better than bigram diversity in predicting task performance, thus supporting a predictability hypothesis of statistical learning.

However, given that the current experiments were designed with the explicit aim of increasing the likelihood of detecting a statistical priming effect, it would be irresponsible to draw any definitive conclusions from the data collected thus far. Furthermore, the discrepancy between the current methodology and that of previous studies raises questions about the validity of the findings presented herein. Over the next two chapters I will therefore be presenting further replications of the four previously detailed experiments in order to improve the validity of my findings.

CHAPTER OVERVIEW

Over the course of this chapter I:

- Addressed the methodological limitations highlighted in the previous chapter
- Failed to exactly replicate the effects shown in the previous experiments using Bayesian multi-level modelling
- Questioned the efficacy of transitional probability as a predictor of statistical learning performance
- Provided tentative support for a predictive hypothesis of statistical learning

5 ADJUSTED TIMINGS 1

CHAPTER OVERVIEW

The aim of this chapter is to:

- Repeat Experiments 1 & 2 using timings more typically seen in primed lexical decision paradigms
- Support the findings of Experiments 1 & 2 by replicating the results using a new participant-set
- Expand the theoretical explanations of statistical regularities in lexical decision performance

5.1 PREPARATION

The following code excerpt initialises the packages necessary to run the analyses in this chapter and introduces some global settings in the interest of reproducibility.

```
library(formatR)
library(readr)
library(brms)
library(rstanarm)
library(GGally)
set.seed(100)
```

5.2 EXPERIMENTS

Over the past two chapters I have attempted to demonstrate that repurposing existing language tasks, specifically lexical decision tasks, is a valid approach to investigating statistical learning phenomena in large scale naturalistic corpora; an endeavour I feel has been mostly successful. Results from the previous experiments suggest that participants are sensitive to the statistical regularities within the British National Corpus and that they can implicitly access these to more efficiently perform an explicit discrimination task. Nonetheless, there is some discrepancy between the methodologies implemented and those more commonly used in lexical decision. Many studies use shorter display times for primes and a smaller interval between prime and target (e.g., Ferre et al., 2015; Kusonose et al., 2016; Ortells et al., 2016; Yap et al., 2013). Despite this, the evidence for an existing statistical priming effect is sufficient to highlight the suitability of the task. However, in the interest of scientific rigour, it was decided to treat the previous experiments as a proof of concept and to replicate them using more typical timings. Data from Experiments 1-8 will then be aggregated and used in a meta-analysis in Chapter 7.

5.3 EXPERIMENT 5

5.3.1 Participants

Thirty-one participants (25 females) aged between 18 and 41 years (M= 20.77, SD= 4.17) were recruited from Nottingham, UK. All participants reported English as their first language and were screened for language difficulties. Participants took part in

both Experiment 5 and 6 and received research credits in exchange for their participation where applicable.

5.3.2 **Materials**

The experimental stimuli consisted of ninety bigrams and ninety non-word stimuli between three and eight letters long. Non-word stimuli were created using entries from the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002) and only non-words between with legal orthographic structures (in English) were used. Each non-word was paired with a unique real word prime chosen pseudo-randomly from the BNC - primes were constrained to not appear more than once across the two experiments. As described in Chapter 3, a list of 12,293,349 unique bigrams were extracted from the BNC and filtered to exclude items with a frequency of less than .1 per million. Any bigrams containing acronyms, initialisations, contractions, hyphenations, non-standard or non-English words, names, numbers expressed as digits, or words with fewer than three letters were also excluded from the stimulus list. Measures of frequency (http://ucrel.lancs.ac.uk/bncfreq/flists.html), concreteness (Brysbaert et al., 2014), and number of letters for the target word were obtained for each bigram. Bigram diversity and transitional probability were also calculated but were not constrained during stimuli selection. The bigrams used in the experiment were identical to those used in Experiment 1 and were selected to include an equal number of high, low, and zero frequency items; group descriptive statistics can be seen in table 5.1 and example stimuli are displayed in table 5.2.

Table 5.1: Descriptive statistics for high, low, and zero frequency bigrams used as stimuli in Experiment 5

Level	Bigram_frequency	Bigram_diversity	Target_frequency	Letters	Concreteness	Transitional_probability
High Frequency	507.52 (1337.30)	206.09 (474.19)	143.57 (110.55)	5.57 (1.18)	3.46 (.96)	.01 (.06)
Low Frequency	11.00 (1.71)	208.62 (487.11)	143.91 (112.95)	5.56 (1.18)	3.48 (.96)	.01 (.06)
Zero frequency	.00 (.00)	203.80 (472.04)	147.46 (115.90)	5.57 (1.18)	3.46 (.96)	.00 (.00)
Log-transformed values						
High frequency	6.23 (7.20)	5.33 (6.16)	4.97 (4.70)	-	-	-4.61 (2.81)
Low frequency	2.40 (.54)	5.34 (6.19)	4.97 (4.73)	-	-	-4.61 (2.81)
Zero frequency	-13.82 (.00)	5.32 (6.16)	4.99 (4.75)	-	-	-13.82 (.00)

A constant of .000001 was added to the zero values before transformation

Table 5.2: Examples of stimuli used in Experiment 5, including descriptive statistics

Prime	Target	Group	Bigram_Diversity	Bigram_Frequency	Transitional_Probability	Letters	Concreteness	Word_Frequency
eighth	army	High Frequency	21.00	187.00	0.15	4.00	4.70	114.41
bile	acid	High Frequency	17.00	234.00	0.20	4.00	4.25	49.68
fleet	street	High Frequency	25.00	305.00	0.14	6.00	4.75	196.14
rugby	union	High Frequency	39.00	428.00	0.12	5.00	3.38	176.07
good	practice	High Frequency	850.00	461.00	0.01	8.00	2.52	171.14
daily	post	High Frequency	74.00	512.00	0.07	4.00	4.30	93.39
cash	flow	High Frequency	134.00	573.00	0.07	4.00	3.72	52.44
award	title	High Frequency	50.00	1,815.00	0.12	5.00	3.32	97.90
wide	range	High Frequency	115.00	2,745.00	0.23	5.00	3.22	204.27
make	sure	High Frequency	376.00	4,530.00	0.06	4.00	1.73	245.95
always	accept	Low Frequency	567.00	10.00	0.00	6.00	3.03	98.07
craggy	face	Low Frequency	1.00	10.00	0.08	4.00	4.87	349.78
interest	account	Low Frequency	149.00	10.00	0.00	7.00	3.08	158.91
local	access	Low Frequency	568.00	10.00	0.00	6.00	2.71	109.40
people	achieve	Low Frequency	679.00	10.00	0.00	7.00	2.29	67.68
rustic	style	Low Frequency	1.00	10.00	0.04	5.00	2.67	107.25
time	across	Low Frequency	569.00	10.00	0.00	6.00	3.07	252.03
puck	fair	Low Frequency	1.00	11.00	0.18	4.00	2.39	92.10
coiled	spring	Low Frequency	1.00	13.00	0.07	6.00	3.89	59.83
canned	food	Low Frequency	3.00	14.00	0.07	4.00	4.80	189.92
abase	number	Zero Frequency	0.00	0.00	0.00	6.00	3.30	493.85
building	food	Zero Frequency	177.00	0.00	0.00	4.00	4.80	189.92
drubs	nudge	Zero Frequency	0.00	0.00	0.00	5.00	4.47	1.53
geese	wits	Zero Frequency	4.00	0.00	0.00	4.00	1.76	4.00
lifer	hugs	Zero Frequency	0.00	0.00	0.00	4.00	4.14	1.03
oval	hipster	Zero Frequency	9.00	0.00	0.00	7.00	2.50	190.60
rethinks	scaly	Zero Frequency	0.00	0.00	0.00	5.00	4.22	0.75
snuffles	model	Zero Frequency	0.00	0.00	0.00	5.00	4.53	133.35
tides	mauve	Zero Frequency	4.00	0.00	0.00	5.00	4.00	2.22
way	agree	Zero Frequency	263.00	0.00	0.00	5.00	2.31	81.81

5.3.3 **Procedure**

Participants were presented with a real-word prime drawn from the initial position of a bigram (with half of the bigrams being word pairs with the bigram being zero, low, or high frequency; and half being a word-nonword pair). The prime (first word of the bigram) remained on the screen for 75ms before being immediately replaced with the target (second word of the bigram); the target was presented for a maximum of 1500ms during which time participants were required to press either 'z' or 'm' on a standard QWERTY keyboard; key mapping was systematically varied so that half of all participants used 'z' to indicate a word and 'm' to indicate a non-word whilst half responded with 'm' for words and 'z' for non-words. A fixation point was presented in the centre of the screen for 500ms prior to each trial. Prime-Target pairs were presented in two blocks each containing forty-five bigram trials – comprised of equal numbers of high, low and zero frequency items - and forty-five non-word trials. The order in which the blocks were presented was counterbalanced and individual trials were randomised for each participant.

5.4 **RESULTS**

5.4.1 **Data preparation**

Data was read into R and assessed for normality; bigram frequency, transitional probability, and response time were log-transformed prior to the analysis to achieve an approximation of a normal distribution; a small constant was added to all the

values to avoid errors resulting from trying to calculate log(0). Correlations were also run between each of the predictors to highlight any potential problems with multicollinearity (figure 5.1).

```
df5 <- read_csv("Exp5_data.csv")

ggpairs(data = df5, columns = c(4:5, 8, 10, 13)

df5$log_word_freq <- log(df5$word_freq + 1e-06)

df5$log_bigram_freq <- log(df5$bigram_freq + 1e-06)

df5$log_trans_prob <- log(df5$bigram_freq + 1e-06)

df5$log_response_time <- log(df5$response_time + 1e-06)
```

Means, standard deviations and inter-quartile range were also calculated for each of the variables and are shown in table 5.3.

Table 5.3: Descriptive statistics for Experiment 5

Variable	Mean	SD	Min	Max	Range	IQR
age	20.780	4.190	18.000	41.000	23.000	1.000
bigram_freq	510.640	1343.150	0.000	8465.000	8465.000	311.000
concreteness	3.460	0.960	1.680	4.970	3.290	1.730
diversity	207.610	477.000	0.000	3442.000	3442.000	153.000
word_freq	149.140	115.580	0.510	493.850	493.340	133.630
letters	5.570	1.180	3.000	8.000	5.000	1.000
response_time	659.920	208.440	208.000	1498.000	1290.000	241.000
trans_prob	0.010	0.060	0.000	0.710	0.710	0.000

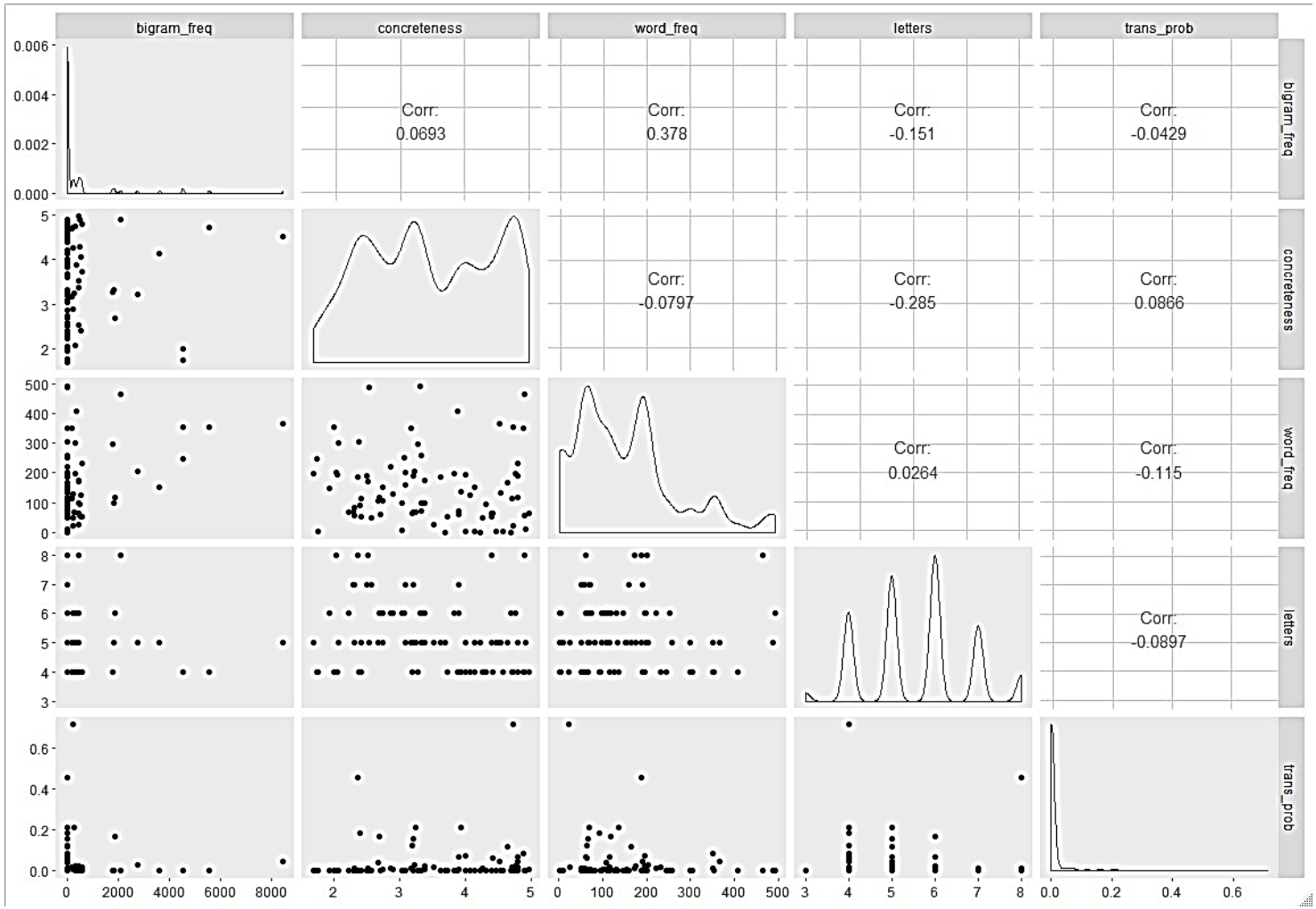

Figure 5.1: Correlation matrix for Experiment 5.

Response accuracy for all participants was greater than eighty percent and was comparable for both word and non-word trials. All non-word trials were removed prior to analysis and data was trimmed to exclude incorrect responses as well as those made faster than 200ms or more extreme than three standard deviations from the participant mean (as in Madan, Shafer, Chan, & Singhal, 2016), following this procedure 2.29% of the remaining correct trials were removed; this did not change the pattern of results.

5.4.2 **Specifying the models**

As in all previous experiments, baseline and covariate only models were used for comparative purposes. In addition, three random-intercept models were run using individual trial data (N = 7957) to predict log-transformed response times in a lexical-decision task. Individual participants and items were included as group-level effects. Bigram frequency and transitional probability were included as population level effects, both individually and together. Target-word frequency, concreteness, target-word length, and participant age were also included as covariates.

```
base_model_5 <- brm(log_response_time ~ 1, data = df5,  save_all_pars
    = TRUE, silent = TRUE, refresh = 0)
cov_model_5 <- brm(log_response_time ~ age + concreteness + letters +
    log_word_freq, data = df5, save_all_pars = TRUE, silent = TRUE,
    refresh = 0)
model_5a <- brm(log_response_time ~ bigram_freq + age + concreteness
    +  letters + word_freq + (1 | subject) + (1 | item), data = df5,
    save_all_pars = TRUE, silent = TRUE, refresh = 0)
```

```
model_5b <- brm(log_response_time ~ trans_prob + age + concreteness +
    letters + word_freq + (1 | subject) + (1 | item), data = df5,
    save_all_pars = TRUE, silent = TRUE, refresh = 0)
model_5c <- brm(log_response_time ~ bigram_freq + trans_prob + age +
    concreteness + letters + word_freq + (1 | subject) + (1 | item),
    data = df5, save_all_pars = TRUE, silent = TRUE, refresh = 0)
```

Model comparison was performed using leave-one-out cross-validation with the loo() function in R. Information criteria for all the models are displayed in Table 5.4.

```
cv_base5 <- loo(base_model_5)
cv_cov5 <- loo(cov_model_5)
cv_m5a <- loo(model_5a)
cv_m5b <- loo(model_5b)
cv_m5c <- loo(model_5c)
```

Table 5.4: Leave-one-out information criteria (LOOIC) for the models from Experiment 5

Model	Population-level	Group-level	LOOIC (SD)
Base	None	None	3523.90 (130.20)
Covariate	Age, letters, word frequency, concreteness	participant, item	1067.00 (73.80)
Model A	Age, letters, word frequency, concreteness, bigram frequency	participant, item	444.50 (71.4)
Model B	Age, letters, word frequency, concreteness, transitional probability	participant, item	363.80 (58.60)
Model C	Age, letters, word frequency, concreteness, bigram diversity, transitional probability	participant, item	364.30 (58.90)

Cross-validation shows that the transitional probability model (B) is the best model but is only marginally better than Model C – the combination model - based on the information criteria. The large standard deviation for the LOOIC also makes it

impossible to differentiate between the three experimental models with any degree of confidence.

5.4.3 **Bayes factors**

Since we were unable to adequately discriminate between the models through cross validation it was decided that Bayes factors would be used to weigh the evidence in favour of each model against each other model. These model comparisons were performed using the built-in Bayes factor calculator in brms and can be seen in Table 5.5.

```
bf_covbase5 <- bayes_factor(cov_model_5, base_model_5, silent = TRUE)
bf_5abase <- bayes_factor(model_5a, base_model_5, silent = TRUE)
bf_5bbase <- bayes_factor(model_5b, base_model_5, silent = TRUE)
bf_5cbase <- bayes_factor(model_5c, base_model_5, silent = TRUE)
bf_acov5 <- bayes_factor(model_5a, cov_model_5, silent = TRUE)
bf_bcov5 <- bayes_factor(model_5b, cov_model_5, silent = TRUE)
bf_ccov5 <- bayes_factor(model_5c, cov_model_5, silent = TRUE)
bf_5ba <- bayes_factor(model_5b, model_5a, silent = TRUE)
bf_5ca <- bayes_factor(model_5c, model_5a, silent = TRUE)
bf_5cb <- bayes_factor(model_5c, model_5b, silent = TRUE)
```

Table 5.5: Bayes factors comparing the statistical models based on data from Experiment 5

Model	Base	Covariate	A (Bigram frequency)	B (Transitional probability)
Covariate	>999			
A (bigram frequency)	>999	>999		
B (Transitional probability)	>999	>999	>999	
C (Combination)	>999	>999	>999	1.07

The Bayes factor comparisons shown in table 5.5 indicate that each of the experimental models performs better than both the base and covariate only models. It is also clear that both Model B and Model C – the transitional probability and combined models, respectively – are more likely than the bigram frequency model (Model A). However, there is insufficient evidence to differentiate between the transitional probability and combined models, just as there was with cross-validation (above). The Bayes factor of 1.07 suggests that there is slightly more evidence in favour of the combined model over the transitional probability model but, based on Raftery's (1995) guidelines this could be considered as weak, at best. As such, it must be concluded that both models are equally likely given the data and are thus set out in more detail below.

5.4.4 **Model summaries**

A summary of the model can be obtained using the summary() command.

```
summary(model_5b)
```

Table 5.6 shows that, in the transitional probability model (B) there is no meaningful effect of transitional probability on response time. This is not entirely consistent with Experiments 1 and 3 which favoured the bigram frequency model and showed a positive effect of transitional probability, respectively. Although this combination of results fails to provide conclusive evidence against transitional probability being a reasonable predictor of statistical learning, it does support the narrative that statistical learning paradigms should be giving more consideration to explanations and metrics outside of the traditional transitional probability hypothesis.

However, the accuracy of this model is questionable given that we do not see any of the expected covariate effects. In fact, opposite effects to those that would be predicted are evident for age and word length; this, in addition to the null effect of word frequency is surprising and might be considered cause for a more cautious interpretation of the data presented here.

Table 5.6: Summary of Model B, a variable intercept model based on the data from Experiment 5 which includes transitional probability as a fixed effect

Model B	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Group-level effects:						
Item	0.11	0.01	0.09	0.14	1,070.00	1.00
Participant	0.07	0.01	0.05	0.09	1,161.00	1.00
Population-level effects:						
Intercept	6.33	0.17	6.00	6.66	1,001.00	1.00
Age	[-.01, .00]	<.01	-0.01	<.00	1,929.00	1.00
Concreteness	0.01	0.02	-0.02	0.05	863.00	1.00
Letters	-0.01	0.01	-0.04	0.02	1,019.00	1.00
Word frequency	0	0.02	-0.04	0.05	867.00	1.00
Transitional probability	0	0.03	-0.06	0.07	723.00	1.00
Family specific parameters:						
Sigma	0.26	0	0.25	0.27	4,000.00	1.00

Estimate and Est.Error are equivalent to unstandardized coefficients and Std.Error respectively; Rhat is a measure of model convergence, values much greater than one indicate a failure to converge; Eff.Sample is the effective number of independent samples drawn by the MCMC after adjusting for autocorrelation.

```
summary(model_5c)
```

A summary of Model C (transitional probability) is shown in Table 5.7 This model combined the effects of transitional probability and bigram frequency and shows comparatively large effects of both. We also see the expected effects for number of letters and concreteness but unusual effects of age and target word frequency. It is also worth mentioning that the effective sample sizes in this model are consistently higher than those in the transitional probability model (B, above). Since effective sample size represents an estimate of the effective number of samples drawn from the Monte Carlo simulation after adjusting for autocorrelation – a value of 4000 indicates no correlation whereas a value of zero would indicate 100% correlation between the data points - higher values can be considered a more accurate representation of the

data. In Model B, the transitional probability model, the effective sample size is particularly low – although not necessarily problematic - for transitional probability; the same is not true of the combined model presented below (Model C). Given this discrepancy in effective sample size, I would be more inclined to favour the combined model over the transitional probability model.

Looking at the combined model, the effects of transitional probability and bigram frequency mirror those found in previous experiments. Experiment 1 highlighted a negative effect of bigram frequency which is repeated in the current model. Moreover, Experiment 3 showed a positive effect of transitional probability which can also be seen in the combination model. The effects of the two variables are more pronounced in the current data than in previous experiments, however.

Table 5.7: Model C: A variable intercept model incorporating transitional probability and bigram frequency

Model C	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Group-level effects:						
Item	0.1	0.01	0.08	0.13	1,229.00	1.00
Participant	0.07	0.01	0.05	0.09	1,586.00	1.00
Population-level effects:						
Intercept	6.11	0.19	5.72	6.49	1,844.00	1.00
Age	[-.01, .00]	<.01	-0.01	0	4,000.00	1.00
Concreteness	0.01	0.02	-0.03	0.04	1,754.00	1.00
Letters	-0.01	0.01	-0.03	0.02	1,728.00	1.00
Word frequency	0.02	0.02	-0.02	0.06	1,714.00	1.00
Bigram frequency	-0.11	0.05	-0.21	[-.01, .00]	1,407.00	1.00
Transitional probability	0.46	0.23	0.01	0.91	1,411.00	1.00
Family specific parameters:						
Sigma	0.26	<.01	0.25	0.27	4,000.00	1.00

Estimate and Est.Error are equivalent to unstandardized coefficients and Std.Error respectively; Rhat is a measure of model convergence, values much greater than one indicate a failure to converge; Eff.Sample is the effective number of independent samples drawn by the MCMC after adjusting for autocorrelation.

5.5 DISCUSSION

Experiment 5 was intended to replicate the findings of Experiments 1 and 3 using timings more typically seen in lexical decision. Model comparison using both leave-one-out cross-validation and Bayes factors showed no meaningful difference between the transitional probability and combination models. Further examination of the two most likely models highlights a discrepancy in the observed effects of transitional probability and bigram frequency. In the transitional probability only model, transitional probability is shown to be a slight negative predictor of response time; these findings should be interpreted with caution however, since the 95% credibility intervals include zero and the model shows quite low effective sample sizes for a number of predictors. Conversely, the combination model shows a strong positive association between transitional probability and response time and a strong negative effect of bigram frequency.

Given the null effect of transitional probability in Model B (The transitional probability model) and the opposing effects of bigram frequency and transitional probability in the combined model (C), it is suggested that the tenuous contribution of transitional probability in Model B may actually be masking the effect of bigram frequency - since transitional probability necessarily encapsulates the frequency of the bigram as well as the individual word frequency – and that, when separating the two, we see the true effects. At this point, this is a purely speculative position but one that is somewhat supported by the proof-of-concept experiments in previous chapters.

5.6 **EXPERIMENT 6**

The design and procedure were identical to the first experiment with the exception that bigram diversity was manipulated rather than bigram frequency. The nature of bigram diversity is such that the manipulation in this experiment focuses on the prime rather than the target word of the bigram.

5.6.1 **Participants**

Thirty-one participants (25 females) aged between 18 and 41 years (M= 20.77, SD= 4.17) were recruited from Nottingham, UK. All participants reported English as their first language and were screened for language difficulties. Participants took part in both Experiment 5 and 6 and received research credits in exchange for their participation where applicable.

5.6.2 **Materials**

Measuring bigram diversity required examining the number of words that follow a prime word ('followers') in the BNC. For example, armed is followed by forty unique words in the BNC and therefore has forty followers. The stimulus-list for experiment six comprised of forty-five bigrams and forty-five non-word pairs (non-words paired with a real-word prime) which were identical to those used in Experiment 4. These were selected from a list of unique bigrams between three and eight letters long extracted from the BNC and filtered to remove names, acronyms, initialisations, hyphenations, and numbers expressed as digits. Non-word stimuli were created using entries from the ARC Nonword Database and only non-words between with acceptable English orthographic structures were used; for example, since the letter combination

qa does not occur in English it did not appear in any of the non-words for this experiment. Each non-word was paired with a real word prime which was constrained to not appear more than once across the two experiments but was otherwise randomly selected from the BNC. Bigrams were selected to include an equal number of high, low, and zero diversity items, examples of which can be seen in Table 5.9. Bigram frequency was not controlled across stimuli since attempting to do so resulted in a prohibitively small stimulus-pool, as such bigram frequency was free to vary across items. Group descriptive statistics can be seen in table 5. 8.

Table 5.8: Group descriptive statistics for high, low, and zero diversity items in Experiment 6

Level	Bigram_frequency	Bigram_diversity	Target_frequency	Letters	Concreteness	Transitional_probability
High Diversity	376.848925	483.5146123	152.09 (120.93)	5.46 (1.09)	2.93 (.95)	0.002 (.61)
Low Diversity	20.71221971	1.989522297	150.52 (117.47)	4.92 (.17)	3.98 (.30)	0.38 (.86)
Zero Diversity	.00 (.00)	.00 (.00)	150.72 (116.19)	5.05 (.25)	3.19 (.34)	.00 (.00)
Log-transformed values						
High Diversity	5.95 (1.18)	6.20 (1.51)	5.02 (4.78)	-	-	-6.24 (0.49)
Low Diversity	3.04 (.70)	0.69 (.58)	5.01 (4.77)	-	-	-0.98 (.86)
Zero Diversity	-13.82 (.00)	-13.82 (.00)	5.02 (4.76)	-	-	-13.82 (.00)

A constant of .000001 was added to the zero values before transformation

Table 5.9: Example high, low, and zero diversity items used in Experiment 6

Prime	Target	Group	Bigram_Diversity	Bigram_Frequency	Transitional_Probability	Letters	Concreteness	Word_Frequency
that	place	High Diversity	2,074.00	367.00	0.00	5.00	3.48	48,651.00
were	almost	High Diversity	1,170.00	600.00	0.00	6.00	1.66	31,588.00
will	gain	High Diversity	870.00	151.00	0.00	4.00	2.24	5,218.00
could	happen	High Diversity	617.00	350.00	0.00	6.00	1.78	8,760.00
very	deep	High Diversity	542.00	146.00	0.00	4.00	3.38	10,700.00
even	among	High Diversity	372.00	186.00	0.00	5.00	2.38	22,864.00
common	object	High Diversity	155.00	47.00	0.00	6.00	3.66	6,325.00
total	lack	High Diversity	152.00	80.00	0.00	4.00	2.04	10,068.00
clearly	defined	High Diversity	128.00	363.00	0.02	7.00	2.07	5,898.00
living	alone	High Diversity	112.00	239.00	0.01	5.00	2.86	13,265.00
strict	sense	Low Diversity	18.00	64.00	0.03	5.00	2.61	21,935.00
musty	smell	Low Diversity	2.00	16.00	0.11	5.00	3.70	3,755.00
croquet	club	Low Diversity	1.00	11.00	0.08	4.00	3.78	16,465.00
dimmer	switch	Low Diversity	1.00	11.00	0.15	6.00	4.07	3,316.00
hallowed	ground	Low Diversity	1.00	11.00	0.08	6.00	4.77	16,200.00
loony	left	Low Diversity	1.00	34.00	0.23	4.00	3.70	47,089.00
revolve	around	Low Diversity	1.00	72.00	0.57	6.00	1.96	45,286.00
shady	corner	Low Diversity	1.00	12.00	0.04	6.00	4.61	7,500.00
snare	drum	Low Diversity	1.00	10.00	0.10	4.00	4.96	985.00
whacking	great	Low Diversity	1.00	12.00	0.28	5.00	1.81	45,217.00
orate	red	Zero Diversity	0.00	0.00	0.01	3.00	4.24	15,136.00
warren	kept	Zero Diversity	0.00	0.00	0.00	4.00	2.79	14,306.00
opulent	recent	Zero Diversity	0.00	0.00	0.01	6.00	2.50	15,858.00
conflate	art	Zero Diversity	0.00	0.00	0.01	3.00	4.17	15,587.00
comely	answer	Zero Diversity	0.00	0.00	0.04	6.00	2.89	14,421.00
taxes	lead	Zero Diversity	0.00	0.00	0.03	4.00	4.10	14,555.00
sable	field	Zero Diversity	0.00	0.00	0.02	5.00	4.26	15,298.00
hoary	despite	Zero Diversity	0.00	0.00	0.04	7.00	1.33	14,592.00
detritus	works	Zero Diversity	0.00	0.00	0.91	5.00	3.79	14,528.00
ley	poor	Zero Diversity	0.00	0.00	0.16	4.00	2.70	15,125.00

5.6.3 **Procedure**

Participants were presented with a real-word prime drawn from the initial position of a bigram (with half of the bigrams being word pairs with the bigram being zero, low, or high frequency; and half being a word-nonword pair). The ‘prime’ (first word of the bigram) remained on the screen for 75ms before being immediately replaced with the ‘target’ (second word of the bigram); the target was presented for a maximum of 1500ms during which time participants were required to press either ‘z’ or ‘m’ on a standard QWERTY keyboard; key mapping was systematically varied so that half of all

participants used ‘z’ to indicate a word and ‘m’ to indicate a non-word whilst half responded with ‘m’ for words and ‘z’ for non-words. A fixation point was presented in the centre of the screen for 500ms prior to each trial. Prime-Target pairs were presented in two blocks each containing forty-five bigram trials and forty-five non-word trials. The order in which the blocks were presented was counterbalanced and individual trials were randomised for each participant.

5.7 **RESULTS**

Accuracy was comparable for both word and non-word trials. Data from experiment two was trimmed and analysed using the same procedure as the first experiment, a total of 2.04% of correct trials were removed (this did not change the pattern of results). All response time data were log-transformed; mean RTs for each participant were then analysed using a Bayesian multi-level regression. Individual trial data ($N = 2170$) was used to predict log-transformed response times in a lexical-decision task using three random-intercept models. Individual participants and items were included as group-level effects. Bigram diversity and transitional probability were included as population-level effects, both singly and individually. Target-word frequency, concreteness, target-word length, and participant age were also included as covariates. Leave-one-out cross-validation statistics were used to compare model fit, with smaller values considered indicators of goodness-of-fit. Log-transformed values were used for bigram diversity, word frequency, transitional probability and response time; a constant of .000001 was added to all values to avoid errors resulting from items with values equal to zero.

5.7.1 **Data preparation**

Data was read into R and analysed in the same manner as Experiment 1, correlation between predictors was examined and the results displayed in figure 5.2. The Bigram frequency, transitional probability, and response time variables were log transformed prior to the analysis; a small constant was added to all the values to avoid errors resulting from trying to calculate log(0).

```
df6 <- read_csv("Exp6_data.csv") ggpairs(data = df6, columns =
     c(5, 7:8, 10, 13)) +  theme(panel.grid = element_blank())
df6$word_freq <- log(df6$word_freq + 1e-06)
df6$bigram_freq <- log(df6$bigram_freq + 1e-06)
df6$trans_prob <- log(df6$bigram_freq + 1e-06)
df6$response_time <- log(df6$response_time + 1e-06)
```

Table 5.10: Descriptive statistics for Experiment 6

Variable	Mean	SD	Min	Max	Range	IQR
age	20.780	4.190	18.000	41.000	23.000	1.000
bigram_freq	297.320	978.720	0.000	8465.000	8465.000	151.000
concreteness	3.400	1.030	1.330	5.000	3.670	1.760
diversity	286.890	719.620	0.000	5217.000	5217.000	214.000
word_freq	150.130	112.630	0.510	493.850	493.340	123.480
letters	5.470	1.170	3.000	8.000	5.000	1.000
response_time	644.260	198.570	208.000	1498.000	1290.000	228.750
trans_prob	0.030	0.090	0.000	0.710	0.710	0.000

The descriptive statistics for each of the variables used in the analyses in Experiment 6 are displayed in table 5.10. Shown in the table are the means, standard deviations, minimum and maximum values of each variable along with the range and interquartile range.

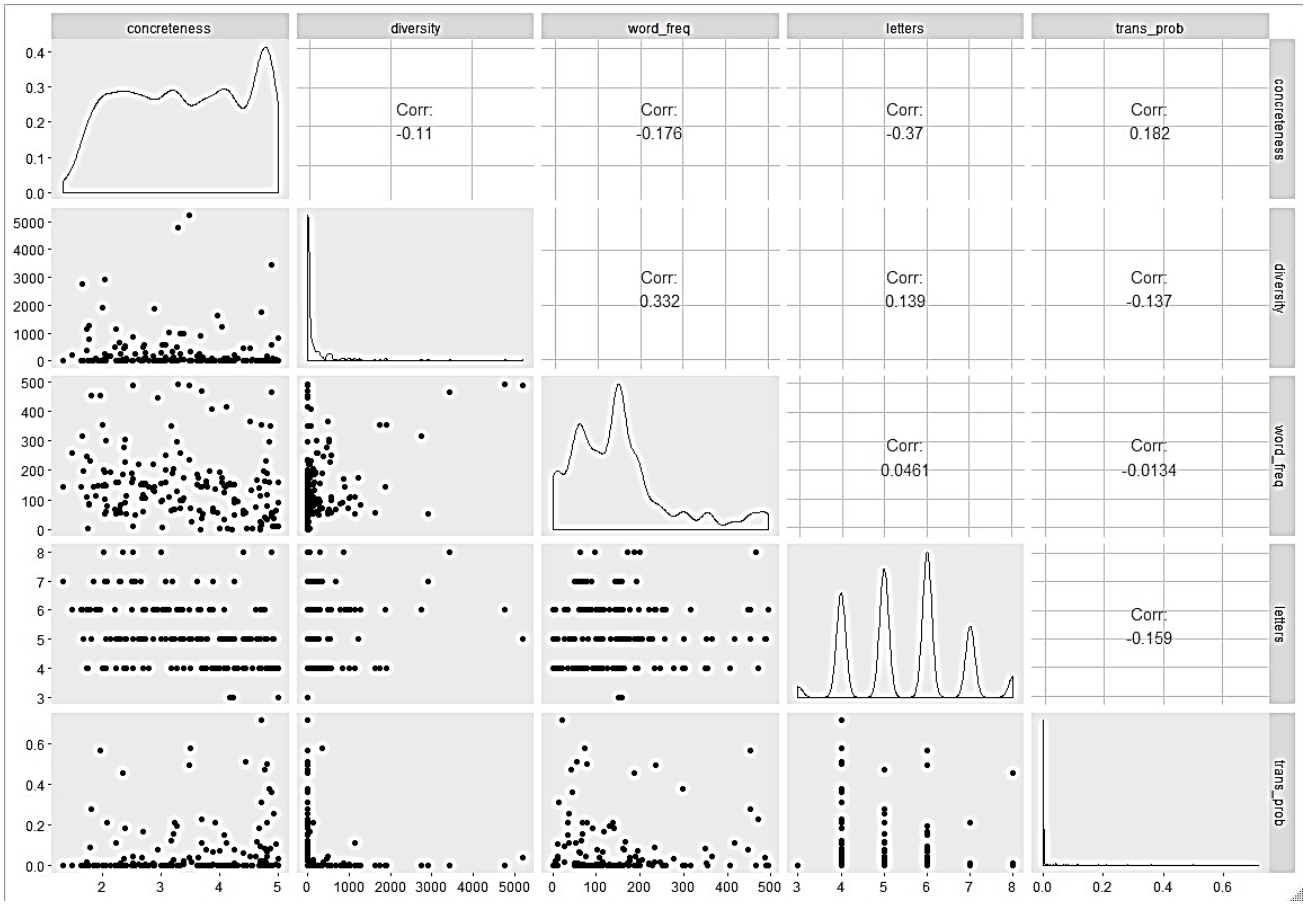

Figure 5.2: Correlation matrix for Experiment 6

5.7.2 **Specifying the models**

Models were run in the same way as previous experiments and consist of a baseline model, covariate model, and three experimental models. Models A, B, and C examined bigram diversity, transitional probability, and both variables respectively; all models included participant age, target word frequency, concreteness, and number of letters as population-level effects and participant and item as group-level effects.

```
base_model_6 <- brm(response_time ~ 1, data = df6, save_all_pars =
    TRUE, silent = TRUE, refresh = 0)
cov_model_6 <- brm(response_time ~ age + concreteness + letters
    +  word_freq, data = df6, save_all_pars = TRUE, silent =
    TRUE,  refresh = 0)
 model_6a <- brm(response_time ~ diversity + age + concreteness
    + letters + word_freq + (1|subject) + (1|item), data = df6,
    save_all_pars = TRUE, silent = TRUE, refresh = 0)
 model_6b <- brm(response_time ~ trans_prob + age +
    concreteness + letters + word_freq + (1|subject) +
    (1|item), data = df6,  save_all_pars = TRUE, refresh = 0)
model_6c <- brm(response_time ~ diversity + trans_prob + age +
    concreteness + letters + word_freq + (1|subject) +
    (1|item),  data = df6, save_all_pars = TRUE, silent = TRUE,
    refresh = 0)
```

5.7.3 **Cross-validation**

Model comparison was performed using leave-one-out cross-validation with the loo() function in R. Information criteria for all the models are displayed in Table 5.11.

```
cv_base6 <- loo(base_model_6)

cv_cov6 <- loo(cov_model_6)

cv_m6a <- loo(model_6a)

cv_m6b <- loo(model_6b)

cv_m6c <- loo(model_6c)
```

Table 5.11 shows that the bigram diversity model is slightly better at predicting the data than the remaining models. However, we once again encounter the problem of high standard deviation in the leave-one-out information criteria which makes it impossible to meaningfully discriminate between the models. Although this has been a recurring theme throughout this thesis the decision was made to continue using LOOIC as the initial metric of model comparison since it is the most comprehensive measure of model fit available and despite being unable to discriminate between the three experimental models has proven effective at demonstrating the improvement of these models over the baseline and covariate only models. Additionally, the large standard deviation around the information criteria forces a more conservative interpretation of the model comparisons and allows the selection of one model over another only if there is a clear and substantial improvement in LOOIC. As such, I will continue to run and report cross-validation statistics in the remaining chapters of this work.

Table 5.11: Cross-validation information criteria for statistical models based on data from Experiment 6

Model	Population-level	Group-level	LOOIC (SD)
Base	None	None	41225.1 (153.7)
Covariate	Age, letters, word frequency, concreteness	participant, item	1446.7 (107.7)
Model A	Age, letters, word frequency, concreteness, bigram diversity	participant, item	543.30 (100.90)
Model B	Age, letters, word frequency, concreteness, transitional probability	participant, item	547.80 (82.70)
Model C	Age, letters, word frequency, concreteness, bigram diversity, transitional probability	participant, item	544.40 (82.50)

5.7.4 **Bayes factors**

Due to the lack of discrimination between the models based on cross-validation, Bayes factors were used for model comparison and can be seen in Table 5.12.

```
bf_covbase6 <- bayes_factor(cov_model_6, base_model_6, silent = TRUE)
bf_6abase <- bayes_factor(model_6a, base_model_6, silent = TRUE)
bf_6bbase <- bayes_factor(model_6b, base_model_6, silent = TRUE)
bf_6cbase <- bayes_factor(model_6c, base_model_6, silent = TRUE)
bf_acov6 <- bayes_factor(model_6a, cov_model_6, silent = TRUE)
bf_bcov6 <- bayes_factor(model_6b, cov_model_6, silent = TRUE)
bf_ccov6 <- bayes_factor(model_6c, cov_model_6, silent = TRUE)
bf_6ba <- bayes_factor(model_6b, model_6a, silent = TRUE)
bf_6ca <- bayes_factor(model_6c, model_6a, silent = TRUE)
bf_6cb <- bayes_factor(model_6c, model_6b, silent = TRUE)
```

Table 5.12: Bayes factor comparisons for statistical models based on data from Experiment 6

Model	Base	Covariate	A (Bigram diversity)	B (Transitional probability)
Covariate	>999			
A (Bigram diversity)	>999	>999		
B (Transitional probability)	>999	>999	>999	
C (Combination)	>999	>999	>999	<.001

Based on the Bayes factors set out in table 5.12, we can see that each of the experimental models is more likely, given the evidence, than both the baseline and covariate models. We also see that the transitional probability model (B) is better than both the bigram diversity (A) and combined (C) models by a margin of greater than 999 (since the comparison shows the strength of evidence for C over B as less than .001, we can obtain the inverse Bayes factor by doing 1/.001). This means that, as in Experiment 4, there is a greater strength of evidence for the transitional probability model than any of the other models presented here, based on the observed data.

5.7.5 **Model summary**

```
summary(model_6b)
```

A summary of the transitional probability model is set out in table 5.13, this model was judged as most likely based on Bayes factor analysis (above).

Table 5.13: Summary of Model B, the transitional probability model

-	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Group-level effects						
Item	0.1	0.01	0.08	0.12	1,148.00	1.01
Participant	0.05	0.01	0.03	0.07	1,516.00	1.00
Population-level effects:						
Intercept	6.54	0.12	6.31	6.77	1,052.00	1.00
Age	[-.01, .00]	< .01	-0.01	0.00	2,138.00	1.00
Concreteness	0	0.01	-0.02	0.03	867.00	1.00
Letters	-0.01	0.01	-0.03	0.01	1,057.00	1.00
Word frequency	-0.02	0.01	-0.04	0.01	1,172.00	1.00
Transitional probability	-0.02	0.02	-0.07	0.03	1,051.00	1.01
Family specific parameters:						
Sigma	0.26	<.01	-0.04	0.01	1,172.00	1.00

Estimate and Est.Error are equivalent to unstandardized coefficients and Std.Error respectively; Rhat is a measure of model convergence, values much greater than one indicate a failure to converge; Eff.Sample is the effective number of independent samples drawn by the MCMC after adjusting for autocorrelation.

Table 5.11 shows a minor negative association between transitional probability and response time. However, the wide credibility interval (which also includes zero) indicates that this result should be interpreted with caution.

5.8 **DISCUSSION**

Experiment 6 examined the effects of bigram diversity and transitional probability on response time in a lexical decision task. Model comparison suggests that the transitional probability model is the most likely model given the observed data. Further examination of the model shows that transitional probability has a weak negative relationship with response time, albeit not one that could be described as 'significant' as it is commonly understood. This is congruent with Experiment 4, in the previous chapter, which suggested a non-meaningful effect of transitional probability but of a greater magnitude than the effect seen in the current data.

5.9 GENERAL DISCUSSION

In Experiments 5 and 6, I attempted to replicate the findings highlighted by the Experiments (1-4) shown in previous chapters using experimental timings more typical of the primed lexical decision paradigm. The original experiments were designed in such a way as to afford participants what I believed to be the greatest possible opportunity to encode, and therefore benefit from, the statistical primes without significant deviation from the experimental architecture usually seen in such tasks. This resulted in longer prime exposure times than are typical in published research as well as delay between prime and target which, on reflection, could have allowed lexical activation to decay prior to the target onset. Since these were originally intended as proof of concept for statistical priming it was unclear as to whether participants would pick up on what are relatively weak statistical associations drawn from a large naturalistic language corpus. With these prototype timings, the original experiments were somewhat successful in demonstrating sensitivity to the priming effect provided by the distributional statistics inherent to the British National Corpus; as such, the studies were repeated using a shorter exposure time for the prime and no delayed onset for the target.

The experiments presented in this chapter suggest that the distributional statistics of a language still influence task performance in a primed lexical decision task when the prime is presented for a much shorter period. Furthermore, the most likely models for Experiments 5 and 6 (above) are congruent with the interpretations of the previous four experiments. In Experiment 5, which contrasted the effects of bigram frequency and transitional probability on word recognition speed, I showed that a combined model including both metrics was the best model at describing the data.

This highlighted a negative association between bigram frequency and response time as well as a positive relationship between transitional probability and response time. This is an interesting development since it suggests that there is more to be gained by increased exposure to a bigram than by stronger predictivity, at least in the current task.

Conventional thinking in statistical learning theory is that learners are using transitional probability to extract patterns from any given input and use these to inform beliefs about the nature of the stimulus. This is most prominent in studies of word segmentation (e.g., Saffran, Aslin, & Newport, 1996) where infants are presumed to use differences in conditional probability as a way of discriminating between within- and across- item transitions (hence the term transitional probability) in order to accurately parse words from speech streams in the absence of alternative cues (e.g., syllable stress or utterance boundaries). However, earlier in this work I argued that transitional probability constitutes a complex mental calculation that is unlikely to scale to naturalistic language-sets. Furthermore, transitional probabilities demonstrated in artificial grammars provide unrealistic predictive cues which do not adequately represent the way in which humans interact with languages. It was, and still is, my assertion that bigram frequency represents a better tool for understanding language patterns due to its less complex nature and is a more intuitive representation of how language is used.

To clarify, bigram frequency represents the number of times a bigram is encountered in a given subset of language. In this way, it represents a snapshot of bigram usage at any given time. This is arguably more useful than transitional probability which represents the likelihood of one item appearing after another – at least for the current

paradigm in which we are interested in previously learnt associations, this may be different for newly acquired information where predictability could be more beneficial. This is more in line with the way in which humans understand and describe their environment (e.g., McDowell et al., 2018; Tversky and Kahneman, 1973). Moreover, the intuitive way in which we interact with language can be at odds with the actual distributional properties as represented by transitional probability.

Consider the bigrams Premier League and insights into, both have a transitional probability within the British National Corpus of .35 which suggests that when presented with either Premier or insights learners should be able to predict the second part of the bigram with equal accuracy. However, it would be difficult to argue that insights into is as recognisable a bigram as Premier League and we might intuit that one is more likely to occur in a given subset of language than the other. In this example, Premier League occurs over three times more frequently (879 compared to 270 occurrences) than insights into demonstrating that although transitional probabilities incorporate a frequency component, in some cases they serve to obfuscate this information.

If we accept the combined model, then Experiment 5 supports a bigram frequency hypothesis since the data suggests that higher frequency bigrams lead to faster recognition of words, but it also suggests that transitional probability might be interfering with the recognition of words. But what about the transitional probability only model (Model B)? This model was shown to be indistinguishable from the combined model on both LOOIC and Bayes factor comparison and shows transitional probability as having a non-meaningful, negative relationship with response time. As I mentioned above, there is some question as to the reliability of this model compared

to Model C (Combined model) which leads me to favour the latter model overall; but if we consider them together, we see something interesting. In the absence of bigram frequency as a predictor, transitional probability appears to account for a small increase in task performance. However, when bigram frequency is introduced to the model, we see reversal of the transitional probability effect and a new effect of bigram frequency improving response time. This may suggest that transitional probability is masking a frequency effect which is better accounted for by bigram frequency.

Once again, we see that transitional probability does not perform as expected given the strength of published evidence behind it. Although the studies presented here are far from conclusive, they should prompt us to ask whether transitional probability is, realistically, the best metric of statistical distribution given that there appears to be little benefit beyond that provided by a raw frequency metric.

Experiment 6 was equally successful at replicating the effects shown in the proof of concept chapters in that there was, once again, no meaningful effect of either transitional probability or bigram diversity. This is somewhat disheartening given the documented effects of predictability and contextual diversity in language tasks but, considering the results from the bigram frequency experiments is not particularly surprising. In those experiments, as discussed above, we see no real benefit of the predictability component of transitional probability and so to see the same null effects in Experiment 6 is also somewhat encouraging since it goes some way towards supporting that hypothesis. That said, it is becoming clear that bigram diversity is unlikely to constitute a meaningful metric in describing the statistical regularities of a stimulus-set. However, in the interest of completeness – and for the sake of the planned meta-analysis – I shall still be conducting the planned final experiment

comparing transitional probability and bigram diversity. These experiments, presented in the next chapter, attempt to build on the findings from Experiments 5 and 6 with the target stimuli held constant across each level of bigram frequency and diversity, respectively.

CHAPTER SUMMARY

In this chapter I:

- Repeated Experiments 1 & 2 using timings more typically seen in primed lexical decision paradigms
- Showed that bigram frequency may be a better metric of statistical distribution than transitional probability in predicting word recognition performance
- Suggested that transitional probability may be masking an effect of frequency
- Questioned the value of the predictive component of transitional probability
- Concluded that bigram diversity is unlikely to constitute a meaningful descriptor of statistical regularity

6 ADJUSTED TIMINGS 2

CHAPTER OVERVIEW

In Chapter 6, I:

- Repeat Experiments 5 & 6 whilst holding target words constant at each level of bigram frequency and bigram diversity
- Will expand on the theoretical interpretations set out in the previous chapter based on the data from Experiments 7 & 8

6.1 PREPARATION

The following code excerpt initialises the packages necessary to run the analyses in this chapter and introduces some global settings in the interest of reproducibility.

```
library(formatR)
library(readr)
library(brms)
library(rstanarm)
library(GGally)
set.seed(100)
```

6.2 EXPERIMENTS

In Chapter 4, I discussed how a large amount of variation in response times could be attributed to differences in individual items. To recap, although efforts were made to balance the stimuli across the three levels of bigram frequency and diversity in their respective experiments, individual items may differ both qualitatively and experientially for each participant. Two separate items cannot therefore be treated as equivalent even when perfectly balanced on every dimension and treating them as such is known as the language-as-fixed-effects fallacy (Clark, 1973).

In order to overcome this issue and effectively reduce variation between items the stimuli for Experiment 3 were adjusted in such a way that the target word was held constant at the high, low, and zero frequency levels. This was deemed successful at reducing the inter-item variability but has the potential to introduce practise effects since participants were exposed to each target on multiple trials. Conducting a replication of Experiment 3 will allow me to contrast the results of the two paradigms and identify whether there is any improvement in task performance when the target word is held constant across the different levels of bigram frequency.

Experiment 7 is therefore a replication of Experiment 5 but with the targets held constant across the different levels of bigram frequency. Similarly, Experiment 8 seeks to replicate the findings of Experiment 6 by once again holding the targets constant across levels. Thus, these experiments replicate those presented in Chapter 4 using the newly modified timings. Interpretations as to the role of bigram frequency, bigram diversity, and transitional probability will then be based upon the findings of each set of experiments.

6.3 EXPERIMENT 7

A primed lexical decision task in which the target words were held constant across different levels of bigram frequency was used to assess the relative impacts of bigram frequency and transitional probability response times. Data from the experiment was used to inform a number of variable-intercept models, the best of which was selected using leave-one-out cross-validation and Bayes factor comparisons.

6.3.1 Participants

Fifty participants (44 females) aged between 18 and 53 years (M= 22.29, SD= 9.44) were recruited from Nottingham, UK. All participants reported English as their first language and were screened for language difficulties. Participants received research credits in exchange for their participation where applicable.

6.3.2 Materials

The experimental stimuli consisted of one-hundred and eighty bigrams and one hundred and eighty non-words. These were drawn from the same pool of 12,293,349 unique bigrams extracted from the British National Corpus and used in each of the previous experiments. Non-word stimuli were drawn from the ARC nonword Database (Rastle et al, 2002) and constrained to be between three and eight letters long and contain only legal orthographic structures in English. Each non-word was paired with a unique real word prime to form a non-word bigram. Measures of frequency, concreteness, and number of letters were also obtained for use as covariates. Since the targets were identical across levels, the main constraint was identifying targets that occurred as part of both low and high frequency bigrams –

zero frequency bigrams, by definition, did not occur in the British National Corpus and were created by pairing a real word with the targets from the other levels and using a lookup function to ensure the bigram was not present in the stimulus pool. This resulted in a stimulus-set comprising sixty sets of three targets, each with a high, low, and zero frequency bigram, examples of which can be seen in Table 6.2, with group descriptive statistics shown in table 6.1.

Table 6.1: Group descriptive statistics for high, low, and zero bigram frequency

Level	Bigram_frequency	Bigram_diversity	Target_frequency	Letters	Concreteness	Transitional_probability
High Frequency	507.52 (1337.30)	206.09 (474.19)	143.57 (110.55)	5.57 (1.18)	3.46 (.96)	.01 (.06)
Low Frequency	11.00 (1.71)	208.62 (487.11)	143.91 (112.95)	5.56 (1.18)	3.48 (.96)	.01 (.06)
Zero frequency	.00 (.00)	203.80 (472.04)	147.46 (115.90)	5.57 (1.18)	3.46 (.96)	.00 (.00)
Log-transformed values						
High frequency	6.23 (7.20)	5.33 (6.16)	4.97 (4.70)	-	-	-4.61 (2.81)
Low frequency	2.40 (.54)	5.34 (6.19)	4.97 (4.73)	-	-	-4.61 (2.81)
Zero frequency	-13.82 (.00)	5.32 (6.16)	4.99 (4.75)	-	-	-13.82 (.00)

A constant of .000001 was added to the zero values before transformation

Table 6.2: Example stimuli for Experiment 7

Prime	Target	Group	Bigram_Diversity	Bigram_Frequency	Transitional_Probability	Letters	Concreteness	Word_Frequency
talking	about	High Frequency	69.000	5376.000	0.002	5.000	1.770	1971.160
those	aged	High Frequency	909.000	309.000	0.001	4.000	3.140	47.030
most	basic	High Frequency	1090.000	268.000	0.001	5.000	2.260	112.270
she	bent	High Frequency	1430.000	142.000	0.310	4.000	3.620	23.930
some	bread	High Frequency	1928.000	112.000	0.000	5.000	4.920	37.700
eldest	daughter	High Frequency	14.000	111.000	0.014	8.000	4.790	94.430
clearly	defined	High Frequency	215.000	363.000	0.001	7.000	2.070	58.980
soon	enough	High Frequency	144.000	164.000	0.005	6.000	1.330	325.940
blonde	hair	High Frequency	9.000	122.000	0.024	4.000	4.970	144.530
scattered	over	High Frequency	23.000	92.000	0.034	4.000	2.460	1322.760
second	about	Low Frequency	498.000	10.000	0.001	5.000	1.770	1971.160
someone	aged	Low Frequency	160.000	10.000	0.001	4.000	3.140	47.030
full	basic	Low Frequency	355.000	10.000	0.001	5.000	2.260	112.270
being	bent	Low Frequency	1223.000	10.000	0.000	4.000	3.620	23.930
eating	bread	Low Frequency	72.000	10.000	0.005	5.000	4.920	37.700
wife	daughter	Low Frequency	173.000	10.000	0.001	8.000	4.790	94.430
thus	defined	Low Frequency	216.000	10.000	0.002	7.000	2.070	58.980
humble	enough	Low Frequency	7.000	10.000	0.020	6.000	1.330	325.940
silky	hair	Low Frequency	3.000	11.000	0.040	4.000	4.970	144.530
basin	over	Low Frequency	15.000	10.000	0.021	4.000	2.460	1322.760
mud	about	Zero Frequency	24.000	0.000	0.000	5.000	1.770	1971.160
morning	aged	Zero Frequency	143.000	0.000	0.000	4.000	3.140	47.030
pop	basic	Zero Frequency	51.000	0.000	0.000	5.000	2.260	112.270
fan	bent	Zero Frequency	26.000	0.000	0.000	4.000	3.620	23.930
crouch	bread	Zero Frequency	4.000	0.000	0.000	5.000	4.920	37.700
corn	daughter	Zero Frequency	18.000	0.000	0.000	8.000	4.790	94.430
census	defined	Zero Frequency	18.000	0.000	0.000	7.000	2.070	58.980
anxiety	enough	Zero Frequency	32.000	0.000	0.000	6.000	1.330	325.940
poem	hair	Zero Frequency	31.000	0.000	0.000	4.000	4.970	144.530
praise	over	Zero Frequency	21.000	0.000	0.000	4.000	2.460	1322.760

6.3.3 **Procedure**

Participants were presented with series of trials in which a real-word prime drawn from the initial position of a bigram appeared on the screen for 75ms before being immediately replaced with the target, which consisted of the second word of the same bigram. The target was presented for a maximum of 1500ms during which time participants were required to press either ‘z’ or ‘m’ on a standard QWERTY keyboard; key mapping was systematically varied based on participant number so that odd

numbered participants used ‘z’ to indicate a word and ‘m’ to indicate a non-word whilst even-numbered participants responded with ‘m’ for words and ‘z’ for nonwords. A central fixation point was presented for 500ms prior to each trial. These timings replace the longer, less typical 150ms prime-exposure time and remove the delay between the display of prime and target words. As noted previously, these new timings are representative of those more widely seen in lexical decision experiments (e.g., Ferre et al., 2015; Kusonose et al., 2016). Prime-Target pairs were presented in four blocks each containing forty-five bigram trials. Each block contained fifteen high, low and zero frequency items and forty-five non-word trials, for a total of ninety items per block. The blocks were presented in a counterbalanced order and individual trials were randomised for each participant.

6.4 **RESULTS**

6.4.1 **Data preparation**

Data was trimmed to exclude incorrect responses as well as those made faster than 200ms or more extreme than three standard deviations from the participant mean (as in Madan, Shafer, Chan, & Singhal, 2016), a total of 1.96% of correct trials were removed (this did not change the pattern of results). Once again, participants showed high levels of accuracy for both word and non-word trials (>80%). All response time data were log-transformed; response times for each participant were then analysed using a Bayesian multi-level regression.

```
df7 <- read_csv("Exp7_data.csv") ggpairs(data = df7, columns =
     c(2:3, 5, 7, 13)) +  theme(panel.grid = element_blank())
df7$log_word_freq <- log(df7$word_freq + 1e-06)
df7$log_bigram_freq <- log(df7$bigram_freq + 1e-06)
df7$log_trans_prob <- log(df7$bigram_freq + 1e-06)
df7$log_response_time <- log(df7$response_time + 1e-06)
```

Correlations between predictors were examined and no evidence of multicollinearity was found. Figure 6.1 shows the distributions for each predictor as well as the correlation coefficients. Also calculated were descriptive statistics for each variable, these are shown in table 6.3.

Table 6.3: Descriptive statistics for Experiment 7

Variable	Mean	SD	Min	Max	Range	IQR
age	23.040	6.070	18.000	53.000	35.000	4.000
bigram_freq	398.630	2603.960	0.000	40008.000	40008.000	154.000
concreteness	3.110	1.100	1.220	4.970	3.750	1.840
word_freq	462.220	1593.280	0.000	11153.820	11153.820	132.650
letters	4.980	1.440	3.000	8.000	5.000	2.000
response_time	445.530	77.430	217.000	785.000	568.000	100.000
trans_prob	0.010	0.060	0.000	0.810	0.810	0.000

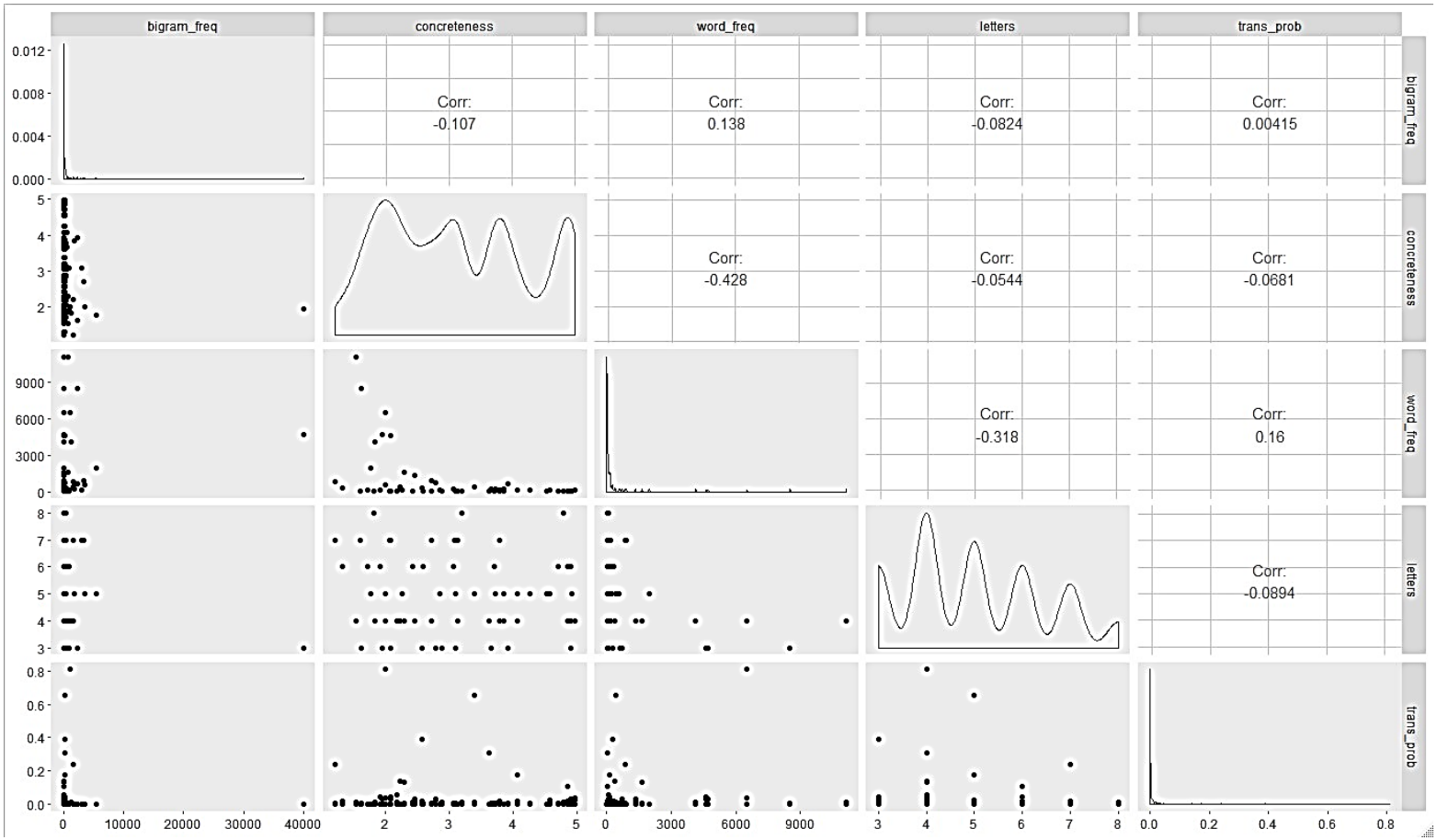

Figure 6.1: Correlation matrix for Experiment 7

6.4.2 **Specifying the models**

Log-transformed response times were modelled using Individual trial data (N = 11,083) in five random-intercept models. Individual participants and items were included as group-level effects. Bigram frequency and transitional probability were included as population-level effects, both singly and individually. Target-word frequency, concreteness, target-word length, and participant age were also included as covariates. Baseline and covariate models were used for comparative purposes alongside three experimental models: Model A, the bigram frequency model; Model B, the transitional probability model; and Model C, the combined model.

In Chapter 4 I highlighted some of the problems associated with using Bayes factors for model comparison; particularly, that they can vary wildly based on the chosen priors. This becomes problematic if there is insufficient new data from which to draw conclusions since the prior distributions will have a more pronounced effect on the posterior distribution in small datasets than large ones. Although I don't believe this to be an issue in the current datasets due to the large number of observations, the accuracy of the analyses can still be improved by the application of conjugate priors. That is, priors that more accurately represent the expected distribution of the data. Unfortunately, I was unable to specify more accurate priors in earlier experiments owing to the novel use of the lexical decision paradigm to investigate statistical priming effects – statistical learning research is focused primarily on the acquisition of new information and the manipulation of distributional statistics to facilitate learning. This meant that there was insufficient data available to predict the likely effect sizes, particularly for bigram frequency and bigram diversity which have not previously been studied. As such, the decision was made to use the non-informative default

priors built-in to the brms package for all predictors. However, the experiments presented in previous chapters allow for a more accurate specification of the prior distribution based on the effect sizes observed in those analyses.

The following conjugate priors were therefore placed on each of the predictors:

Bigram frequency = N(-.01, .01), transitional probability = N(0, .01), age = N(.01, .01), word frequency = N(0, .01), concreteness = N(0, .01), and number of letters in the target word = N(.01, .01). Prior distributions were selected based on the mean of the observed posterior distributions of all models in the previous experiments with slightly wider standard deviations. No priors were placed on the baseline model since it does not include any predictor variables.

6.4.3 Define priors

```
priors_cov7 <- c(prior("normal(0, .01)", class = "b",  coef =
    log_word_freq),  prior("normal(0, .01)", class = "b", coef
    = concreteness), prior("normal(.01, .01)", class = "b",
    coef = letters), prior("normal(.01, .01)", class = "b",
    coef = age))
priors_model_a7 <- c(prior("normal(0, .01)", class = "b",  coef =
    log_word_freq),  prior("normal(0, .01)", class = "b", coef =
    concreteness), prior("normal(.01, .01)", class = "b", coef =
    letters),  prior("normal(.01, .01)", class = "b", coef = age),
    prior("normal(-.01, .01)", class = "b", coef = log_bigram_freq))
```

```
priors_model_b7 <- c(prior("normal(0, .01)", class = "b",  coef
    = log_word_freq),  prior("normal(0, .01)", class = "b", coef
    = concreteness), prior("normal(.01, .01)", class = "b", coef
    = letters),  prior("normal(.01, .01)", class = "b", coef =
    age), prior("normal(0, .01)", class = "b", coef =
    log_trans_prob))
priors_model_c7 <- c(prior("normal(0, .01)", class = "b",  coef
    = log_word_freq), prior("normal(0, .01)", class = "b", coef
    = concreteness), prior("normal(.01, .01)", class = "b", coef
    = letters),  prior("normal(.01, .01)", class = "b", coef =
    age), prior("normal(-.01,.01)", class = "b", coef =
    log_bigram_freq), prior("normal(0, .01)", class = "b", coef
    = log_trans_prob))
```

6.4.4 Run Models

```
base_model_7 <- brm(log_response_time ~ 1, data = df7,
    save_all_pars = TRUE, silent = TRUE, refresh = 0)
cov_model_7 <- brm(log_response_time ~ age + concreteness +
    letters + log_word_freq, data = df7, save_all_pars = TRUE,
    prior = priors_cov7, silent = TRUE, refresh = 0)
model_7a <- brm(log_response_time ~ log_bigram_freq + age +
    concreteness + letters + log_word_freq + (1 | subject) + (1
    | item), data = df7,  save_all_pars = TRUE, prior =
    priors_model_a7, silent = TRUE, refresh = 0)
```

```
model_7b <- brm(log_response_time ~ log_trans_prob + age +
    concreteness + letters + log_word_freq + (1 | subject) +
    (1 | item), data = df7,  save_all_pars = TRUE, prior =
    priors_model_b7, silent = TRUE, refresh = 0)
model_7c <- brm(response_time ~ bigram_freq + trans_prob + age +
    concreteness + letters + word_freq + (1 | subject) + (1 |
    item),  data = df7, prior = priors_model_c7, save_all_pars =
    TRUE,  silent = TRUE, refresh = 0)
```

6.4.4.1 Cross-validation

Model comparison was performed using leave-one-out cross-validation with the `loo()` function in R. Information criteria for all the models are displayed in Table 6.3.

```
cv_base7 <- loo(base_model_7)
cv_cov7 <- loo(cov_model_7)
cv_m7a <- loo(model_7a)
cv_m7b <- loo(model_7b)
cv_m7c <- loo(model_7c)
```

Cross-validation statistics show that the baseline model is by far the poorest at predicting the data and that the bigram frequency model (A) has a much lower LOOIC than the covariate and the other experimental models, making this the best model at predicting new data – assuming that data was drawn from an identical distribution. Closer examination of the standard deviation for each model shows that there is no meaningful difference between the covariate model, the transitional probability

model, and the combined model. Considering these differences, it is appropriate to select the bigram frequency model as the best fit for the observed data without the need for Bayes factor comparisons. In the interest of consistency, and to further confirm these results, Bayes factors were still calculated for model comparison.

Table 6.4: Leave-one-out Cross-validation Information Criteria for models based on data from Experiment 7

Model	Population-level	Group-level	LOOIC (SD)
Base	None	None	3523.90 (65.10)
Covariate	Age, letters, word frequency, concreteness	participant, item	-3499.40 (112.9)
Model A	Age, letters, word frequency, concreteness, bigram frequency	participant, item	-5440.70 (97.5)
Model B	Age, letters, word frequency, concreteness, transitional probability	participant, item	-3863.60 (97.5)
Model C	Age, letters, word frequency, concreteness, bigram diversity, transitional probability	participant, item	-3861.5 (97.1)

6.4.5 **Bayes factors**

Bayes factors were calculated as a confirmatory measure and used to compare models from Experiment 7, these can be seen in Table 6.5.

```
bf_covbase7 <- bayes_factor(cov_model_7, base_model_7, silent = TRUE)
bf_7abase <- bayes_factor(model_7a, base_model_7, silent = TRUE)
bf_7bbase <- bayes_factor(model_7b, base_model_7, silent = TRUE)
bf_7cbase <- bayes_factor(model_7c, base_model_7, silent = TRUE)
bf_acov7 <- bayes_factor(model_7a, cov_model_7, silent = TRUE)
bf_bcov7 <- bayes_factor(model_7b, cov_model_7, silent = TRUE)
bf_ccov7 <- bayes_factor(model_7c, cov_model_7, silent = TRUE)
bf_7ba <- bayes_factor(model_7b, model_7a, silent = TRUE)
bf_7ca <- bayes_factor(model_7c, model_7a, silent = TRUE)
bf_7cb <- bayes_factor(model_7c, model_7b, silent = TRUE)
```

Table 6.5: Bayes factors comparing statistical models based on data from Experiment 7

Model	Base	Covariate	A (Bigram frequency)	B (Transitional probability)
Covariate	>999			
A (Bigram frequency)	>999	>999		
B (Transitional probability)	>999	>999	<.001	
C (Combination)	>999	>999	<.001	0.05

As can be seen in table 6.5, there is strong evidence for the bigram frequency model over the baseline (>999), covariate (>999), transitional probability (1/.001 = 1000) models, and reasonable evidence versus the combined model (1/.001 = 1000). This confirms the conclusions from cross-validation and allows for a more confident interpretation of the results.

6.4.6 **Model summary**

Based on leave-one-out cross-validation and confirmatory Bayes factor comparisons, the bigram frequency model is the most likely given the observed data; this model is set out in more detail in table 6.6.

```
summary(model_7a)
```

Table 6.6: Summary of Model A, the bigram frequency model. All values are presented on a logarithmic scale where such was used in the analysis.

-	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Group-level effects:						
Item	0.09	0.01	0.07	0.11	912.00	1.00
Participant	0.05	0.01	0.04	0.06	954.00	1.00
Population-level effects:						
Intercept	6.03	0.06	5.9	6.15	1,172.00	1.00
Age	0	[-.01, 0]	[-.01, 0]	[-.01, 0]	4,341.00	1.00
Concreteness	0	0.01	-0.01	0.02	1,160.00	1.00
Letters	0	0.01	-0.01	0.01	815.00	1.00
Word frequency	0	0.01	-0.01	0.01	900.00	1.00
Bigram frequency	[-.01, 0]	[-.01, 0]	[-.01, 0]	[-.01, 0]	4,048.00	1.00
Family specific parameters:						
Sigma	0.15	[-.01, 0]	0.15	0.15	10,786.00	1.00

Estimate and Est.Error are equivalent to unstandardized coefficients and Std.Error respectively; Rhat is a measure of model convergence, values much greater than one indicate a failure to converge; Eff.Sample is the effective number of independent samples drawn by the MCMC after adjusting for autocorrelation.

6.5 **DISCUSSION**

Experiment 7 builds on the findings of Experiment 5 by once again showing bigram frequency to be a negative predictor of response times in a primed lexical decision task, although not of the same magnitude seen in the previous experiment. This once again demonstrates that transitional probability is not as certain a metric of statistical learning as published literature would suggest. This may be related to the specific paradigm in use here or could represent a deeper issue for Statistical Learning Theory more generally. For example, this could be an example of publication bias in the experiments being reported or of transitional probability masking a simpler frequency-based effect. Since the current studies do not assess the role of transitional

probability in the acquisition of new information it is not possible to be clearer as to the nature of the discrepancy between these findings and those reported in statistical learning paradigms more widely.

However, In Chapter 2, I postulated that individuals develop stronger lexical representations for items that are encountered in a wider range of contexts. This is congruent with work by Hurtado et al. (2008; also, Jones & Rowland, 2017; Rowe, 2008) which shows that diversity in care-giver speech improves children's vocabulary acquisition. Furthermore, I suggested that having multiple contextual references for a linguistic item could lead to the development of context-independent lexical representations. This is incongruent with the transitional probability hypothesis assumed by statistical learning theory since higher transitional probabilities are associated with greater predictability which, as previously discussed, may result in the development of more context-dependent lexical representations. This would explain the lack of effect in the current paradigm since such representations would be more difficult to apply to novel situations. It becomes necessary to investigate whether the effect of transitional probability is absent when participants are required to learn new information.

With respect to bigram frequency, the results herein are congruent with an experiential model of learning such as those discussed by Bybee (1998) and Tomasello (2000), amongst others. Higher bigram frequencies can be said to represent greater linguistic experience – since participants are likely to have encountered the bigram numerous times in everyday interactions – and therefore we can presume that they constitute stronger lexical representations than those encountered less frequently.

6.6 **EXPERIMENT 8**

The design and procedure were identical to the Experiment 7 with the exception that bigram diversity was manipulated rather than bigram frequency. The nature of bigram diversity is such that the manipulation in this experiment focuses on the prime rather than the target word of the bigram.

6.6.1 **Participants**

Thirty-two participants (22 females) aged between 19 and 22 years (M= 20.31, SD= 1.03) were recruited from Nottingham, UK. All participants reported English as their first language and were screened for language difficulties. Participants received research credits in exchange for their participation where applicable.

6.6.2 **Materials**

The experimental stimuli consisted of ninety bigrams and ninety non-word stimuli. Stimuli were selected in the same way as previous experiments with the exception that the target words were held constant across the high, low, and zero diversity items to reduce the potential variance stemming from individual targets. This resulted in a stimulus-set comprising thirty sets of three targets, each with a high, low, and zero diversity bigram, examples of which can be seen in Table 6.7. Bigram selection was limited by the requirement that the target-words remain constant across the three levels, which resulted in a much smaller stimulus-pool from which to select the bigrams. Moreover, only four of the target words in the reduced stimulus pool did not begin with the letter A; these were therefore removed to avoid a potential distinctiveness effect. Descriptive statistics for each level of bigram diversity can be

seen in table 6.6. It should be noted that although the manipulation in this experiment is related to the prime, it was not possible to hold the prime constant across levels of bigram frequency. This is because changing the target word does not alter the number of followers the prime has; for example, the word modern has 102 followers in the British National Corpus, this does not change whether the bigram is modern language, modern age, or even modern potato since the bigram diversity is inherent to the prime and is unaffected by the target word. However, this does mean that it is possible to hold the target word constant without compromising the range of bigram diversity in the experiment.

Table 6.7: Group descriptive statistics for levels of bigram diversity in Experiment 8

Level	Bigram_frequency	Bigram_diversity	Target_frequency	Letters	Concreteness	Transitional_probability
High Diversity	376.848925	483.5146123	152.09 (120.93)	5.46 (1.09)	2.93 (.95)	0.002 (.61)
Low Diversity	20.71221971	1.989522297	150.52 (117.47)	4.92 (.17)	3.98 (.30)	0.38 (.86)
Zero Diversity	.00 (.00)	.00 (.00)	150.72 (116.19)	5.05 (.25)	3.19 (.34)	.00 (.00)
Log-transformed values						
High Diversity	5.95 (1.18)	6.20 (1.51)	5.02 (4.78)	-	-	-6.24 (0.49)
Low Diversity	3.04 (.70)	0.69 (.58)	5.01 (4.77)	-	-	-0.98 (.86)
Zero Diversity	-13.82 (.00)	-13.82 (.00)	5.02 (4.76)	-	-	-13.82 (.00)

A constant of .000001 was added to the zero values before transformation

Table 6.8: Example stimuli for Experiment 8

Prime	Target	Group	Bigram_Diversity	Bigram_Frequency	Transitional_Probability	Letters	Concreteness	Word_Frequency
their	absence	High Diversity	1,375.00	153.00	0.00	7.00	2.31	5,780.00
who	abuse	High Diversity	597.00	13.00	0.00	5.00	2.71	3,597.00
trade	accord	High Diversity	110.00	11.00	0.00	6.00	1.57	1,159.00
common	action	High Diversity	155.00	25.00	0.00	6.00	2.86	22,099.00
under	active	High Diversity	201.00	21.00	0.00	6.00	3.32	7,290.00
large	adult	High Diversity	252.00	10.00	0.00	5.00	4.40	5,078.00
modern	age	High Diversity	102.00	59.00	0.00	3.00	2.86	21,857.00
former	air	High Diversity	122.00	11.00	0.00	3.00	4.11	19,076.00
kept	alive	High Diversity	125.00	100.00	0.00	5.00	3.14	4,254.00
few	and	High Diversity	143.00	281.00	0.00	3.00	1.52	2,682,878.00
virtual	absence	Low Diversity	6.00	15.00	0.01	7.00	2.31	5,780.00
racial	abuse	Low Diversity	7.00	22.00	0.02	5.00	2.71	3,597.00
lake	accord	Low Diversity	29.00	21.00	0.01	6.00	1.57	1,159.00
reflex	action	Low Diversity	4.00	35.00	0.05	6.00	2.86	22,099.00
chronic	active	Low Diversity	12.00	40.00	0.01	6.00	3.32	7,290.00
mature	adult	Low Diversity	15.00	12.00	0.01	5.00	4.40	5,078.00
bygone	age	Low Diversity	3.00	24.00	0.16	3.00	2.86	21,857.00
ambient	air	Low Diversity	2.00	16.00	0.08	3.00	4.11	19,076.00
eaten	alive	Low Diversity	25.00	14.00	0.02	5.00	3.14	4,254.00
acne	and	Low Diversity	1.00	10.00	0.07	3.00	1.52	2,682,878.00
ribbed	absence	Zero Diversity	0.00	0.00	0.00	7.00	2.31	5,780.00
opulent	abuse	Zero Diversity	0.00	0.00	0.00	5.00	2.71	3,597.00
nemesis	accord	Zero Diversity	0.00	0.00	0.00	6.00	1.57	1,159.00
xylophone	action	Zero Diversity	0.00	0.00	0.00	6.00	2.86	22,099.00
wicket	active	Zero Diversity	0.00	0.00	0.00	6.00	3.32	7,290.00
saccade	adult	Zero Diversity	0.00	0.00	0.00	5.00	4.40	5,078.00
jink	age	Zero Diversity	0.00	0.00	0.00	3.00	2.86	21,857.00
exclaim	air	Zero Diversity	0.00	0.00	0.00	3.00	4.11	19,076.00
habitat	alive	Zero Diversity	0.00	0.00	0.00	5.00	3.14	4,254.00
drubs	and	Zero Diversity	0.00	0.00	0.00	3.00	1.52	2,682,878.00

6.6.3 **Procedure**

The procedure was identical to Experiment 7 except that Prime-Target pairs were presented in four blocks, two of which contained twenty-three bigram trials and twenty-two non-word trials and two of which contained twenty-two bigram trials and twenty-three non-word trials.

6.6.4 **Results**

Data from Experiment 8 was trimmed and analysed using the same procedure as the previous experiments, a total of 2.04% of correct trials were removed (this did not change the pattern of results); accuracy was comparable for both words and nonwords. All response time data were log-transformed; mean response times for each participant were then analysed using a Bayesian multi-level regression.

6.6.5 **Data preparation**

Data was read into R and analysed in the same manner as previous experiments. Log transformed values were used for bigram diversity, word frequency, transitional probability, and response time; a constant of .000001 was added to all values to avoid errors resulting from items with values equal to zero. Correlations between predictors were examined using the ggpairs function from the GGally (Schloerke et al., 2018) package in R and are shown in figure 6.2; descriptive statistics are shown in table 6.9.

```
df8 <- read_csv("Exp8_data.csv") ggpairs(data = df8, columns =
    c(3, 5:6, 8, 13)) + theme(panel.grid = element_blank())
df8$log_word_freq <- log(df8$word_freq + 1e-06)
df8$log_diversity <- log(df8$diversity + 1e-06)
df8$log_trans_prob <- log(df8$trans_prob + 1e-06)
df8$log_response_time <- log(df8$response_time + 1e-06)
```

Table 6.9: Experiment 8 descriptive statistics

Variable	Mean	SD	Min	Max	Range	IQR
age	20.320	1.010	19.000	22.000	3.000	1.000
bigram_frequency	26.640	53.440	0.000	336.000	336.000	24.000
diversity	148.640	355.150	0.000	2048.000	2048.000	143.000
word_freq	116684.160	461149.540	966.000	2682878.000	2681912.000	23213.000
letters	5.130	1.230	0.000	8.000	8.000	1.000
response_time	603.950	111.640	413.000	884.000	471.000	173.000
trans_prob	0.010	0.050	0.000	0.350	0.350	0.000

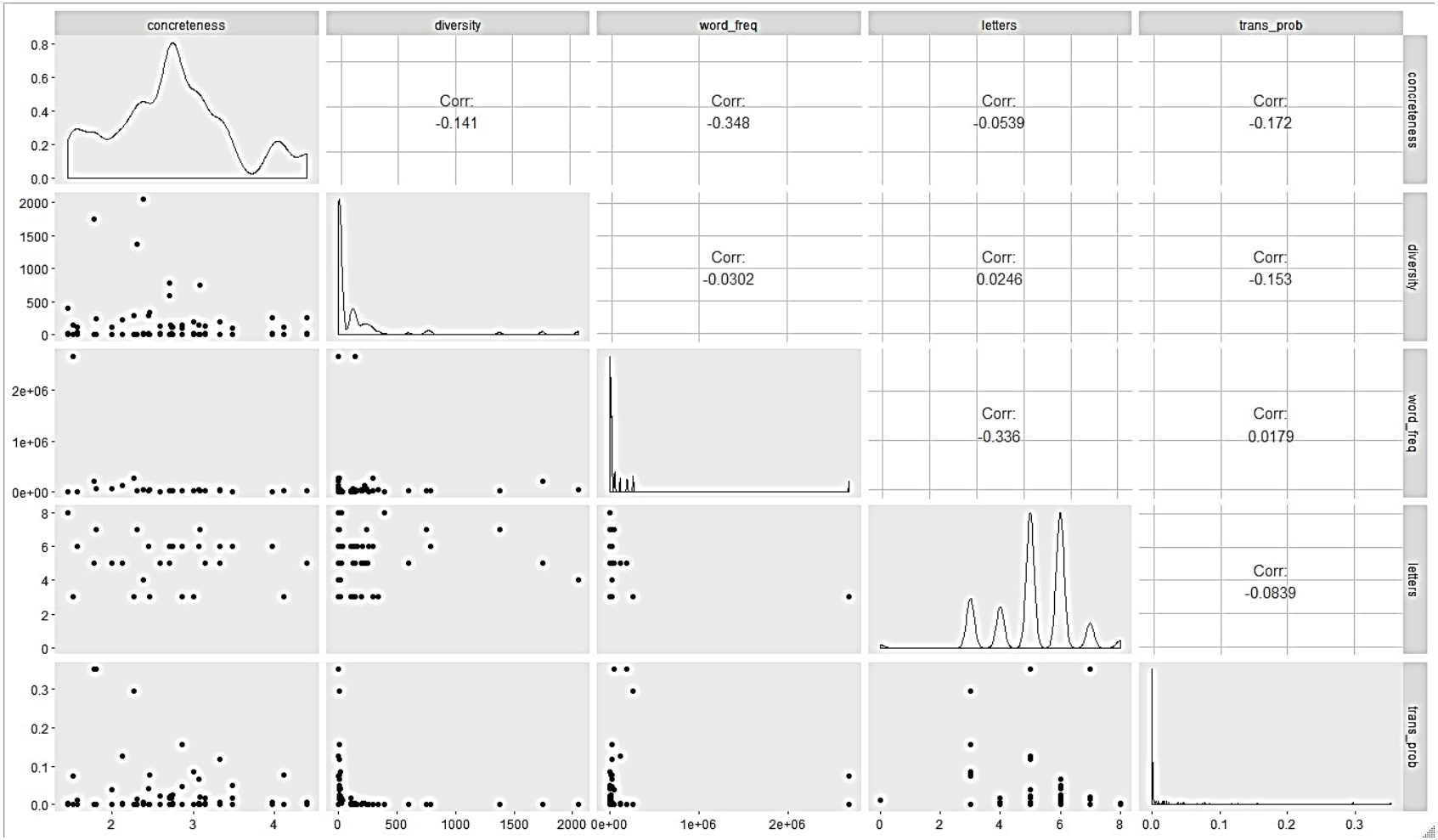

Figure 6.2: Correlation matrix for Experiment 8

6.6.6 **Specifying the models**

Log-transformed response times were modelled using Individual trial data (N = 4,169) in five random-intercept models. Individual participants and items were included as group-level effects. Bigram diversity and transitional probability were included as population-level effects, both individually and together. Target-word frequency, concreteness, target-word length, and participant age were also included as covariates. Models were run in the same way as Experiment 1 and consist of a baseline model, covariate only model, and three experimental models. Models A, B, and C examined bigram diversity, transitional probability, and both variables respectively; all models included participant age, target word frequency, concreteness, and number of letters as population-level effects and participant and item as group-level effects. Conjugate priors based on previous data were applied to each of the variables and covariates to improve the efficiency of the Monte Carlo simulation as follows: Bigram frequency = N(-.01, .01), transitional probability = N(0, .01), age = N(.01, .01), word frequency = N(0, .01), concreteness = N(0, .01), and number of letters in the target word = N(.01, .01). These priors are identical to those used in Experiment 7.

6.6.7 Define priors

```
priors_cov8 <- c(prior("normal(0, .01)", class = "b",  coef =
    log_word_freq),  prior("normal(0, .01)", class = "b", coef
    = concreteness), prior("normal(.01, .01)", class = "b",
    coef = letters), prior("normal(.01, .01)", class = "b",
    coef = age))
```

```
priors_model_a8 <- c(prior("normal(0, .01)", class = "b",  coef =
    log_word_freq),  prior("normal(0, .01)", class = "b", coef =
    concreteness), prior("normal(.01, .01)", class = "b", coef =
    letters),  prior("normal(.01, .01)", class = "b", coef =
    age), prior("normal(-.01, .01)", class = "b", coef =
    log_diversity))
priors_model_b8 <- c(prior("normal(0, .01)", class = "b",  coef =
    log_word_freq), prior("normal(0, .01)", class = "b", coef =
    concreteness), prior("normal(.01, .01)", class = "b", coef =
    letters),  prior("normal(.01, .01)", class = "b", coef =
    age), prior("normal(0, .01)", class = "b", coef =
    log_trans_prob))
priors_model_c8 <- c(prior("normal(0, .01)", class = "b",  coef =
    log_word_freq), prior("normal(0, .01)", class = "b", coef =
    concreteness), prior("normal(.01, .01)", class = "b", coef =
    letters),  prior("normal(.01, .01)", class = "b", coef =
    age), prior("normal(-.01, .01)", class = "b", coef =
    log_diversity), prior("normal(0, .01)", class = "b", coef =
    log_trans_prob))
```

6.6.8 Run Models

```
base_model_8 <- brm(log_response_time ~ 1, data = df8,  save_all_pars
    = TRUE, silent = TRUE, refresh = 0)
```

```
cov_model_8 <- brm(log_response_time ~ age + concreteness + letters +
    log_word_freq, data = df8, save_all_pars = TRUE, prior = priors_cov8,
    silent = TRUE, refresh = 0)

model_8a <- brm(log_response_time ~ log_diversity + age + concreteness +
    letters + log_word_freq + (1|subject) + (1|item), data = df8,
    save_all_pars = TRUE, prior = priors_model_a8, silent = TRUE,
    refresh = 0)

model_8b <- brm(log_response_time ~ log_trans_prob + age + concreteness +
    letters + log_word_freq + (1|subject) + (1|item), data = df8,
    save_all_pars = TRUE, prior = priors_model_b8, silent = TRUE,
    refresh = 0)

model_8c <- brm(log_response_time ~ log_diversity + log_trans_prob + age +
    concreteness + letters + log_word_freq + (1|subject) + (1|item),  data
    = df8, save_all_pars = TRUE, prior = priors_model_c8,  silent = TRUE,
    refresh = 0)
```

6.6.9 **Cross-validation**

Model comparison was performed using leave-one-out cross-validation with the `loo()` function in R. Information criteria for all the models are displayed in table 6.10.

```
cv_base8 <- loo(base_model_8)
cv_cov8 <- loo(cov_model_8)
cv_m8a <- loo(model_8a)
cv_m8b <- loo(model_8b)
cv_m8c <- loo(model_8c)
```

Table 6.10: Leave-one-out cross-validation information criteria for statistical models based on the data from Experiment 8

Model	Population-level	Group-level	LOOIC (SD)
Base	None	None	-2227.6 (67.3)
Covariate	Age, letters, word frequency, concreteness	participant, item	-1992.0 (47.8)
Model A	Age, letters, word frequency, concreteness, bigram diversity	participant, item	-1988.7 (47.6)
Model B	Age, letters, word frequency, concreteness, transitional probability	participant, item	-2000.1 (47.2)
Model C	Age, letters, word frequency, concreteness, bigram diversity, transitional probability	participant, item	-2001.6 (47.1)

Cross validation shows that the baseline model, which includes no predictors, is better than all the other models and that there is very little difference between the information criteria for the covariate, bigram diversity, transitional probability, and combined models. The difference in LOOIC is large enough that Bayes factor comparison is unnecessary but will be used to confirm what is an entirely unexpected result.

6.6.10 **Bayes factors**

Bayes factors were used for model comparison and can be seen in table 6.11.

```
bf_covbase8 <- bayes_factor(cov_model_8, base_model_8, silent = TRUE)
bf_8abase <- bayes_factor(model_8a, base_model_8, silent = TRUE)
bf_8bbase <- bayes_factor(model_8b, base_model_8, silent = TRUE)
bf_8cbase <- bayes_factor(model_8c, base_model_8, silent = TRUE)
bf_acov8 <- bayes_factor(model_8a, cov_model_8, silent = TRUE)
bf_bcov8 <- bayes_factor(model_8b, cov_model_8, silent = TRUE)
bf_ccov8 <- bayes_factor(model_8c, cov_model_8, silent = TRUE)
bf_8ba <- bayes_factor(model_8b, model_8a, silent = TRUE)
```

```
bf_8ca <- bayes_factor(model_8c, model_8a, silent = TRUE)

bf_8cb <- bayes_factor(model_8c, model_8b, silent = TRUE)
```

Table 6.11: Bayes factors comparing statistical models A, B, and C as well as the Base and Covariate only models

Model	Base	Covariate	A (Bigram diversity)	B (Transitional probability)
Covariate	<.001			
A (Bigram diversity)	<.001	<.001		
B (Transitional probability)	<.001	<.001	883.95	
C (Combination)	<.001	<.001	296.35	0.32

Surprisingly, the baseline model is more likely than all the experimental models and the covariate models. This suggests that neither bigram diversity nor transitional probability influence response times in a lexical decision task. This is contrary to what would be expected given the wealth of evidence in favour of transitional probability and confirms my earlier assertion that bigram diversity does not constitute a meaningful distributional statistic upon which learning can be scaffolded.

6.6.11 **Model summary**

The baseline model outperformed all other models when compared using cross validation and Bayes factors. This model treats response time as a constant value and assumes no effects for any of the predictors. The intercept term for the model is 6.39 on the logarithmic scale with an estimated error of less than .01.

It could be argued that the results seen in this and the previous bigram diversity experiments (4 & 6) may be an artefact of the stimuli used. Since the same stimuli were used in all three experiments, we might expect to see consistent effects for each of the predictors. This is not the case for transitional probability where we see inconsistent effects across the experiments, with some showing a negative effect and others showing no effect. However, if we consider that the transitional probability coefficient in each model is drawn from a distribution in which 95% of possible values fall between the upper and lower credible intervals then we can see that the null result shown in this experiment is plausible given the distribution of possible results in Experiments 4 and 6.

6.7 DISCUSSION

The data from Experiment 8 revealed that none of the models were able to outperform the baseline model when compared using cross-validation or Bayes factors. Given the pattern of the results seen so far in this series of experiments, this is not a particularly surprising result for bigram diversity which has performed poorly throughout. That said, I was mildly surprised to find that the covariate only model was also outperformed by the baseline since each of the covariates have a well-documented record of influencing response times in word recognition paradigms such as the one participants completed here, though it is less clear whether their effects would hold true when the targets remain constant across conditions since, in this case, the covariates are also held constant. Most surprisingly – though still somewhat consistent with the developing narrative – is the recurrent null effect of transitional probability across the set of experiments (6 & 8) – Experiments 2 and 4 also show this

effect but were conducted with different stimuli and/or timings. Throughout this work I have highlighted the strength of published evidence supporting transitional probabilities as the driving force in statistical learning and asserted the need to consider alternative metrics, but to see them consistently underperform in these experiments was somewhat unexpected.

Theoretically speaking, the consistent lack of effect from both transitional probability and bigram diversity implies that there is little to be gained from increased predictivity or contextual diversity, at least as it applies to accessing previously learnt information in a lexical decision task.

6.8 GENERAL DISCUSSION

In this chapter I set out to extend and support the findings set out in Chapter 5. Experiments 7 & 8 replicate those experiments (5 & 6) whilst holding the target items constant across each of the levels of bigram frequency and bigram diversity in order to reduce inter-item variability.

In Experiment 7, bigram frequency was shown to be a negative predictor of response times in a statistically primed lexical decision task. This is congruent with the data from Experiment 5 and supports a frequency-based account of statistical learning in which higher bigram frequencies represent greater linguistic experience. Moreover, Experiment 8 strengthens the interpretation that bigram diversity does not represent a meaningful predictor of task performance in the current paradigm.

Taken together, it becomes apparent that transitional probability is not as certain a driver of statistical learning as it would seem based on the published literature. This

has been a recurrent theme throughout this thesis and reinforces my initial argument that there needs to be a reconsideration of whether transitional probabilities should be considered the default measure of statistical regularity. Furthermore, the encouraging performance of bigram frequency in predicting task performance lends credence to the argument that perhaps a simpler, frequency-based mechanism provides a better explanation of how individuals used statistical regularities in language to scaffold their learning. There is also some suggestion – though too little to challenge the accepted narrative, at this time – that transitional probability may be masking such a frequency effect.

It must be considered however, that these results are a product of a specific, novel paradigm which has hitherto not been applied to statistical learning. As such, it is plausible that these findings are an artefact of the unconventional nature of the task rather than representative of more generalised statistical learning mechanisms. Since the task evaluates the effect of naturally occurring statistical relationships between previously learnt information any application to the acquisition of new information can only be speculative and must be applied cautiously.

In the next chapter I perform a meta-analysis using the data from Experiments 1-8 to get a more complete picture of the effects of bigram frequency, bigram diversity, and transitional probability.

CHAPTER SUMMARY

Over the course of this chapter, I:

- Conducted Experiments 7 & 8 using similar paradigms to Experiments 5 & 6 but with the target words held constant across the different levels of bigram diversity and bigram frequency
- Strengthened the argument that a frequency-based mechanism of statistical learning might be more plausible that one based on transitional probability
- Reaffirmed the conclusion that bigram diversity is unlikely to constitute a meaningful descriptor of statistical regularity

7 META-ANALYSIS

CHAPTER OVERVIEW

In this chapter I:

- Perform a meta-analysis using aggregated data from Experiments 1-8
- Use leave-one-out cross-validation and Bayes factors to select the best statistical model of the data
- Interpret the effects of bigram frequency, bigram diversity, and transitional probability in light of the meta-analysis

7.1 PREPARATION

The following libraries need to be loaded in order to complete the analyses in this chapter. set.seed() is set to 100 to ensure reproducibility.

```
library(brms)
library(readr)
library(hexbin)
set.seed(100)
```

7.2 META-ANALYSIS

The experiments presented in last two chapters examined the plausibility of bigram frequency, bigram diversity, and transitional probability as predictors of task performance in a suite of lexical decision tasks. In Experiments 5 and 7, which examined bigram frequency, the data show a small but meaningful negative contribution of bigram frequency to response time. This suggests that participants may be drawing on the existing statistical associations within bigrams in order to improve their word recognition performance. These findings are congruent with a frequency-based mechanism of statistical learning as set out in earlier chapters. The impact of bigram diversity is less clear, however, as Experiment 6 shows a positive relationship between bigram diversity and response time, but this is not supported by data from Experiment 8. This echoes the findings of the proof-of-concept studies in Chapters 3 and 4 and, at this point, it seems unlikely that learners are utilising bigram diversity to facilitate word recognition in any meaningful way. Unlike the other two metrics, transitional probability is included as a predictor in all four of the

experiments (5-8) but only shows an effect when compared with bigram diversity in Experiment 6. This lack of consistent performance lends further support to the argument that transitional probability may not be the best measure of statistical regularity in statistical learning paradigms. The coefficients for each of the Experiments (5-8) are shown in table 7.1.

Table 7.1: Coefficient estimates for all variables from Experiments 5-8.

Experiment	Age	Concreteness	Letters	Word frequency	Bigram frequency	Bigram diversity	Transitional probability
5 (Bigram frequency)	-0.01	0.01	-0.01	0.02	-0.11	-	0.46
6 (Bigram diversity)	-0.01	0.00	-0.01	-0.02	-	0	-0.02
7 (Bigram frequency)	0.00	0.00	0.00	0.00	-0.01	-	0.00
8 (Bigram diversity)	0.00	0.00	0.00	0.00	-	0	0.00

It is recognised however that these studies use relatively small sample sizes ($n < 50$) so, to allow for a more robust estimation of the effect sizes a meta-analysis of the existing data was conducted. Experiments 1 to 4 – which were presented as proof of concept studies – are similar enough to the later experiments in their design that the data from those experiments will also be used in this analysis. The data from Experiments 1-8 were therefore aggregated for use in the following meta-analysis. Since only the predictor of interest - bigram frequency in Experiments 1, 3, 5, and 7 or bigram diversity in Experiments 2, 4, 6, and 8 - was manipulated in each experiment it was possible to include both bigram frequency and bigram diversity for all individual trials, even if they were not analysed in the original experiments.

7.2.1 **Participants**

Data was aggregated for the 129 participants who participated in Experiments 1-8. There were ninety-nine female and thirty male participants aged between 18 and 41 years (M = 21.22, SD = 3.77). Since Experiments 1-6 were conducted in pairs (i.e., 1 & 2, 3 & 4, 5 & 6) some participants took part in more than one experiment. In these cases, only a single participant number was allocated. There were therefore 129 unique participants whose data was included in the meta-analysis.

7.3 **RESULTS**

The pre-trimmed data from Experiments 1-8 were used in the meta-analysis (N = 16,864). All response time data were log-transformed; mean RTs for each participant were then analysed using a Bayesian multi-level regression. Log-transformed values were also used for bigram diversity, word frequency, and transitional probability. Once again, a constant of .000001 was added to avoid errors resulting from values equal to zero.

```
dfm <- read_csv("Meta_raw.csv") ggpairs(data = dfm, columns =
     c(3:4, 8, 12:14)) +  theme(panel.grid = element_blank())
dfm$log_bigram_freq <- log(dfm$bigram_freq + .000001)
dfm$log_diversity <- log(dfm$diversity + .000001)
dfm$log_trans_prob <- log(dfm$trans_prob + .000001)
dfm$log_word_freq <- log(dfm$word_freq + .000001)
dfm$log_response_time <- log(dfm$response_time + .000001)
```

Correlations between predictors were examined using ggpairs() and are displayed in figure 7.1; no evidence of multicollinearity was observed between the predictors.

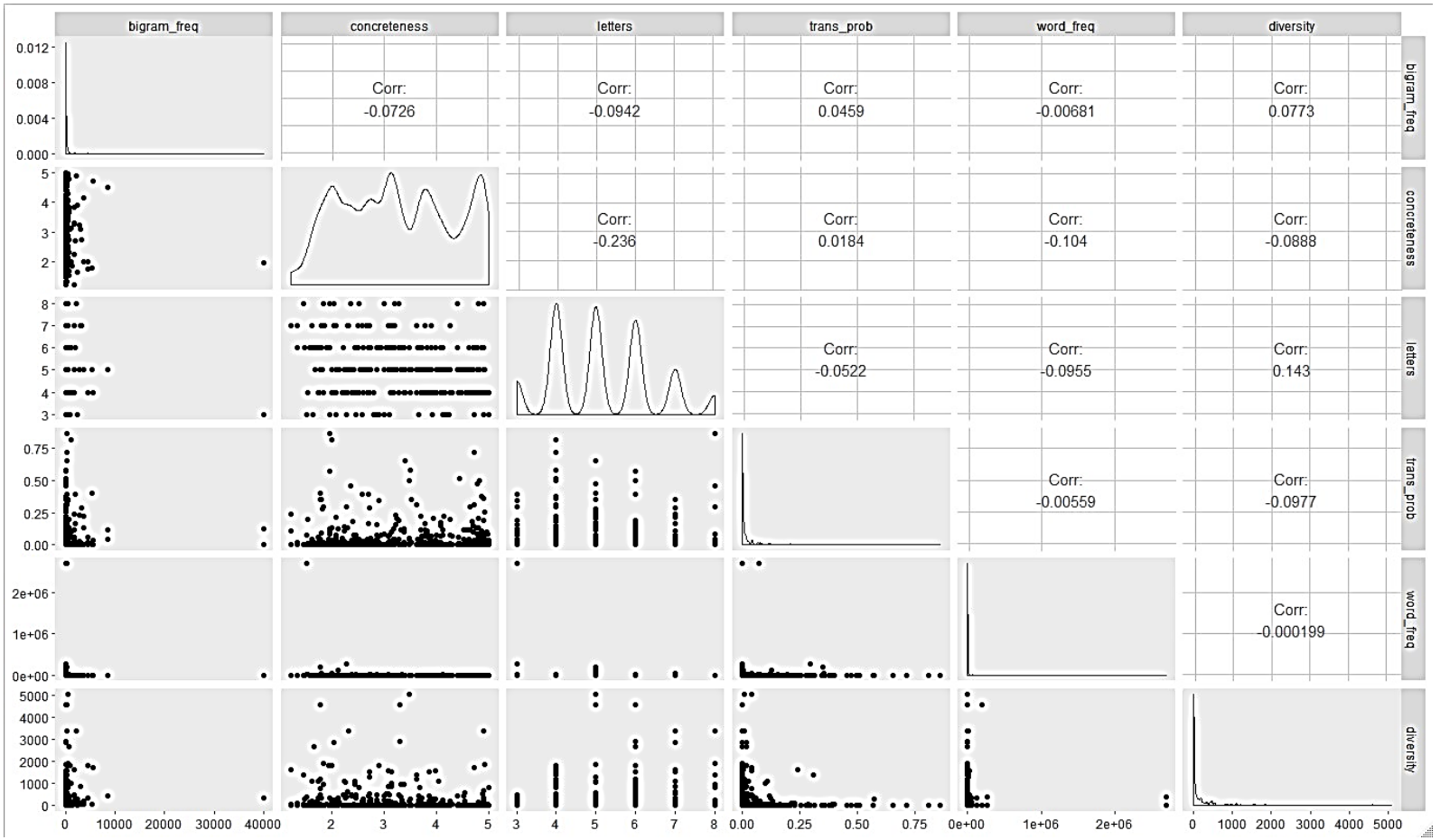

Figure 7.1: Correlation matrix for the meta-analysis of the data from Experiments 1-8.

Eight random-intercept models used individual participants and items as group-level effects and bigram frequency, bigram diversity and transitional probability as population-level effects, both individually and together. Target-word frequency, concreteness, target-word length, and participant age were also included as covariates in every model except the baseline. Normally distributed priors chosen based on the effect sizes of Experiments 1 to 8 (table 7.2) were defined for each model: Age N(0, .01), concreteness N(-.01, .02), letters N(.01, .02), word frequency N(-.01, .02), bigram frequency N(-.03, .04), bigram diversity N(0, .01), and transitional probability N(.05, .20). Note that the means and standard deviations of the effect sizes were used as a guide rather than being directly 'plugged-in' to the analysis. Given the large number of datapoints in the current analysis this is unlikely to be a problem since the prior distributions will be overwhelmed by the data when forming the posterior distribution, but we can still take steps to minimise any potential issues by increasing the variance in the priors to include a wider range of potential values. As such, the values for the priors shown above do not exactly match those seen in the table below. The priors for the variables were the same in each model with the exception of the baseline model which contains no predictors, and therefore has no prior distributions.

Table 7.2: Coefficients from Experiments 1-8 plus means and standard deviations

Experiment	Age	Concreteness	Letters	Word frequency	Bigram frequency	Bigram diversity	Transitional probability
1 (Bigram frequency)	0.01	-0.02	0.03	-0.03	-0.01	-	0.00
2 (Bigram diversity)	0.01	-0.01	0.02	-0.03	-	0	0.00
3 (Bigram frequency)	0.00	-0.01	0.01	-0.02	0	-	0.01
4 (Bigram diversity)	0.00	-0.01	0.02	-0.02	-	0	-0.04
5 (Bigram frequency)	-0.01	0.01	-0.01	0.02	-0.11	-	0.46
6 (Bigram diversity)	-0.01	0.00	-0.01	-0.02	-	0	-0.02
7 (Bigram frequency)	0.00	0.00	0.00	0.00	-0.01	-	0.00
8 (Bigram diversity)	0.00	0.00	0.00	0.00	-	0	0.00
Mean	0.00	-0.01	0.01	-0.01	-0.03	0	0.05
SD	0.01	0.01	0.02	0.02	0.04	0	0.17

Means and standard deviations were calculated using effect sizes from four experiments with the exception of bigram frequency which was calculated using effect sizes from Experiments 1, 3, 5, & 7 and bigram diversity which uses effect sizes from Experiments 2, 4, 6, & 8. All values are rounded to two decimal places.

Priors were specified using the build in prior() function in brms and can be seen below for each model individually; priors were only defined for variables included in the model.

7.3.1 **Define priors**

```
priors_cov <- c(prior("normal(-.01, .02)", class = "b",  coef =
    log_word_freq),  prior("normal(-.01, .02)", class = "b", coef
    = concreteness),  prior("normal(.01,  .02)", class = "b",
    coef = letters),  prior("normal(0, .01)", class = "b", coef =
    age))
```

```
priors_model_a <- c(prior("normal(-.01, .02)", class = "b",  coef =
    log_word_freq), prior("normal(-.01, .02)", class = "b", coef =
    concreteness), prior("normal(.01, .02)", class = "b", coef =
    letters), prior("normal(0, .01)", class = "b", coef = age),
    prior("normal(-.03, .04)", class = "b", coef =
    log_bigram_freq), prior("normal(0, .01)", class = "b", coef =
    log_diversity), prior("normal(.05, .20)", class = "b", coef =
    log_trans_prob))

priors_model_b <- c(prior("normal(-.01, .02)", class = "b",  coef =
    log_word_freq),  prior("normal(-.01, .02)", class = "b", coef =
    concreteness), prior("normal(.01, .02)", class = "b", coef =
    letters), prior("normal(0, .01)", class = "b", coef = age),
    prior("normal(-.03, .04)", class = "b", coef = log_bigram_freq),
    prior("normal(0, .01)", class = "b", coef = log_diversity))

priors_model_c <- c(prior("normal(-.01, .02)", class = "b",  coef =
    log_word_freq),  prior("normal(-.01, .02)", class = "b", coef =
    concreteness), prior("normal(.01, .02)", class = "b", coef =
    letters),  prior("normal(0, .01)", class = "b", coef = age),
    prior("normal(-.03, .04)", class = "b", coef =
    log_bigram_freq), prior("normal(.05, .20)", class = "b", coef =
    log_trans_prob))
```

```
priors_model_d <- c(prior("normal(-.01, .02)", class = "b",  coef
    = log_word_freq), prior("normal(-.01, .02)", class = "b",
    coef = concreteness), prior("normal(.01, .02)", class = "b",
    coef = letters), prior("normal(0, .01)", class = "b", coef =
    age), prior("normal(0, .01)", class = "b", coef =
    log_diversity), prior("normal(-.05, .20)", class = "b", coef =
    log_trans_prob))


priors_model_e <- c(prior("normal(-.01, .02)", class = "b",  coef
    = log_word_freq), prior("normal(-.01, .02)", class = "b",
    coef = concreteness), prior("normal(.01, .02)", class = "b",
    coef = letters),  prior("normal(0, .01)", class = "b", coef =
    age), prior("normal(-.03, .04)", class = "b", coef =
    log_bigram_freq))


priors_model_f <- c(prior("normal(-.01, .02)”, class = "b",  coef
    = log_word_freq), prior("normal(-.01, .02)", class = "b",
    coef = concreteness), prior("normal(.01, .02)", class = "b",
    coef = letters),  prior("normal(0, .01)", class = "b", coef =
    age), prior("normal(0, .01)", class = "b", coef =
    log_diversity))


priors_model_g <- c(prior("normal(-.01, .02)", class = "b",  coef
    = log_word_freq),  prior("normal(-.01, .02)", class = "b", coef
    = concreteness), prior("normal(.01, .02)", class = "b", coef =
    letters), prior("normal(0, .01)", class = "b", coef = age),
    prior("normal(-.05, .20)", class = "b", coef = log_trans_prob))
```

Once the priors had been defined, the models were run using brms in the same manner as in previous chapters.

7.3.2 **Run models**

```
base_model_meta <- brm(log_response_time ~ 1, data = dfm,  save_all_pars
    = TRUE)
covariate <- brm(log_response_time ~ age + concreteness + letters +
    log_word_freq + (1|subject) + (1|item), data = dfm,
    prior = priors_cov, save_all_pars = TRUE)
model_a <- brm(log_response_time ~ age + concreteness + letters +
    log_word_freq + log_trans_prob + log_bigram_freq + log_diversity +
    (1|subject) + (1|item), data = dfm, prior = priors_model_a,
    save_all_pars = TRUE)
model_b <- brm(log_response_time ~ age + concreteness + letters +
    log_word_freq + log_bigram_freq + log_diversity + (1|subject) +
    (1|item), data = dfm, prior = priors_model_b, save_all_pars = TRUE)
model_c <- brm(log_response_time ~ age + concreteness + letters +
    log_word_freq + log_trans_prob + log_bigram_freq + (1|subject) +
    (1|item), data = dfm, prior = priors_model_c, save_all_pars = TRUE)
model_d <- brm(log_response_time ~ age + concreteness + letters +
    log_word_freq + log_trans_prob + log_diversity + (1|subject) +
    (1|item), data = dfm, prior = priors_model_d, save_all_pars = TRUE)
```

```
model_e <- brm(log_response_time ~ age + concreteness + letters +
    log_word_freq + log_bigram_freq + (1|subject) + (1|item),
       data = dfm, prior = priors_model_e, save_all_pars = TRUE)
model_f <- brm(log_response_time ~ age + concreteness + letters +
       log_word_freq + log_diversity + (1|subject) + (1|item), data = dfm,
       prior = priors_model_f, save_all_pars = TRUE)
model_g <- brm(log_response_time ~ age + concreteness + letters +
       log_word_freq + log_trans_prob + (1|subject) + (1|item), data = dfm,
       prior = priors_model_g, save_all_pars = TRUE)
```

7.3.3 **Cross-validation and Bayes factors**

As with the previous experiments, leave-one-out cross-validation and Bayes factors were used to compare model fit.

```
cv_base <- loo(base_model_meta)
cv_cov <- loo(covariate)
cv_a <- loo(model_a)
cv_b <- loo(model_b)
cv_c <- loo(model_c)
cv_d <- loo(model_d)
cv_e <- loo(model_e)
cv_f <- loo(model_f)
cv_g <- loo(model_g)
```

Table 7.3 shows the cross-validation information criteria for each of the models. We once again see large standard error values across all the models, which limits the extent to which we can reasonably select the best model using this metric alone. However, there are some clear differences between some of the models. The inclusion of any predictors drastically improves the performance of the models compared to the baseline. Additionally, there is a clear improvement in model performance for the covariate, combined transitional probability and bigram frequency model (C), the bigram frequency model (E), and the transitional probability model (G); this lends further credence to the conclusion that bigram diversity is a poor predictor of task performance in this paradigm since it does not appear in any of the better performing models, and any models containing bigram diversity perform far worse than the covariate model at predicting the data. However, it is impossible to select between these four models based on cross-validation criteria alone, although there is a slight preference for Model C (transitional probability and bigram frequency) over the other three models.

Table 7.3: Leave-one-out cross-validation information criteria for the base, covariate, and experimental (AG) models based on data from the meta-analysis.

Model	Population-level	Group-level	LOOIC (SD)
Base			12005.5 (256.6
Covariate	Age, concreteness, letters, word frequency	Participant, Item	-3424.4 (278.3)
A	Age, concreteness, letters, word frequency, transitional probability, bigram frequency, bigram diversity	Participant, Item	-34.6 (237.6)
B	Age, concreteness, letters, word frequency, bigram frequency, bigram diversity	Participant, Item	-23.0 (237.8)
C	Age, concreteness, letters, word frequency, transitional probability, bigram frequency	Participant, Item	-3484.2 (278.0)
D	Age, concreteness, letters, word frequency, transitional probability, bigram diversity	Participant, Item	-24.0 (237.8)
E	Age, concreteness, letters, word frequency, bigram frequency	Participant, Item	-3448.2 (278.3)
F	Age, concreteness, letters, word frequency, bigram diversity	Participant, Item	-25.6 (237.8)
G	Age, concreteness, letters, word frequency, trasitional probability	Participant, Item	-3426.7 (278.3)

Bayes factors were then used to compare each model with each other model and can be seen in table 7.4, below. Although cross-validation ruled out several of the models, I have included the Bayes factor comparisons between all models for completeness and to confirm that the four models selected based on the leave-one-out information criteria are, in fact, performing better than the remaining models.

```
b1 <- bayes_factor(model_g, base_model_meta)
b2 <- bayes_factor(model_g, covariate)
b3 <- bayes_factor(model_g, model_a)
b4 <- bayes_factor(model_g, model_b)
b5 <- bayes_factor(model_g, model_c)
b6 <- bayes_factor(model_g, model_d)
b7 <- bayes_factor(model_g, model_e)
b8 <- bayes_factor(model_g, model_f)
b9 <- bayes_factor(model_f, base_model_meta)
b10 <- bayes_factor(model_f, covariate)
b11 <- bayes_factor(model_f, model_a)
b12 <- bayes_factor(model_f, model_b)
b13 <- bayes_factor(model_f, model_c)
b14 <- bayes_factor(model_f, model_d)
b15 <- bayes_factor(model_f, model_e)
b16 <- bayes_factor(model_e, base_model_meta)
b17 <- bayes_factor(model_e, covariate)
b18 <- bayes_factor(model_e, model_a)
b19 <- bayes_factor(model_e, model_b)
b20 <- bayes_factor(model_e, model_c)
b21 <- bayes_factor(model_e, model_d)
b22 <- bayes_factor(model_d, base_model_meta)
b23 <- bayes_factor(model_d, covariate)
```

```
b24 <- bayes_factor(model_d, model_a)

b25 <- bayes_factor(model_d, model_b)

b26 <- bayes_factor(model_d, model_c)

b27 <- bayes_factor(model_c, base_model_meta)

b28 <- bayes_factor(model_c, covariate)

b29 <- bayes_factor(model_c, model_a)

b30 <- bayes_factor(model_c, model_b)

b31 <- bayes_factor(model_b, base_model_meta)

b32 <- bayes_factor(model_b, covariate)

b33 <- bayes_factor(model_b, model_a)

b34 <- bayes_factor(model_a, base_model_meta)

b35 <- bayes_factor(model_a, covariate)

b36 <- bayes_factor(covariate, base_model_meta)
```

Table 7.4: Comparative Bayes factors for models in the meta-analysis of Experiments 1-8

Model	Base	Covariate	A	B	C	D	E	F
Covariate	> 999							
A - Bigram frequency, bigram diversity, transitional probability	> 999	< .001						
B - Bigram frequency, bigram diversity	> 999	< .001	> 999					
C- Transitional probability, bigram frequency	> 999	> 999	> 999	> 999				
D - Transitional probability, bigram diversity	> 999	< .001	> 999	> 999	< .001			
E - Bigram frequency	> 999	> 999	> 999	> 999	< .001	> 999		
F - Bigram diversity	> 999	< .001	> 999	1.64	< .001	< .001	< .001	
G - Transitional probability	> 999	> 999	> 999	> 999	< .001	> 999	0.01	> 999

All models also include participant age, word concreteness, word frequency, and number of letters in the target word

Looking at table 7.4, it is possible to see that the Bayes factor comparisons support the conclusions drawn from cross-validation since Models C, G, and E, as well as the covariate model all show extremely large Bayes factors when compared to the other models. Most interesting however, are the comparisons between these four models.

We can see that Model G – the transitional probability model – performs well versus the covariate model but poorly against the bigram frequency (E) and combined bigram frequency and transitional probability (C) models; it is also apparent that Model C compares favourably to Model E. This is congruent with the tentative conclusions we were able to draw from cross-validation.

7.3.4 **Model Summary**

Based on both cross-validation and Bayes factor comparison, the combined bigram frequency and transitional probability model appears to be the best model at predicting the aggregated data from the eight experiments presented thus far and is set out in full in table 7.5. As we have come to expect, the effect sizes are relatively small for each of the predictors with most of the variation coming from differences between participants as well as differences between target-words. This is to be expected given the nature of the task, as are the effects of the covariates – participants are slower at recognising longer and less concrete items and faster at responding to higher frequency targets. We also see a similar pattern of effects as in the majority of previous experiments – though these were not universally consistent – in that participants are faster when responding to higher frequency bigrams and slower when responding to bigrams with a higher transitional probability.

Table 7.5: Summary of Model C, the bigram frequency and transitional probability model

-	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Group-level:						
Item	0.08	0.01	0.07	0.09	954.000	1.010
Participant	0.18	0.01	0.17	0.2	270.000	1.020
Population-level:						
Intercept	6.36	0.04	6.28	6.45	520.000	1.000
Age	0	[0, .01]	[-.01, 0]	[0, .01]	641.000	1.000
Concreteness	[-.01, 0]	[0, .01]	-0.01	0.01	1055.000	1.000
Letters	0.01	[0, .01]	[-.01, 0]	0.01	861.000	1.000
log(Word frequency)	-0.01	[0, .01]	-0.02	-0.01	506.000	1.010
log(Bigram frequency)	-0.01	[0, .01]	-0.01	[-.01, 0]	4234.000	1.000
log(Transitional probability)	0.01	[0, .01]	[0, .01]	0.01	4271.000	1.000
Family Specific Parameters:						
Sigma	0.23	[0, .01]	0.22	0.23	7216.000	1.000

7.4 **DISCUSSION**

The meta-analysis presented in this chapter supports the conclusions I have drawn more broadly throughout this work. In a primed lexical decision task, bigram frequency and transitional probability appear to be equally good predictors of task performance. Additionally, bigram diversity – which was included to account for the predictive component of transitional probability not included in bigram frequency – has been revealed as a non-useful metric in predicting task performance. The main point of interest from the meta-analysis however, is the opposing effects of bigram frequency and transitional probability. Data from eight experiments shows that bigram frequency has a facilitatory effect of word recognition speed whereas transitional probability appears to negatively impact lexical decision performance.

These results go some way towards supporting a frequency-based mechanism of statistical learning since we see improved performance for bigrams which occur more frequently in the British National Corpus than we do for those encountered more rarely. However, the arguments set out in Chapter 2 regarding the benefits of a frequency-based mechanism – that is, lower computational difficulty and therefore reduced cognitive load compared to transitional probability – made no predictions regarding the negative impact of transitional probability. It is possible that the effects of transitional probability shown in this model are still reflecting a frequency effect, since words with a higher transitional probability tend to appear less frequently. We can examine this by visualising the data. However, we first need to remove the zero value items for both bigram frequency and transitional probability – these stem from the use of zero-value items for bigram frequency which necessarily have a transitional probability of zero due to their non-occurrence in the British National Corpus – since including them is likely to distort the final figure.

```
dfmsub <- subset(dfm, bigram_freq != 0)

dfmsub <- subset(dfmsub, trans_prob != 0)

x <- dfmsub$bigram_freq

y <- dfmsub$trans_prob
```

In order to visualise the relationship between bigram frequency and transitional probability we can display the values using a hexbin plot from the hexbin package in R (Carr, Lewin-Koh, Maechler, & Sarkar, 2019). A hexbin plot is like a scatterplot but it 'bins' similar values and displays them as graded hexagons; this results in a less messy plot and makes it easier to see where multiple values overlap by giving an indication of how densely the points are clustered.

```
bin <- hexbin(x, y, xbins = 20)
plot(bin, xlab = "Log(Bigram frequency)",
ylab = "Log(Transitional probability)")
```

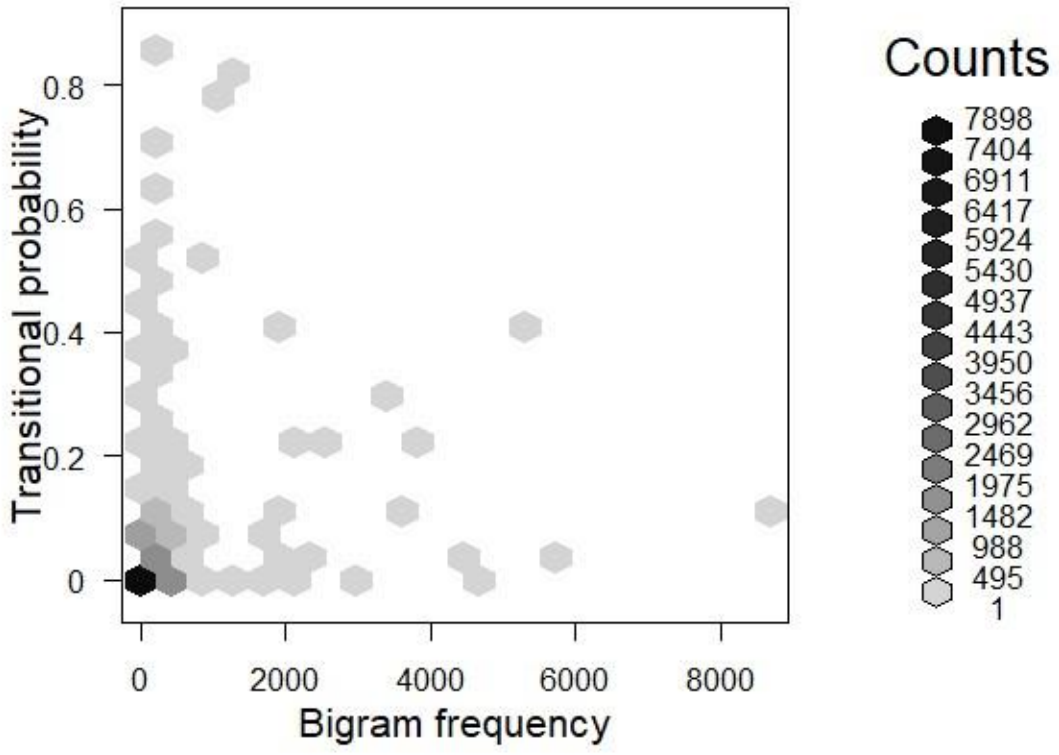

Figure 7.1: Hexbin plot showing the distribution of transitional probability and bigram frequency scores in data from Experiments 1-8 as used in the above meta-analysis; darker hexagons represent more densely clustered points. You can see that higher transitional probabilities are clustered towards to lower end of the bigram frequency scale.

As we can see in figure 7.1, there are very few high frequency, high probability items. In fact, once bigram frequency reaches around 2000 there are no bigrams with a transitional probability of greater than .5. Based on this observation, it could be suggested that transitional probability may still be drawing on a frequency effect in which higher transitional probabilities are representing lower bigram frequencies. This

is not implausible given that cross-validation shows similar information criteria for the transitional probability model (Model G), the bigram frequency model (Model E), and the combined transitional probability and bigram frequency model (Model C) but is not something that can be tested using the current data and, as such, is entirely speculative at this point. Nevertheless, the fact that bigram diversity has shown little value in predicting task performance does provide some support to this hypothesis. As transitional probability represents the probability of one item following another in natural language so too does bigram frequency, though the presumed effects of the two are necessarily inverted. Where high transitional probability represents the probability of being able to predict the target given the prime, high bigram diversity represents greater variability in the potential targets. This being the case, where we see a positive effect of transitional probability, we would expect to see a negative effect of bigram diversity; given that we do not see this inverse effect, it could be inferred that any effect of transitional probability must be frequency-based rather than related to predictability. This makes some sense since transitional probability only has a negative effect on response time in Experiment 4, where the experimental timings were specifically chosen to give participants the best possible opportunity of responding to the statistical priming, and bigram frequency was not included in the model. In all other cases, transitional probability is shown as having either a positive or null effect on response time. We cannot, however, rule out the suggestion that we do not see an effect of bigram diversity because transitional probability is simply doing a better job at capturing the predictability of the stimuli.

As I intimated in Chapter 2, it is impossible to elucidate the exact mechanism by which participants are utilising the distributional statistics within a given stimulus set – at

least with behavioural research – nor do more than speculate as to why transitional probability shows an effect in opposition to that reported in many statistical learning studies. We can, however, conclude that there is more that needs to be done if we are to uncover the mechanisms underlying statistical learning and that more attention should be paid to determining the appropriate metrics for measuring statistical learning performance. Given the novelty of examining statistical learning through previously learnt associations and in using a lexical decision task – a paradigm not usually associated with statistical learning research – to assess the strength of these associations behaviourally, I am reticent to draw clear conclusions about the effects of bigram frequency and transitional probability at this time. I am nevertheless confident in the assertion that transitional probability should not be accepted as the default measure of statistical regularity without first considering a) alternative metrics, and b) whether a more parsimonious mechanism can better explain statistical learning performance.

Throughout this thesis, I have highlighted that the current work aims to assess previously learnt associations. One of the major strengths of this approach is that it allows for the examination of naturalistic language in a way that would not be possible using a more traditional statistical learning paradigm. This begins to address one of the most fundamental criticisms of statistical learning theories – that they may not scale-up to natural language - but should be interpreted with caution. Although the British National Corpus is widely considered to be a good approximation of contemporary British English, it may not be representative of the evolving language experience; as such, it is possible that these findings may be artefactual of the corpus and any conclusions should remain tentative until such results can be independently

replicated and tested with alternative stimulus-sets. Nevertheless, the individual word frequencies within the BNC correlate highly with those in the SUBTLEX-US ($r(55863) = 0.78, p < 0.001$) (Brysbaert & New, 2009) and SUBTLEX-UK ($r(63220) = 0.91, p < 0.001$) corpora (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014) which goes some way towards vindicating the BNC as an appropriate corpus choice and suggests that comparative effects could be expected with alternative corpora. Despite the strength of this approach, it still deviates significantly from more common statistical learning paradigms and cannot reasonably be used to draw conclusions about how learners acquire new information. The next chapter presents two experiments that examine the effects of bigram frequency and transitional probability in a more traditional manner; from this point on, I will no longer be considering bigram diversity as an alternate measure of statistical regularity due to consistent null results demonstrated throughout the experiments.

CHAPTER SUMMARY

In this chapter I:

- Performed a meta-analysis using aggregated data from Experiments 1-8
- Selected the bigram frequency and transitional probability model as the best at predicting the observed data
- Speculated as to the effects of bigram frequency and transitional probability
- Dropped bigram diversity as a metric of statistical regularity

8 Sequence learning

CHAPTER OVERVIEW

In this chapter I will:

- Detail two sequence learning tasks which examine participants' ability to utilise underlying statistical patterns to acquire new information.
- Define and select the most efficient model at predicting participant performance in these tasks using cross-validation and Bayes factor comparisons.
- Directly compare specific trials within each experiment to identify whether transitional (bigram) frequency or transitional probability result in better sequence learning.
- Draw conclusions about the ability of participants to utilise different distributional statistics to acquire patterns within a stimulus-set.
- Build upon the theoretical assertions made in previous chapters

8.1 PREPARATION

The following libraries and settings are required to run the code in this chapter:

```
library(brms)
library(BEST)
library(readr)
library(moments)
library(ggplot2)
set.seed(100)
```

8.2 SEQUENCE LEARNING

Over the course of this work, I have asserted that bigram frequency and bigram diversity may represent better predictors of statistical learning performance than transitional probability. My rationale for this assertion has been that the computational cost of calculating and continually updating probabilistic representations of any given stimulus-set, particularly in natural language, is not commensurate to the benefits of maintaining such a representation. In Chapter 2, I made a case for a simpler, less cognitively effortful mechanism of statistical learning based on frequency of co-occurrence. Furthermore, since bigram diversity was shown to have no effect in any of the lexical decision experiments, there would appear to be little benefit in learners tracking this information. Over the past five chapters, I have presented evidence that suggests this to be the case.

Over the course of Experiments 1, 3, 5, and 7 bigram frequency was shown to have a small but meaningful negative relationship with task performance – as measured by response latency – in a statistically primed lexical decision task in 75% of the experiments (see table 8.1). In addition, transitional probability showed no facilitatory effect in 100% of the same experiments. In fact, in Experiments 3 and 5, transitional probability produced an effect in opposition of that which would be expected given the wealth of literature espousing transitional probability as the primary metric in statistical learning.

It is telling that the only experiments in which transitional probability appears to facilitate word recognition are those in which bigram frequency is not included. This suggests that, in these analyses, transitional probability is likely capturing the effect of bigram frequency – a concept supported by the lack of a predictability benefit shown by bigram diversity. Moreover, when this data is incorporated into the meta-analysis the benefits of transitional probability disappear and an overall effect of bigram frequency becomes apparent.

Although the effects of bigram frequency and transitional probability in table 8.1 appear to be quite small, it should be noted that these are presented on a logarithmic scale where a value of -.01 equates to a one millisecond decrease in reaction time for each one-point change in bigram frequency; so, increasing bigram frequency by one hundred would result in a significant decrease in word recognition speeds.

Table 8.1: Summary of coefficients for the lexical decision experiments and meta-analysis

Experiment	Age	Concreteness	Letters	Word frequency	Bigram frequency	Bigram diversity	Transitional probability
1 (Bigram frequency)	0.01	-0.02	0.03	-0.03	-0.01	-	0.00
2 (Bigram diversity)	0.01	-0.01	0.02	-0.03	-	0	0.00
3 (Bigram frequency)	0.00	-0.01	0.01	-0.02	0	-	0.01
4 (Bigram diversity)	0.00	-0.01	0.02	-0.02	-	0	-0.04
5 (Bigram frequency)	-0.01	0.01	-0.01	0.02	-0.11	-	0.46
6 (Bigram diversity)	-0.01	0.00	-0.01	-0.02	-	0	-0.02
7 (Bigram frequency)	0.00	0.00	0.00	0.00	-0.01	-	0.00
8 (Bigram diversity)	0.00	0.00	0.00	0.00	-	0	0.00
Meta-analysis	0.00	-0.01	0.01	-0.01	-0.01	0	0.01

Means and standard deviations were calculated using effect sizes from four experiments with the exception of bigram frequency which was calculated using effect sizes from Experiments 1, 3, 5, & 7 and bigram diversity which uses effect sizes from Experiments 2, 4, 6, & 8. All values are rounded to two decimal places.

On this data alone, bigram frequency – though not a perfect predictor of task performance – represents a more reliable predictor of performance than the more traditionally used transitional probability; the same is not true of bigram diversity, however. In Experiments 2, 4, 6, and 8 bigram diversity produced consistent null results whereas transitional probability produced effects in 50% of the experiments. Interestingly, in Experiment 4, transitional probability displays a small negative effect – as would be expected based on previous evidence - but in Experiment 6 it shows a much larger positive effect in congruence with the bigram frequency experiments. Taken together, this leads to the conclusion that bigram diversity is unlikely to be driving statistical learning in this task and that, once again, transitional probability

cannot consistently predict task performance. The null effects observed for bigram frequency may also have implications for transitional probability. Since transitional probability arises from the interplay of bigram frequency and bigram diversity, if one of these metrics has no value in predicting task performance then it follows that transitional probability may also have little value beyond the contribution of the remaining metric. In this case, since bigram diversity – and therefore, assumedly, predictability – has no predictive value in these tasks it is unsurprising that transitional probability provides little benefit that cannot already be explained by bigram frequency.

Given the inconsistent effects of the three main predictors, a meta-analysis was conducted by aggregating data from across all the experiments. The results of this analysis support the conclusions drawn for bigram frequency throughout the earlier chapters in showing a small negative relationship with response time. The meta-analysis also showed that transitional probability represented a small positive predictor – equal in size to that of bigram frequency – of response time. This once again supports the assertion that a probabilistic representation of the stimulus-set is not as beneficial as is currently believed and that a frequency-based mechanism may form stronger lexical representations which can be more reliably accessed at a later date.

However, these conclusions are based on a novel approach to statistical learning in which I examine participants' ability to utilise pre-learnt lexical associations rather than whether these metrics can be used to scaffold the acquisition of new information. This departure from traditional statistical learning paradigms, coupled with the inconsistency demonstrated throughout, allows for only tentative conclusions regarding the viability of either a frequency-based or probabilistic account of

statistical learning behaviour. For more generalisable conclusions, the two metrics must be directly contrasted using a more conventional methodology – one in which participants are required to learn new information. To this end, the current chapter presents two sequence learning experiments in which participants' ability to learn an underlying pattern without conscious awareness is examined. Firstly, Experiment 9 tests the premise that participants can benefit from the statistical regularities within the pattern with no conscious awareness of its existence and in the absence of any overt cues. The pattern in Experiment 9 contains eight potential target locations whereas Experiment 10 increases the difficulty of the task by increasing the number of locations to sixteen.

In order to differentiate the effects of transitional probability and transitional frequency in these tasks, key transitions within the sequences of each experiment were identified. These transitions vary on either transitional frequency or transitional probability whilst holding the other metric constant. This will allow us to see whether there is any effect of high versus low transitional probability when transitional frequency is held constant, by repeating the process for transitional frequency, we can infer the effects of each metric independently. It is my expectation, based on the experiments presented thus far, that learning will be more greatly influenced by transitional frequency than by transitional probability.

8.3 BIGRAM FREQUENCY AND TRANSITIONAL FREQUENCY

In Chapter 2 I introduced the term bigram frequency and have referred to this metric throughout this work. Since bigram refers to any pair of written linguistic units, it is not entirely applicable in the current experiments; as such, from this point onwards, I

will be using the terminology transitional frequency. This change in terminology represents the fact that the current experiments, and many other statistical learning paradigms, do not use linguistic stimuli; transitional frequency is therefore functionally identical to bigram frequency – as I have used it in this work - but is more generalisable to non-linguistic stimuli.

8.4 EXPERIMENT 9: EIGHT TARGETS

8.4.1 Participants

An opportunity sample was recruited from Nottingham, UK ($N = 50$); participants were all aged between 18 and 56 ($M = 25.78$, $SD = 10.47$) and reported no visual or motor problems that might interfere with their ability to complete the task.

8.4.2 Design

A repeated-measures design was used to determine whether participants can implicitly learn a sequence using the distributional statistics, when no other cues are present. The independent variable was the type of statistical information (frequency, transitional probability) and the dependent variable was the time taken to transition from one target to another, in milliseconds. Key transitions were pre-selected for comparison to directly examine the effects of high and low transitional probability and transitional frequency.

8.4.3 **Materials**

The experiment was run in OpenSesame 3.1.7 using the 'droid' Backend and displayed on an ASUS T303UA running Windows 10 in tablet mode. The experiment comprised one 16-item practice block, five 37-item sequence blocks, and one 37-item non-sequence block. The transitional probabilities – ranging from zero to one - and transitional frequencies – from one to three - of the items within the sequence were varied systematically throughout each block and can be seen in table 8.2 along with the distance (in pixels) between transitional elements. Apart from the practice block, all items were presented sequentially with no breaks. Eight target areas were presented on a 12.6" screen (resolution 1280px X 800px). Each target measured 200 X 200px and was displayed in one of eight distinct colours (Yellow, cyan, green, red, orange, lilac, blue, and pink) with a vertical/horizontal separation of 96 pixels and a diagonal separation of 135.76 pixels; a black and white star was used to indicate the target square. An example of the display can be seen in figure 8.1. Transitions are hereafter expressed using the notation X -> Y, where X is the first location and Y is the second location; for example, a notation of 1 -> 5 would indicate participants transitioning from location one to location five on the touchscreen.

Table 8.2: Transitional statistics and separation distance for trials in Experiment 9

Transition (X -> Y)	Distance (px)	Transitional frequency	Transitional probability
1 -> 5	96.00	3.00	1.00
2 -> 3	96.00	1.00	0.20
2 -> 4	192.00	2.00	0.40
2 -> 7	135.76	1.00	0.20
2 -> 8	214.66	1.00	0.20
3 -> 1	192.00	1.00	0.20
3 -> 4	96.00	1.00	0.20
3 -> 5	214.66	1.00	0.20
3 -> 6	135.76	1.00	0.20
3 -> 7	96.00	1.00	0.20
4 -> 2	192.00	2.00	0.22
4 -> 3	96.00	3.00	0.33
4 -> 4	0.00	1.00	0.11
4 -> 5	303.56	1.00	0.11
4 -> 7	135.76	2.00	0.22
5 -> 4	303.56	3.00	0.60
5 -> 7	192.00	1.00	0.20
5 -> 8	288.00	1.00	0.20
6 -> 2	96.00	1.00	1.00
7 -> 1	214.66	2.00	0.33
7 -> 3	96.00	1.00	0.16
7 -> 4	135.76	2.00	0.33
7 -> 8	96.00	1.00	0.16
8 -> 2	214.66	2.00	0.66
8-> 7	96.00	1.00	0.33

Transitions use the notation X -> Y, where X is the first target and Y is the second. Distance is the number of pixels separating the closest points of the two targets on the screen

Looking at the transitional probabilities in table 8.2, we can see that some of the transitions have high probabilities. This is typical of artificial grammar tasks and, as previously noted, represents a problem for statistical learning more generally. The high transitional probabilities within the current experiment, although highly inflated compared to those found in natural language, serve two important functions. Firstly, they allow for the comparison of key transitions to directly compare meaningfully different transitional probabilities; and, secondly, they should – if transitional

probabilities are driving learning, as suggested by the bulk of previous evidence – give participants the best chance of learning the pattern – allowing this experiment to act as a proof-of-concept for a more complex experiment with a more comprehensive range of transitional probabilities, including those approaching more naturalistic levels.

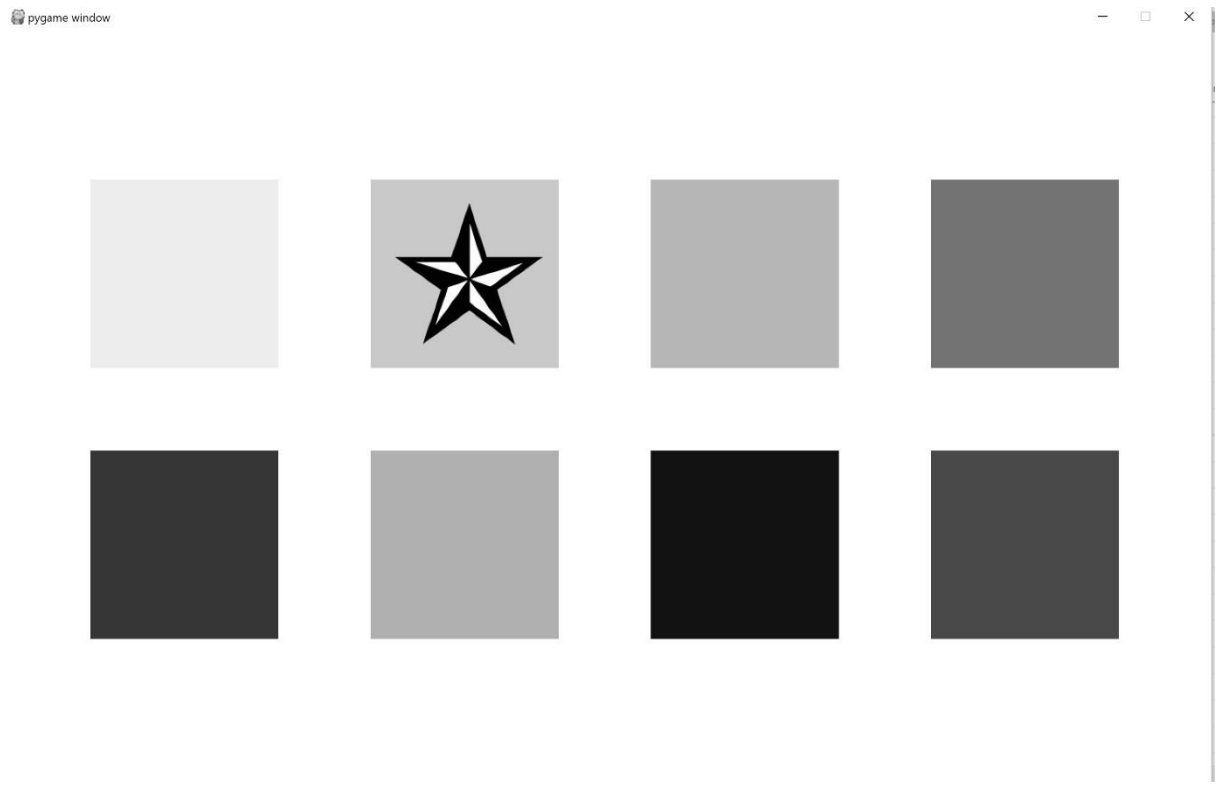

Figure 8.1: Example screenshot of the task participants undertook in Experiment 9. The black and white star indicates the target location. Locations were numbered from top left (1) to bottom right (8).

8.4.4 **Procedure**

Participants were directed to watch for a black and white star to appear in one of the target locations and to tap the star with their RIGHT index finger as soon as it appeared; participants were instructed to do this as quickly as possible. After a short practice, participants completed 222 trials comprising five repetitions of a 37-item sequence and

a final 37-items where the sequence was not present. Participants were not informed of the underlying sequence nor given any feedback throughout the experiment.

8.5 RESULTS

Mean response time by block was plotted to assess whether learning had taken place across the course of the experiment. As can be seen in figure 8.2, participants' overall performance increases over time but is adversely affected during the sixth block, when the underlying sequence is removed. This suggests that participants have become attuned to the transitional relationships between the pairs and that these associations persist even once the sequence has been removed leading to interference between the expected and actual target transitions.

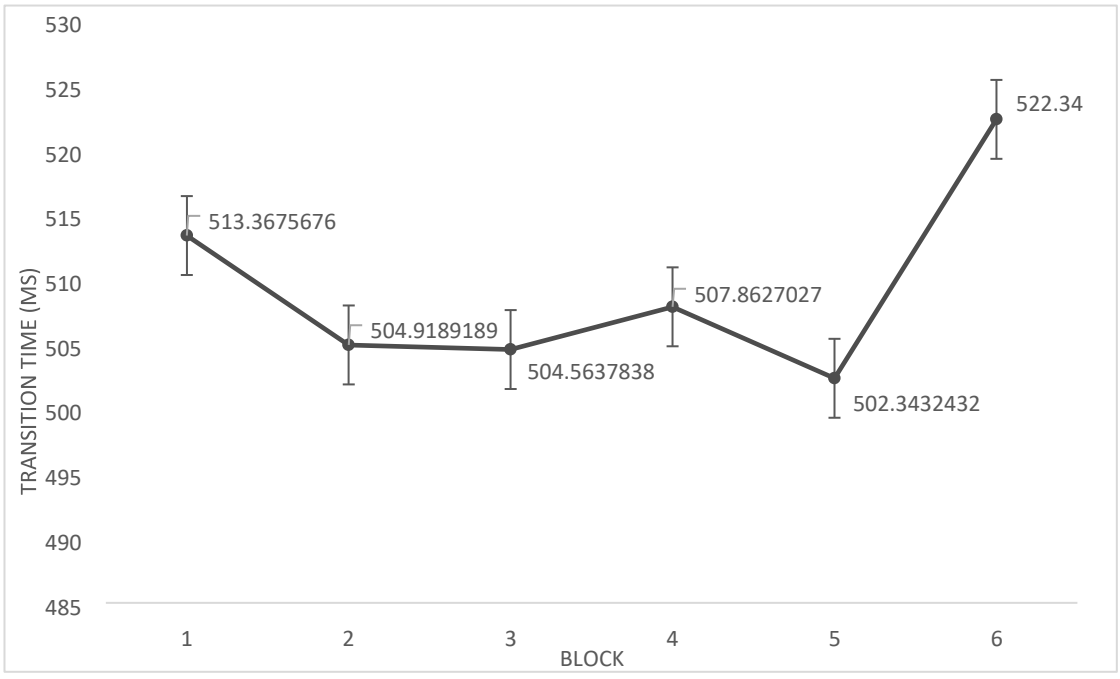

Figure 8.2: Mean response times arranged by block. Performance improves over the course of the learning blocks (1-5) but degrades in the final block (6) once the underlying pattern is removed.

8.5.1 **Data Preparation**

Data was first read into R and the distribution of response times was examined.

```
df9 <- read_csv("exp_9_block_5.csv")

den9 <- density(df9$response_time)

plot(den9, main = "", xlab = "Response time")

skewness(df9$response_time)
```

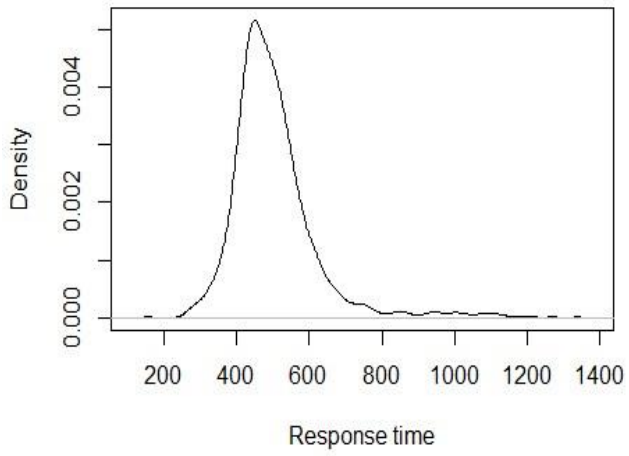

Figure 8.3: Density plot showing the distribution of transition times for Experiment 9; the distribution displays moderate positive skewness (2.30)

Since the transition time data display moderate positive skewness (figure 8.3), a log transformation was applied to the data prior to the analyses (see figure 8.4).

```
df9$log_response_time <- log(df9$response_time)

den9l <- density(df9$log_response_time)
```

```
plot(den9l, main = "", xlab = "Log(Response time)")

skewness(df9$log_response_time)
```

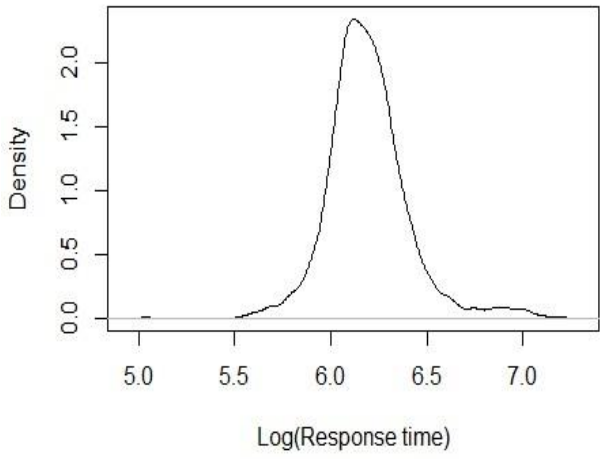

Figure 8.4: Density plot for log-transformed response times in Experiment 10. We can see that the transformation has reduced the skewness (.84) of the distribution to within acceptable parameters.

Data was modelled in R using the brms package. Data was modelled using varying combinations of fixed effects as shown in table 8.3. Model A constitutes a baseline model including participant age and inter-target distance but takes no account of either transitional frequency or transitional probability; Model B includes both age and distance but also includes transitional frequency; similarly, Model C includes the baseline predictors with the addition of transitional probability; finally, Model D includes all four of the predictors. Inter-target distance is defined as the Euclidean distance measured between the closest points of each target. Participant-level differences were also included as a group-level effect in each model. Only correct trials from block five were included in the analyses.

Table 8.3: Population- and group-level parameters for the statistical models used to analyse transition data from Experiment 9

Model	Population-level	Group-level
A	Age, distance	Participant
B	Age, distance, transitional frequency	Participant
C	Age, distance, transitional probability	Participant
D	Age, distance, transitional frequency, transitional probability	Participant

```
model_9a <- brm(log_response_time ~ age + distance + (1|subject_nr),
    data = df9, save_all_pars = TRUE, silent = TRUE)
model_9b <- brm(log_response_time ~ age + freq + distance  +
    (1|subject_nr), data = df9, save_all_pars = TRUE, silent = TRUE)
model_9c <- brm(log_response_time ~ age + tp + distance +
    (1|subject_nr), data = df9, save_all_pars = TRUE, silent = TRUE)
 model_9d <- brm(log_response_time ~ age + freq + tp + distance +
    (1|subject_nr), data = df9, save_all_pars = TRUE, silent = TRUE)
```

8.5.2 **Cross-validation**

As in previous chapters, leave-one-out cross-validation was used to identify which model has the best fit to the data, the results of which can be seen in table 8.4.

```
cv_9a <- loo(model_9a) cv_9b
<- loo(model_9b) cv_9c <-
loo(model_9c) cv_9d <-
loo(model_9d)
```

Table 8.4: Summary of brms model Leave-one out cross-validation statistics

Model	Predictors	LOOIC (SD)
A	Age, distance	-1871.10 (115.40)
B	Transitional frequency, age, distance	-1966.00 (117.60)
C	Transitional probability, distance, age	-1903.10 (114.40)
D	Transitional probability, transitional frequency, distance, age	-1964.20 (117.70)

Cross-validation shows that the transitional frequency model is marginally better than the combined transitional probability and transitional frequency model, as represented by a lower LOOIC, and that these two models are better than both the transitional probability and baseline models. However, large standard deviations mean that we cannot confidently declare any of the models as being better at describing the data from Experiment 9; therefore, comparisons between each model were performed using Bayes factors, as in previous chapters.

8.5.3 **Bayes Factors**

Bayes factors were calculated using the bayes_factor() function from the brms package in R. Models were compared with each other model to show which is most likely under the current data. Table 8.5 shows the comparisons and the associated Bayes factors.

```
bf9.1 <- bayes_factor(model_9b, model_9a)
bf9.2 <- bayes_factor(model_9c, model_9a)
bf9.3 <- bayes_factor(model_9c, model_9b)
bf9.4 <- bayes_factor(model_9d, model_9a)
bf9.5 <- bayes_factor(model_9d, model_9b)
bf9.6 <- bayes_factor(model_9d, model_9c)
```

Table 8.5: Bayes factor comparisons for Experiment 9

Model	A (Baseline)	B (Transitional frequency)	C (Transitional probability)
B (Transitional frequency)	> 999		
C (Transitional probability)	> 999	< .001	
D (Combination)	> 999	0.041	> 999

8.5.4 **Model Summary**

```
Summary(model_9b)
```

A summary of Model B – the transitional frequency model – is shown in table 8.6. As in the previous lexical decision experiments, it is apparent that transitional frequency seems to be scaffolding participants' learning. Also evident is a small effect of participant age and a non-trivial difference in transition time between participants. There also seems to be no difference in transition time based on the distance between the targets, this is likely due to there being only eight targets, all within easy reach of the participants.

Table 8.6: Model summary for the transitional frequency model

-	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Group-level:						
Participant	0.12	0.01	0.1	0.15	711.00	1.00
Population-level:						
Intercept	5.94	0.05	5.85	6.04	354.00	1.01
Age	0.01	[0, .01]	0.01	0.01	531.00	1.01
Transitional Frequency	-0.04	[0, .01]	-0.05	-0.03	7,834.00	1.00
Distance	0.00	[0, .01]	[0, .01]	[0, .01]	3,722.00	1.00
Family Specific:						
Sigma	0.14	[0, .01]	0.14	0.14	5,221.00	1.00

Estimate and Est.Error are equivalent to unstandardized coefficients and Std.Error respectively; Rhat is a measure of model convergence, values much greater than one indicate a failure to converge; Eff.Sample is the effective number of independent samples drawn by the MCMC after adjusting for autocorrelation.

8.6 **KEY COMPARISONS**

Comparisons were drawn between pre-selected target-transitions to further elucidate the relative contributions of frequency and transitional probability. The purpose of these tests was to examine the way in which high or low transitional probabilities or frequencies affect response times when the alternate statistic is held constant. Bayesian Equivalence testing was conducted using the BEST package (Kruschke & Meredith, 2018); this is functionally like conducting paired-samples *t*-tests and has a comparable interpretation. Mean scores for each transitional pair are shown in table 8.7. This notation will be used to represent the transition between locations but also the time taken for that transition. Transitions were chosen to vary on either

transitional frequency or transitional probability and to have a transitional distance of 96px; figure 8.5 depicts the numbered locations as well as the actual transitions.

Table 8.7: Mean transition times, transitional frequency, and transitional probability for each transitional pair in Experiment 9

Transition	Transitional probability	Transitional frequency	Response time (SD)
1->5	1.00	3.00	6.14 (.26)
4->3	0.33	3.00	6.13 (.19)
6->2	1.00	1.00	6.18 (.15)
8->7	0.33	1.00	6.15 (.19)

Transitions use the notation X -> Y, where X is the first target and Y is the second. Response times are given as the log-transformed mean calculated across all individual trials for that transition

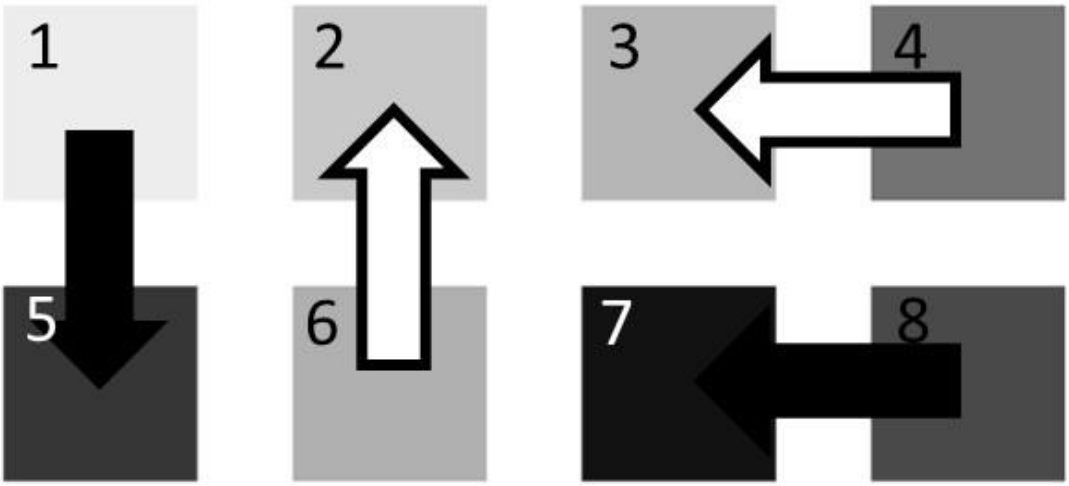

Figure 8.5: Diagram depicting the numbered locations and transitions for the key comparisons in Experiment 9

BEST uses a Monte Carlo Markov Chain to draw samples from each of the distributions to be compared and uses these values to generate a distribution of mean differences. It then calculates the percentages of the resultant distribution that are greater or smaller than zero. This distribution can then be displayed as a density plot and interpreted in relation to the mean difference of means (hereafter referred to as simply difference of means); in this case, a positive value for the difference of means indicates slower responses for the first group in the comparison. For example, the first plot shows a comparison of 1 -> 5 and 4 -> 3, so a positive value for the difference of means would indicate that participants transitioned between locations four and three more quickly than between locations one and five. Also calculated is the percentage of the distribution that falls above or below zero; for the purposes of interpretation, I will be interpreting any percentage value that falls outside of the 95% Highest Density Interval (HDI) as representing a meaningful difference between transition times.

The following code is used to separate the dataset into subsets containing only trials for each of the key comparisons before extracting the response times for use in the equivalence tests.

```
t15 <- subset(df9, key == 15)

t43 <- subset(df9, key == 43)

t62 <- subset(df9, key == 62)

t87 <- subset(df9, key == 87)

t15 <- t15$log_response_time

t43 <- t43$log_response_time

t62 <- t62$log_response_time

t87 <- t87$log_response_time
```

The first test compares transition 1 -> 5 with transition 4 -> 3; both transitions in this comparison have a transitional frequency of three (per block) but differ on transitional probability. This allows us to examine the effects of transitional probability whilst holding transitional frequency constant at the highest level available. Figure 8.6 shows the distribution of differences in means along with the proportion of the distribution that falls above or below zero.

```
t15_43 <- BESTmcmc(t15, t43)

plot(t15_43)

summary(t15_43)
```

It is apparent that, when transitional frequency is held constant at three occurrences per block and transitional probability is compared at the 1.0 and .33 level, there is no meaningful difference between the times taken to transition the first target to the second. Based on this comparison, it can be inferred that when transitional frequency is high, increasing the transitional probability confers no additional benefit in learning the underlying sequence.

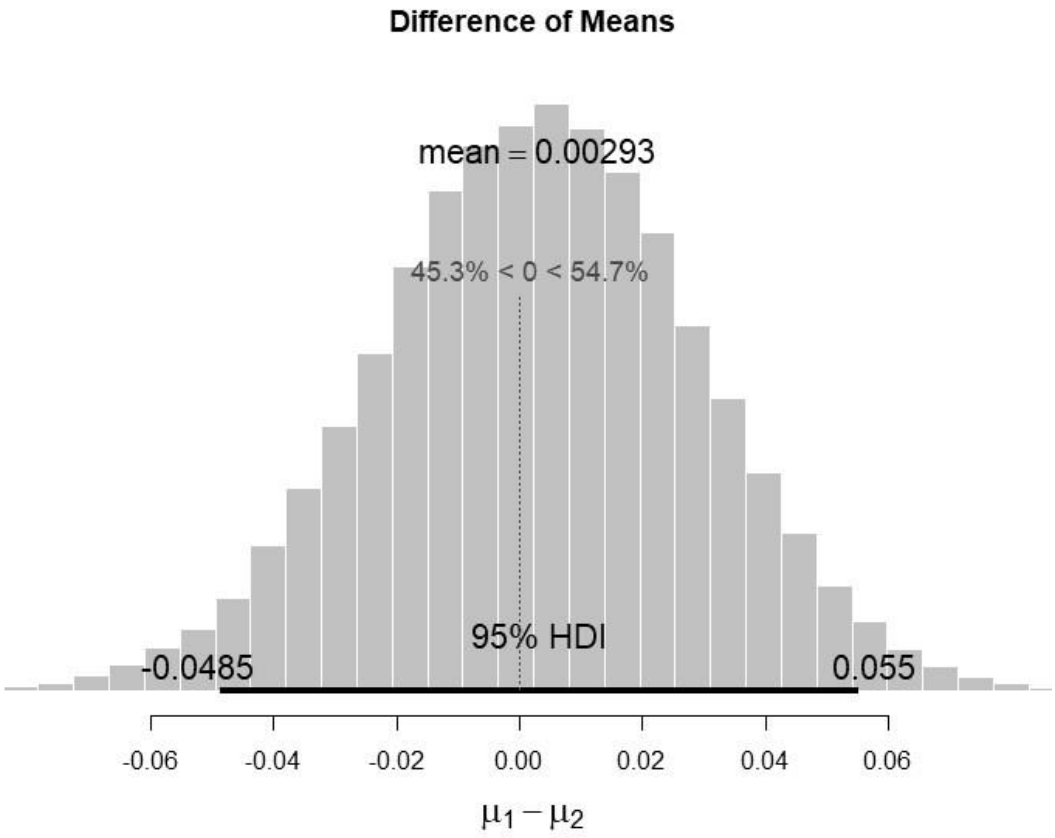

Figure 8.6: Density plot showing the mean difference between transitional pairs 1-5 and 43. The difference of means is almost equally distributed around zero indicating that there is no meaningful difference between the two datasets.

The second comparison of interest is between transitions 1 -> 5 and 6 -> 2. Here we hold transitional probability constant at 1.0 whilst contrasting trials with transitional frequencies of three and one, allowing for the effect of transitional frequency to examined in the same way as transitional probability, above.

```
t15_62 <- BESTmcmc(t15, t62)
plot(t15_62)
summary(t15_62)
```

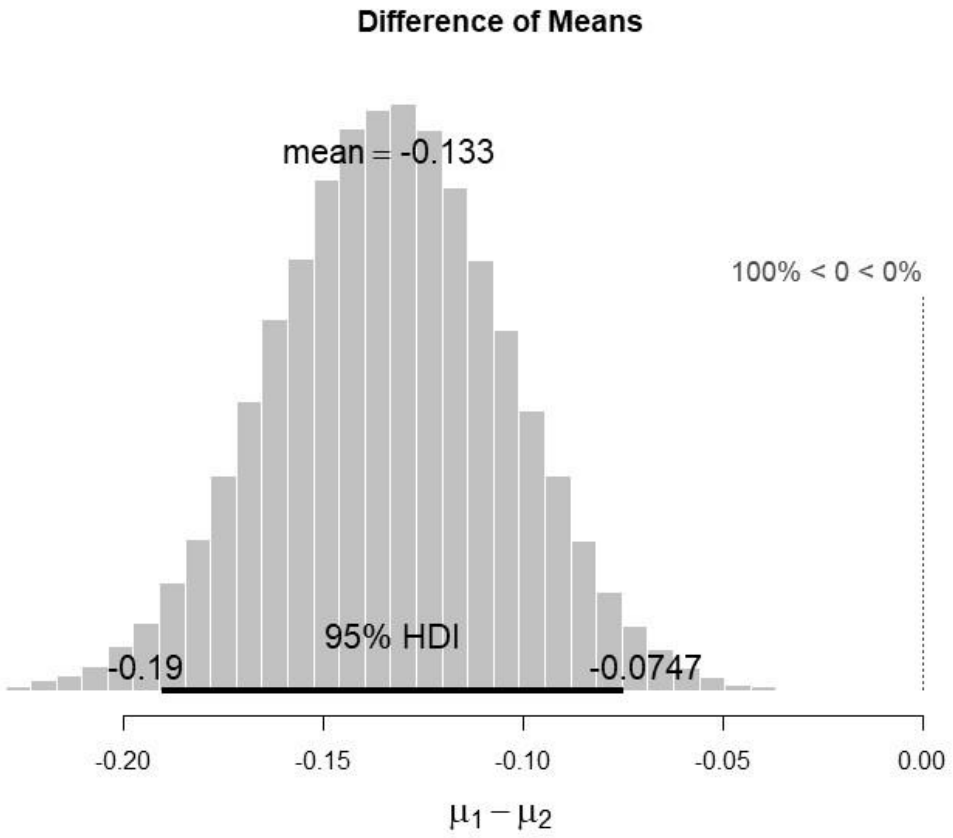

Figure 8.7: Density plot showing the difference in means between transitions 1 -> 5 and 6 > 2. There is a 100% chance that the difference in means is less than zero, indicating that transitions between targets one and five are completed more quickly than those between targets six and two.

In figure 8.7, we see a mean difference of means of -.13; this demonstrates a higher mean transition time between targets six and two than targets one and five. Since these transitions have the same transitional probability and inter-target distance, we can conclude that any differences must be a result of variations in transitional frequency and that higher transitional frequencies – as exemplified by this comparison – result in faster transition times.

Next, a comparison is made between two transitions (6 -> 2 & 8 -> 7) where the transitional frequency is held constant at one and the transitional probability once again takes a value of either .33 or 1.0.

```
t62_87 <- BESTmcmc(t62, t87)
plot(t62_87)
summary(t62_87)
```

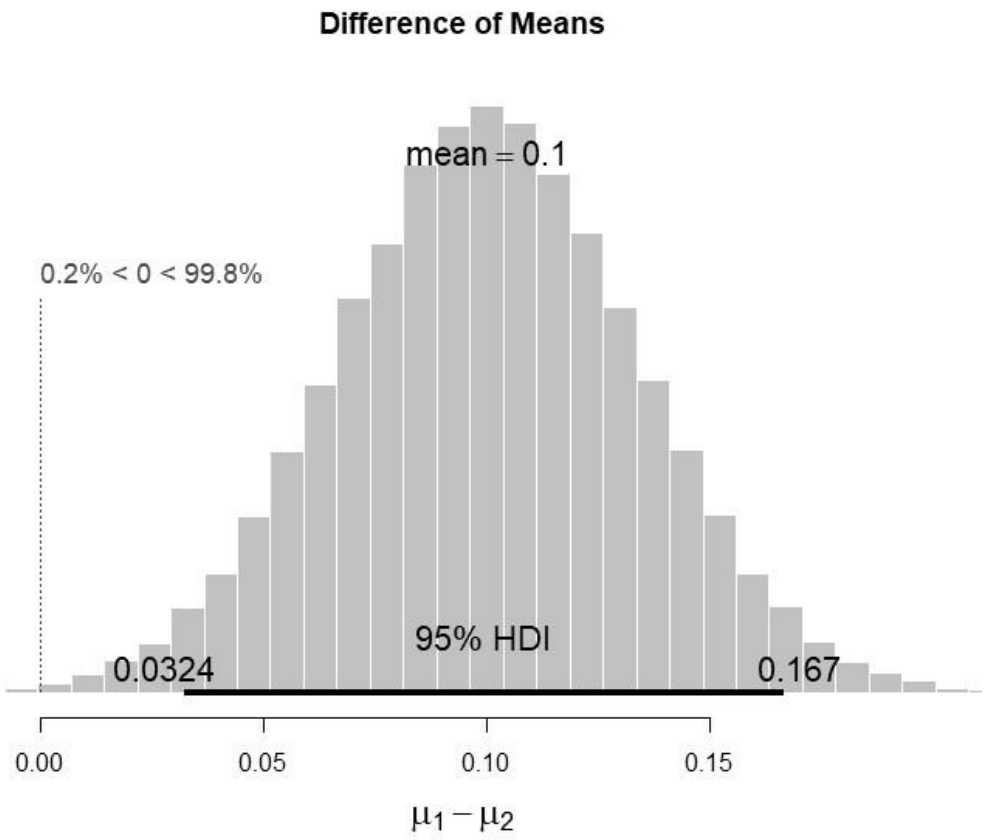

Figure 8.8: Density plot depicting the difference in group means for transitional times between items six and two and eight and seven. The mean difference of means suggests that the transitions between targets six and two are performed the quickest and that the difference between transitions has a 99.8% chance of being greeter than zero.

As can be seen in figure 8.8, there is a high percentage chance that the difference of means between transitions 6 -> 2 and 8 -> 7 is greater than zero. This indicates that the higher transitional probability results in slower transition times when frequency is held constant at one. This would be surprising given the strength of previous evidence but is congruent with the findings of the meta-analysis presented in the previous chapter.

The final comparison between transitions 4 -> 3 and 8 -> 7 examines the effect of transitional frequency when transitional probability is held constant at .33. Figure 8.9 shows that there is an 89.2% chance that the difference of means is less than zero. This suggests that the higher transitional frequency results in participants completing the transition marginally quicker than in the lower transitional frequency pair. However, the results shown here fall inside the highest density interval and are therefore rejected in accordance with the pre-defined cut-off set out above. Table 8.8 shows a summary of the equivalence tests.

```
t43_87 <- BESTmcmc(t43, t87)

plot(t43_87)

summary(t43_87)
```

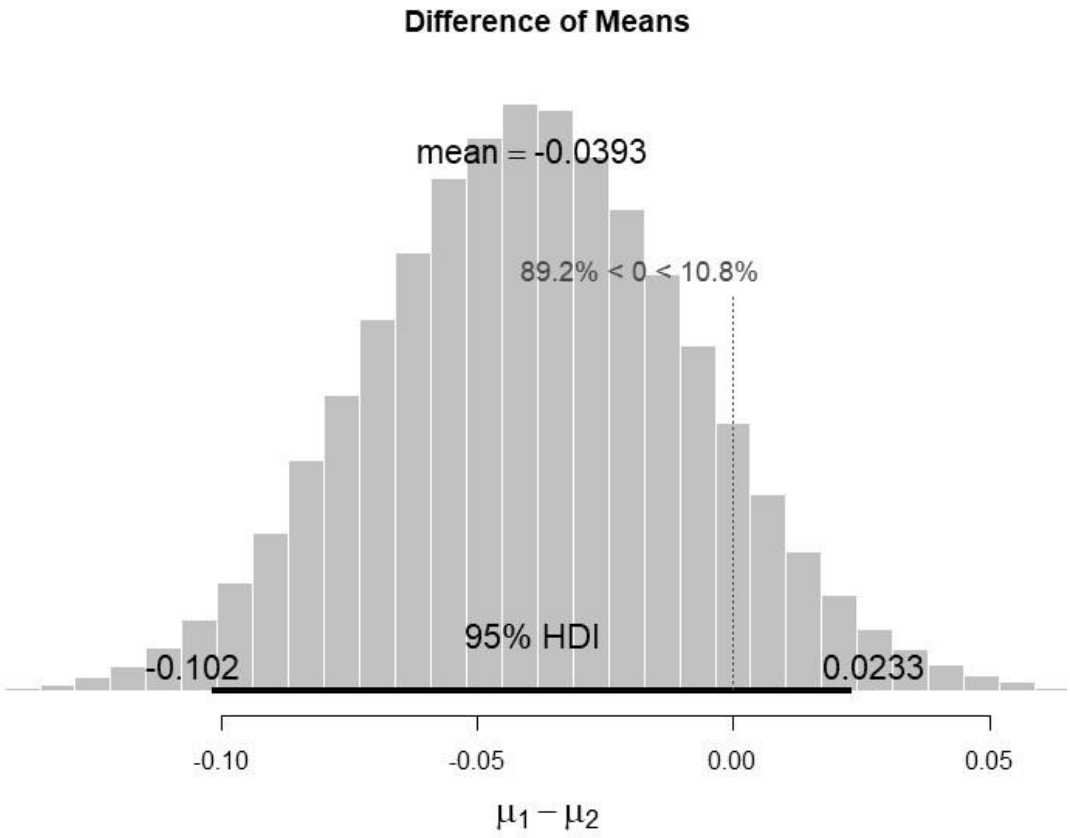

Figure 8.9: Density plot showing the difference in group means between transition 4 -> 3 and 8 -> 7. There is a non-meaningful difference between the two transitions.

Table 8.8: Summary of Bayesian equivalence tests for key comparisons from Experiment 9

Transition 1	Transition 2	M1	M2	M Diff.	SD1	SD2	SD Diff.	Effect size	% > 0	% < 0
1 -> 5	4 -> 3	6.12	6.12	0.00	0.23	0.13	0.10	0.02	54.70	45.30
1 -> 5	6 -> 2	6.11	6.25	-0.13	0.26	0.12	0.15	-0.07	0.00	100.00
6 -> 2	8 -> 7	6.25	6.15	0.10	0.12	0.19	-0.07	0.64	99.83	0.17
4 -> 3	8 -> 7	6.12	6.15	-0.04	0.13	0.17	-0.03	-0.27	10.82	89.18

M1 and M2 refer to the means of transition 1 and 2, respectively; SD 1 and 2 refer to the satandard deviations. M Diff. and SD Diff. refer to the difference in means and standard deviations between the two transitions

8.7 DISCUSSION

The current experiment aimed to investigate the relative contributions of both transitional probability and transitional frequency to performance on a simple sequence learning task. Data shows that the task was effective in eliciting learning from participants and that this learning is driven, in part, by the transitional frequency between target-pairs. Furthermore, results suggest that transitional probability represents a poorer metric of learning performance.

Direct comparisons between high and low transitional frequency and probability trials show that when transitional frequency is high there is no additional benefit in increased transitional probabilities suggesting that participants are more likely to be tracking the frequency of co-occurrence than building a probabilistic representation of the stimulus-set. This is reinforced by the fact that, when transitional probability is held constant at 1.0, higher frequency transitions are performed faster than those with a lower frequency. However, this is not true in cases where transitional probability is held constant at .33 – where the first target transitions to the second target in the transition only a third of the time. In these trials, higher frequency transitions were demonstrably faster albeit not meaningfully so. Finally, in trials where transitional frequency is low, transitional probability has an adverse effect on transition times. This is congruent with the results of the meta-analysis presented in Chapter 7 but is still somewhat surprising given that transitional probability represents the predictability of a transition, so a value of 1.0 is akin to absolute predictability of the next target.

These findings build upon the tentative conclusions discussed in previous chapters that a frequency-based mechanism of statistical learning may be preferable to a probabilistic one. The replication of results across two separate paradigms strengthens the argument that transitional probability may not be the best measure of statistical distribution for understanding statistical learning. Mathematically speaking, transitional probability represents the best descriptor of a given dataset since it captures the frequency of co-occurrence but tempers it with the number of contexts an item can appear in; it also has the advantage of providing a standardised metric that can be applied to any stimulus-set regardless of size.

That said, transitional probability is also a more complex metric to compute and maintain across larger datasets and tends to be unreasonably inflated in small-scale artificial grammar paradigms – though this is an issue of design rather than a problem with transitional probability. These issues are the basis of my argument that transitional (or bigram) frequency may be a better metric for understanding statistical learning performance. This is because it is likely less cognitively effortful to calculate and maintain frequencies than probabilities (see Chapter 2 for a more thorough discussion) and that the extra cognitive load associated with transitional probability is not commensurate to the added benefit of having a fuller, more accurate representation of the stimulus-set.

In the second part of this chapter I present a larger example of the sequence learning experiment with a greater number of targets. The larger range of potential transitions allows for a wider range of transitional probabilities weighted towards the lower end of the scale. This distribution is more akin to that seen in natural language where many

transitions display extremely small transitional probabilities, with only a few, rare transitions having probabilities approaching 1.0 (See figures 8.10 to 8.12).

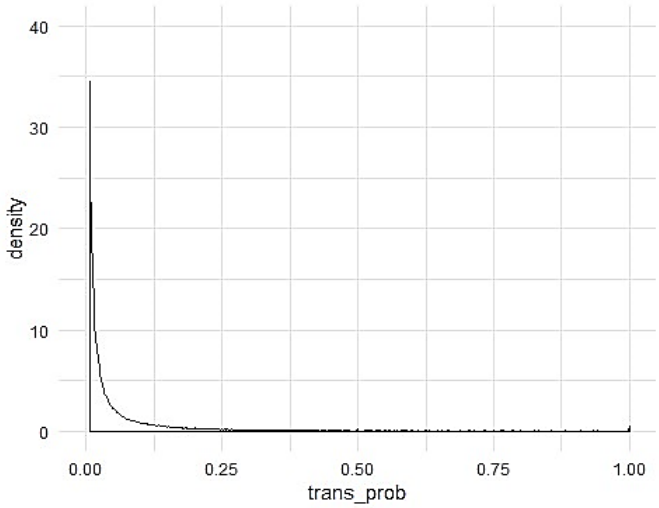

Figure 8.10: Distribution of transitional probabilities in the British National Corpus, note that the bulk of transitions are concentrated below .05 with very few exceeding .25. Repeated from Chapter 2.

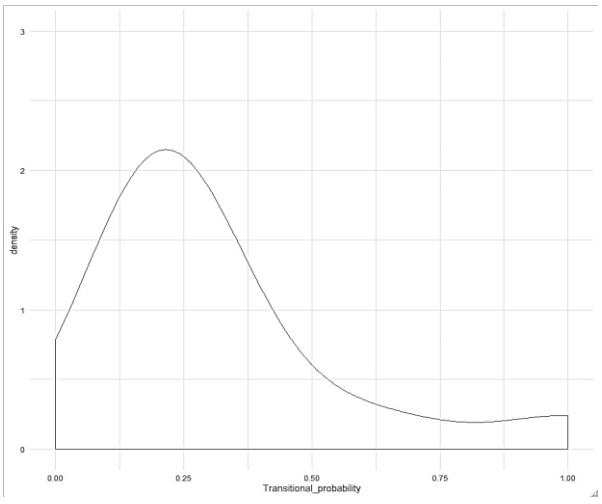

Figure 8.11: Distribution of transitional probabilities in Experiment 9. Note the much heavier tail than that seen in the density plot of transitional probabilities in the British National Corpus.

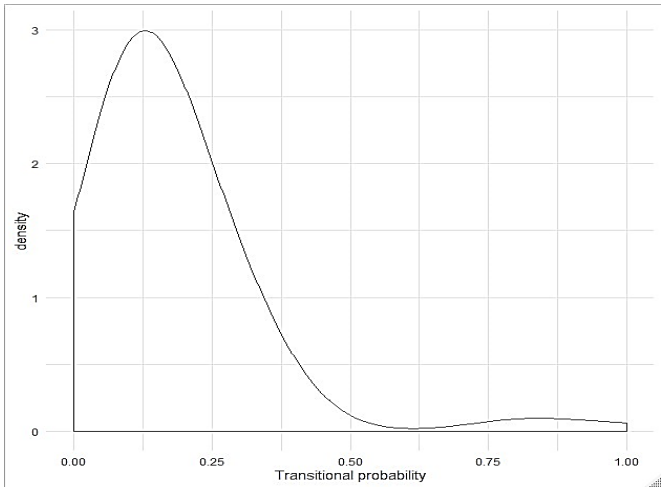

Figure 8.12: Density plot showing the probability of transitions in Experiment 10. Note the smaller tail and lower mean than that seen in Experiment 9, bringing the distribution closer to that seen in the British National Corpus.

In addition, having a more representative sample of transitional probabilities will allow for the direct comparison of values other than .33 and 1.0. This has the advantage of showing that the findings presented above are not a special case related to transitions with specific probabilities. This is particularly important for transitions with a probability of 1.0 which could be considered a special case given that the first target in such a transition is always followed by the second target allowing for perfect prediction. In fact, it is plausible to suggest that such transitions may be encoded as a single item in any representation of the stimulus-set given that they only ever occur in that specific configuration.

In summary, the greater number of targets allows for a longer sequence with smaller and more varied transitional probabilities than in Experiment 9 and enables the comparison of more realistic transitional probabilities than previously available.

8.8 EXPERIMENT 10: SIXTEEN TARGETS

Experiment 10 takes the methodology from Experiment 9 and increases the difficulty of learning the pattern by increasing the number of potential targets to sixteen. In addition, the maximum transitional frequency is raised to four and a more varied range of transitional probabilities introduced.

8.8.1 Participants

An opportunity sample of 50 participants aged between eighteen and forty-eight (M $= 24.73$, SD $= 6.97$) was recruited from Nottingham, UK. The sample was made up of 37 female and 13 male participants all of whom reported no visual or motor difficulties that may interfere with their ability to complete the task.

8.8.2 Design

A repeated-measures design was used to determine whether participants can implicitly learn a sequence and identify the mechanism driving that learning. The independent variable was the statistical information (frequency, transitional probability) and the dependent variable was the time taken to transition from one target to the next, in milliseconds. Key transitions were pre-selected for comparison to directly examine the effects of high and low transitional probability and transitional frequency.

8.8.3 **Materials**

The experiment was run in OpenSesame 3.1.7 using the 'droid' Backend and displayed on an ASUS T303UA running Windows 10 in tablet mode. The experiment comprised one 16 item practice block, five 113-item sequence blocks, and one 112item non-sequence block (See table 8.9). The transitional probabilities and paired frequencies of the items within the sequence were varied systematically throughout each block. Apart from the practice block, all items were presented sequentially with no breaks. Sixteen potential target areas were presented on a 12.6" screen (resolution 1280px X 800px). Each area measured 150 X 150px and was displayed as an empty white box on a black background. Adjacent boxes had a vertical/horizontal separation of 200 pixels and a diagonal separation of 282.84 pixels. Boxes changed colour from black to white to indicate the current target (see Figure 8.13).

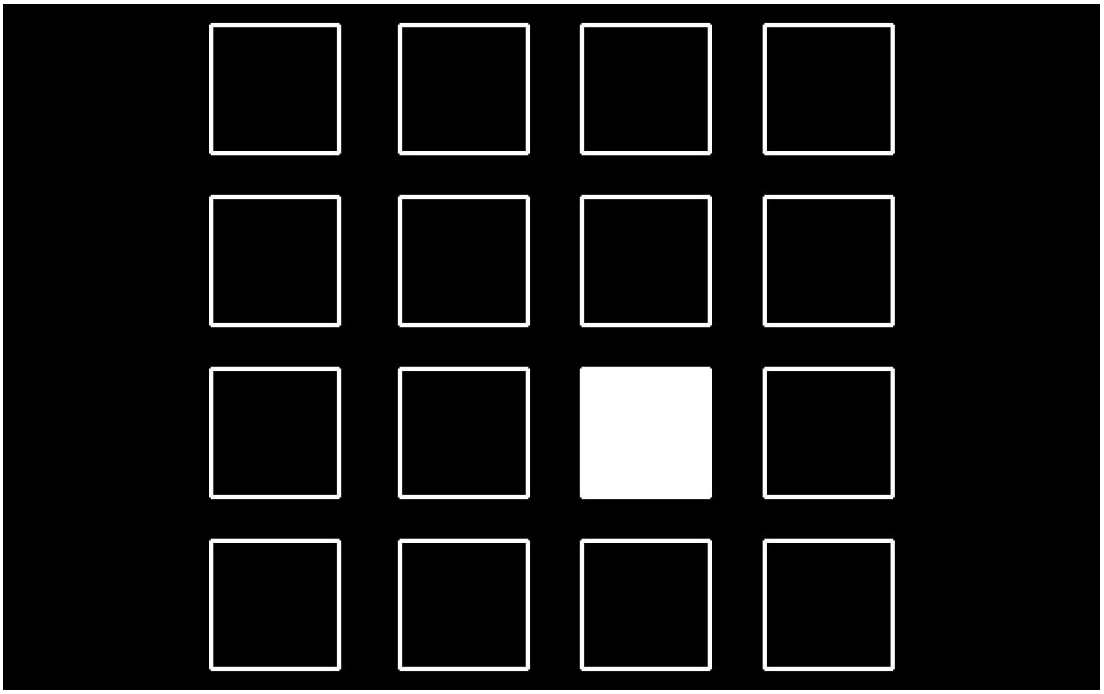

Figure 8.13: Example screenshot of the task participants undertook in Experiment 10. Blocks were labelled from A to P starting at the top left-hand corner and progressing horizontally to the bottom right of the screen; participants were unaware of this labelling.

Table 8.9: Transition table for Experiment 10

Transition (X -> Y)	Distance (px)	Transitional frequency	Transitional probability
A -> B	200.00	2.00	0.33
A -> E	200.00	1.00	0.17
A -> M	600.00	1.00	0.17
A -> O	721.11	2.00	0.33
B -> B	0.00	2.00	0.25
B -> C	200.00	1.00	0.12
B -> E	282.84	1.00	0.12
B -> H	447.21	1.00	0.12
B -> O	632.46	2.00	0.25
B -> P	721.11	1.00	0.12
C -> C	0.00	1.00	0.10
C -> E	447.21	1.00	0.10
C -> G	200.00	1.00	0.10
C -> H	282.84	1.00	0.10
C -> K	400.00	1.00	0.10
C -> L	447.21	1.00	0.10
C -> O	600.00	3.00	0.30
C -> P	632.46	1.00	0.10
D -> A	600.00	1.00	0.33
D -> K	447.21	1.00	0.33
D -> O	632.46	1.00	0.33
E -> A	200.00	1.00	0.08
E -> B	282.84	1.00	0.08
E -> C	447.21	1.00	0.08
E -> D	632.46	1.00	0.83
E -> E	0.00	2.00	0.17
E -> F	200.00	1.00	0.08
E -> H	600.00	1.00	0.08
E -> K	447.21	1.00	0.08
E -> M	400.00	1.00	0.08
E -> O	565.68	1.00	0.08

Transition (X -> Y)	Distance (px)	Transitional frequency	Transitional probability
E -> P	721.11	1.00	0.08
F -> E	200.00	1.00	0.33
F -> M	447.21	1.00	0.33
F -> O	447.21	1.00	0.33
G -> J	282.84	1.00	0.20
G -> K	200.00	1.00	0.20
G -> L	282.84	1.00	0.20
G -> M	565.68	1.00	0.20
G -> N	447.21	1.00	0.20
H -> C	282.84	1.00	0.08
H -> D	200.00	1.00	0.08
H -> E	600.00	1.00	0.08
H -> G	200.00	2.00	0.15
H -> H	0.00	1.00	0.08
H -> I	632.46	1.00	0.08
H - J	447.21	2.00	0.15
H -> L	200.00	1.00	0.08
H -> M	721.11	1.00	0.08
H -> N	565.68	1.00	0.08
H -> P	400.00	1.00	0.08
I -> C	565.68	1.00	1.00
J -> C	447.21	1.00	0.12
J -> H	447.21	1.00	0.12
J -> N	200.00	1.00	0.12
J -> O	282.84	2.00	0.25
K -> B	447.21	1.00	0.12
K -> C	400.00	1.00	0.12
K -> E	447.21	1.00	0.12
K -> G	200.00	1.00	0.12
K -> H	282.84	2.00	0.25
K -> K	0.00	1.00	0.12

Transition (X -> Y)	Distance (px)	Transitional frequency	Transitional probability
K -> M	447.21	1.00	0.12
L -> C	447.21	1.00	0.25
L -> K	200.00	1.00	0.25
L -> M	632.46	1.00	0.25
L -> O	282.84	1.00	0.25
M -> B	632.46	1.00	0.11
M -> E	400.00	2.00	0.22
M -> F	447.21	1.00	0.11
M -> H	721.11	1.00	0.11
M -> J	282.84	1.00	0.11
M -> N	200.00	1.00	0.11
M -> O	400.00	2.00	0.22
N -> H	565.68	4.00	0.80
N -> K	282.84	1.00	0.20
O -> A	721.11	4.00	0.25
O -> C	600.00	2.00	0.12
O -> E	565.68	2.00	0.12
O -> H	447.21	1.00	0.06
O -> J	282.84	1.00	0.06
O -> K	200.00	1.00	0.06
O -> L	282.84	1.00	0.06
O -> M	400.00	1.00	0.06
O -> N	200.00	1.00	0.06
O -> O	0.00	1.00	0.06
O -> P	200.00	1.00	0.06
P -> D	600.00	1.00	0.02
P -> E	721.11	1.00	0.20
P -> F	565.68	1.00	0.20
P -> G	447.21	1.00	0.20
P -> M	600.00	1.00	0.20

8.8.4 **Procedure**

Participants were informed that one of the squares would turn white and that they should with their RIGHT index finger as soon as it appeared; participants were instructed to do this as quickly as possible. After a short practice, participants completed 677 trials comprising five repetitions of a 113-item sequence and a final 112-items where the sequence was not present. Participants were not informed of the underlying sequence nor given any feedback throughout the experiment.

8.9 **RESULTS**

Mean response time by block was plotted to assess whether learning had taken place across the course of the experiment. As can be seen in figure 8.14, participants' overall performance increases over time but is adversely affected during the sixth block, when the underlying sequence is removed. This suggests that participants have become attuned to the transitional relationships between the pairs and that these associations persist even once the relationships have been adjusted.

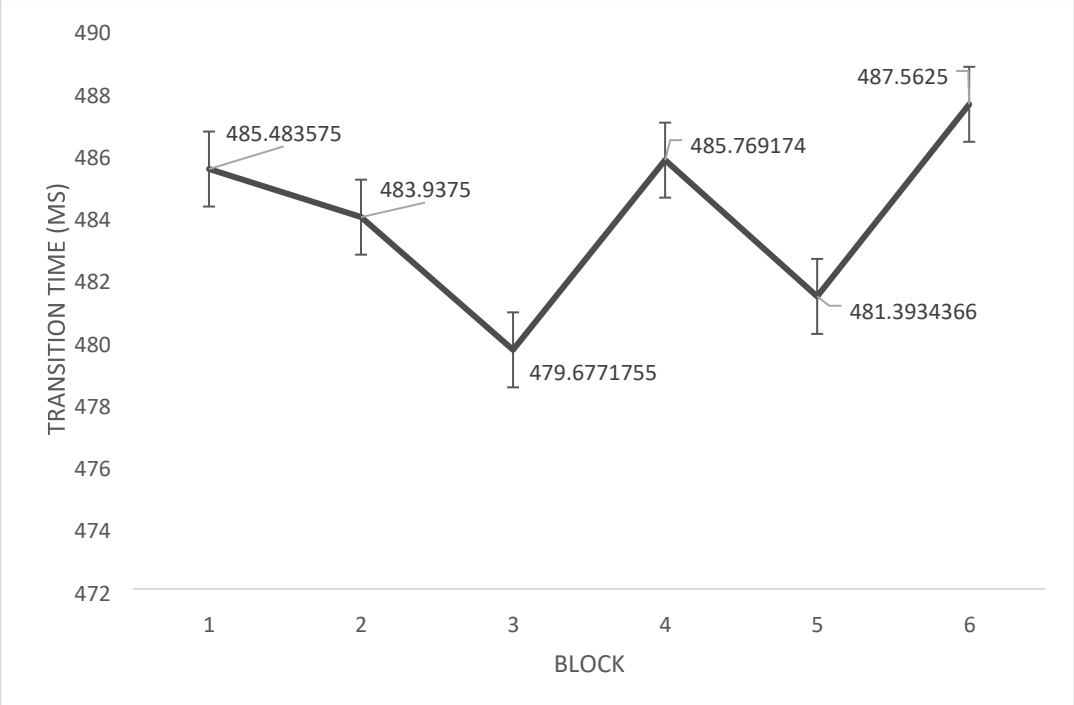

Figure 8.14: Mean response times arranged by block. Performance improves over the course of the learning blocks (1-5) but degrades in the final block (6) once the underlying pattern is removed. The bump in transition times in Block 4 may be due to a fatigue effect which is mitigated in Block 5 once participants realise, they are slowing down.

Data from Experiment 10 was read into R using the readr package and response time data – which signifies the time taken to transition from one target to another – was assessed for normality (see figure 8.15).

```
df10 <- read_csv("exp_10_block_5.csv")
den10 <- density(df10$response_time)
plot(den10, main = "", xlab = "Response time")
skewness(df10$response_time)
```

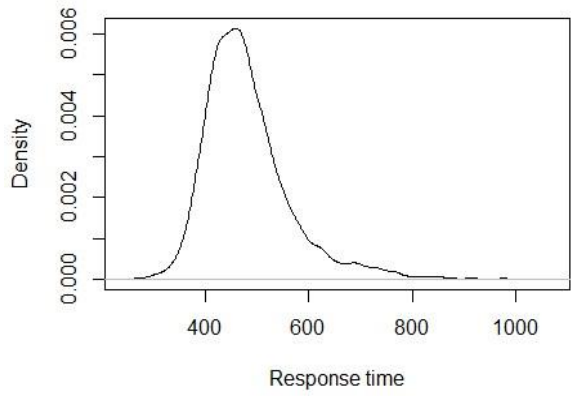

Figure 8.15: Distribution of response times for all trials in Block 5 of Experiment 10; the distribution shows a skewness of 1.42 and was therefore log-transformed to correct to normal.

Response time data was then log-transformed, and skewness reduced from 1.42 to .71 which is within acceptable parameters for the planned analyses.

```
df10$log_response_time <- log(df10$response_time)

skewness(df10$log_response_time)
```

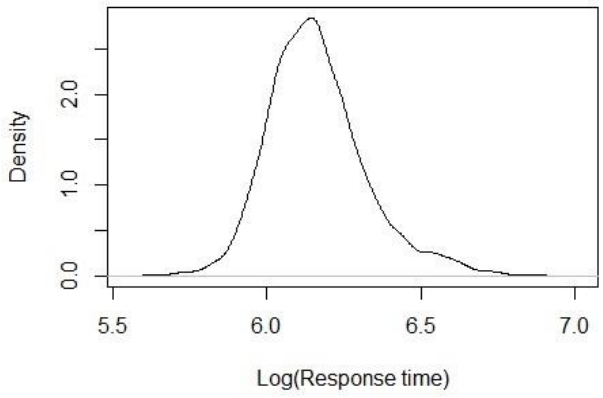

Figure 8.16: Distribution of log-transformed response times for all trials in Block 5 of Experiment 10; the distribution now shows a skewness of .71.

8.9.1 **Multi-level Model**

Only data from block five was included in the analyses. Data was modelled in R using the brms package. Data was modelled using varying combinations of fixed effects as shown in table 8.10. Participant age and inter-target distance were also included as co-variates in the models. Inter-target distance is defined as the Euclidean distance measured between the centre points of each target. Participant-level differences were also included as a random-effect in each model. Leave-one-out cross-validation was then used to identify which model provided the best fit to the data. In addition, Bayes factors were calculated to compare the weight of evidence for each of the models.

Table 8.10: Overview of the varying intercept models designated for Experiment 10

Model	Population-level	Group-level
A	Age, distance	Participant
B	Age, distance, transitional frequency	Participant
C	Age, distance, transitional probability	Participant
D	Age, distance, transitional frequency, transitional probability	Participant

```
model_10a <- brm(log_response_time ~ age + distance + (1|subject_nr),  data
    = df10, save_all_pars = TRUE, silent = TRUE)
 model_10b <- brm(log_response_time ~ age + distance + freq  +
    (1|subject_nr), data = df10, save_all_pars = TRUE, silent = TRUE)
model_10c <- brm(log_response_time ~ age + distance + tp + (1|subject_nr),
     data = df10, save_all_pars = TRUE, silent = TRUE)
model_10d <- brm(log_response_time ~ age + distance + freq + tp +
    (1|subject_nr), data = df10, save_all_pars = TRUE, silent = TRUE)
```

8.9.2 **Cross-validation**

```
cv_10a <- loo(model_10a)
cv_10b <- loo(model_10b)
cv_10c <- loo(model_10c)
cv_10d <- loo(model_10d)
```

Table 8.11 shows that model D has the smallest information criterion and therefore the best fit to the observed data; conversely model A has the largest and represents the worst fit of all the models. However, due to the large standard deviations around the leave-one-out information criteria it is impossible to accurately declare any one model better than the others. In order to discriminate effectively between the four models Bayes factors were used as a comparative tool to identify which model was most likely given the observed data.

Table 8.11: Summary of brms model leave-one out cross-validation statistics for Experiment 10

Model	Population-level	Group-level	LOOIC (SD)
A	Age, distance	Participant	-7420.50 (129.50)
B	Transitional frequency, age, distance	Participant	-7426.50 (129.20)
C	Transitional frequency, distance, age	Participant	-7450.50 (131.20)
D	Transitional probability, transitional frequency, distance, age	Participant	-7494.80 (131.80)

8.9.3 **Bayes Factors**

```
bf10.1 <- bayes_factor(model_10b, model_10a)

bf10.2 <- bayes_factor(model_10c, model_10a)

bf10.3 <- bayes_factor(model_10c, model_10b)

bf10.4 <- bayes_factor(model_10d, model_10a)

bf10.5 <- bayes_factor(model_10d, model_10b)

bf10.6 <- bayes_factor(model_10d, model_10c)
```

Bayes factors were used to compare each model with each other model and can be seen in table 8.12. There is a strong indication that Model D is most likely given the observed data from Experiment 10 and is shown in more detail in table 8.13.

Table 8.12: Bayes factor comparisons for statistical models from Experiment 10

Model	A (Baseline)	B (Transitional frequency)	C (Transitional probability)
B (Transitional frequency)	0.43		
C (Transitional probability)	> 999	> 999	
D (Combination)	> 999	> 999	> 999

8.9.4 **Model Summary**

```
Summary(model_10d)
```

Table 8.13: Summary of Model D, the combined model, based on data from Experiment 10

-	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Group-level:						
Participant	0.09	0.01	0.07	0.11	267.00	1.01
Population-level:						
Intercept	5.85	0.05	5.76	5.94	194.00	1.01
Age	0.01	[0, .01]	0.01	0.01	222.00	1.01
Transitional Frequency	0.02	[0, .01]	0.01	0.02	4,100.00	1.00
Transitional probability	-0.09	0.01	-0.11	-0.07	2,350.00	1.00
Distance	0.00	[0, .01]	[0, .01]	[0, .01]	3,722.00	1.00
Family Specific:						
Sigma	0.14	[0, .01]	0.14	0.14	5,221.00	1.00

Estimate and Est.Error are equivalent to unstandardized coefficients and Std.Error respectively; Rhat is a measure of model convergence, values much greater than one indicate a failure to converge; Eff.Sample is the effective number of independent samples drawn by the MCMC after adjusting for autocorrelation.

Model D, the combined transitional frequency and probability model shows a very different pattern of results to those seen in Experiment 9. In this, more complex, experiment there appears to be a reversal of the roles of transitional probability and transitional frequency. Participants appear to be performing better under high transitional probability, low transitional frequency conditions. This is unusual given the

findings from previous chapters but is entirely in line with the evidence presented by the wider statistical learning literature.

8.10 KEY COMPARISONS

Comparisons were drawn between pre-selected target-transitions to further elucidate the relative contributions of frequency and transitional probability. The purpose of these tests was to examine the way in which high or low transitional probabilities or frequencies affect response times when the alternate statistic is held constant. Bayesian Equivalence testing was conducted using the BEST package. Mean scores for each transitional pair are shown in table 8.14. Due to the longer, more complex sequence used in the current experiment, selecting comparisons with the same transitional frequency, transitional probability, and inter-target distance was more difficult; therefore, transitions were selected to be approximately, rather than exactly, equal on these characteristics. Note that, due to the increased number of targets, letters are used to identify the locations rather than numbers as in Experiment 9; figure 8.17 shows the target locations with their associated letters.

Table 8.14: Mean values for each transitional pair used for comparison in Experiment 10

Transition	Transitional probability	Transitional frequency	Response time (SD)
N -> H	0.80	4.00	6.11
O -> A	0.25	4.00	6.29
E -> D	0.83	1.00	6.15
L -> M	0.25	1.00	6.19

Transitions use the notation X -> Y, where X is the first target and Y is the second. Response times are given as the log-transformed mean calculated across all individual trials for that transition

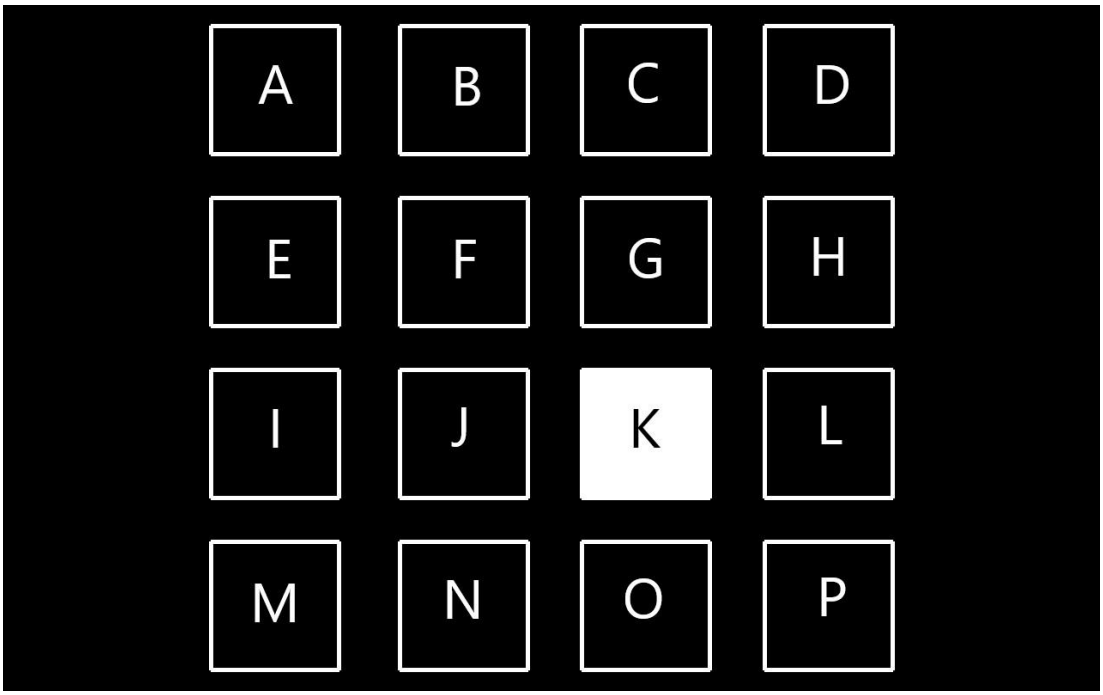

Figure 8.17: Screenshot, including labelled target locations, showing the experimental task. Note that participants were not made aware of the labels and they did not appear during the experiment.

```
tnh <- subset(df10, key == "nh")

toa <- subset(df10, key == "oa")

ted <- subset(df10, key == "ed")

tlm <- subset(df10, key == "lm")

tnh <- tnh$log_response_time

toa <- toa$log_response_time

ted <- ted$log_response_time

tlm <- tlm$log_response_time
```

Comparisons were conducted using the BEST package in R. The first comparison contrasted transitions N -> H and O -> A, both of which have a transitional frequency of four but transitional probabilities of .80 and .25, respectively.

```
tnh_oa <- BESTmcmc(tnh, toa)

plot(tnh_oa)
```

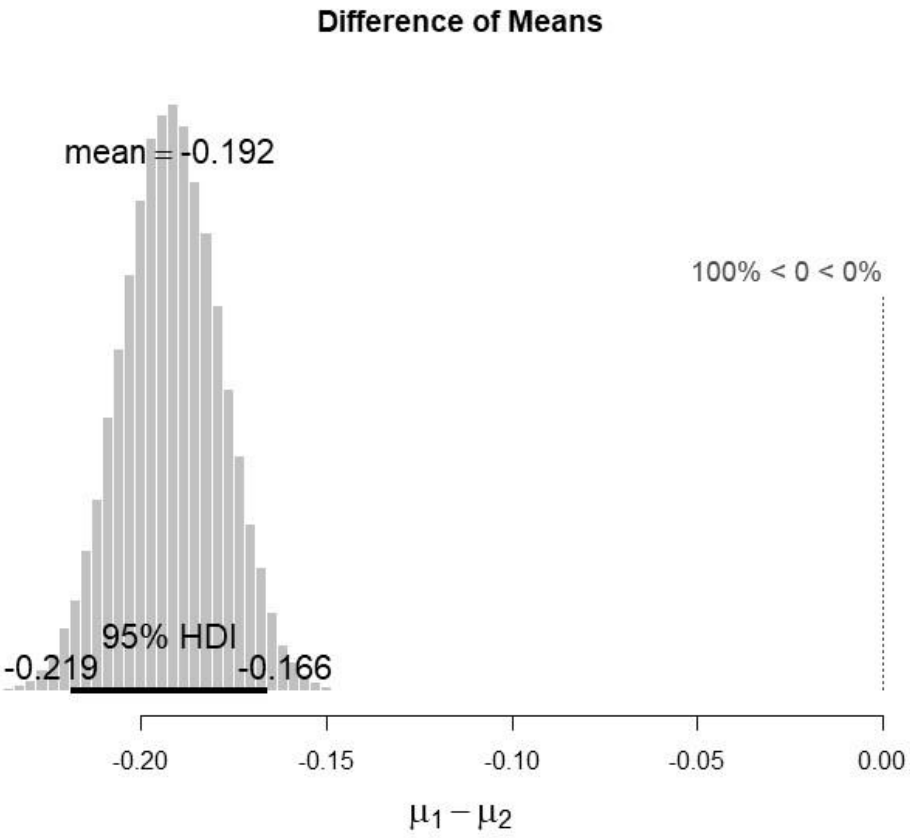

Figure 8.18: Density plot showing the mean difference between transitional pairs N -> H and O -> A. There is a 100% chance that the difference of means between the two transitions is less than zero, indicating that transition N -> H was performed more quickly.

Figure 8.18 shows that participants were quicker transitioning between targets N and H than between targets O and A. This could be attributed to the differences in distance

between the two targets, but this seems unlikely given that, in Model D above, distance is shown to have no meaningful effect on transition times. It is more likely that the difference shown here is, in fact, related to the increased transitional probability for transition N -> H compared to O -> A. It is interesting to see an effect of transitional probability in this comparison since the associated comparison in Experiment 9 showed no impact of transitional probability for high frequency transitions.

The second comparison from Experiment 10 was made between a transitions E -> D and N -> H which have transitional probabilities of .80 and .83 but transitional frequencies of four and one; the results of the comparison can be seen in figure 8.19.

```
tnh_ed <- BESTmcmc(tnh, ted)

plot(tnh_ed)
```

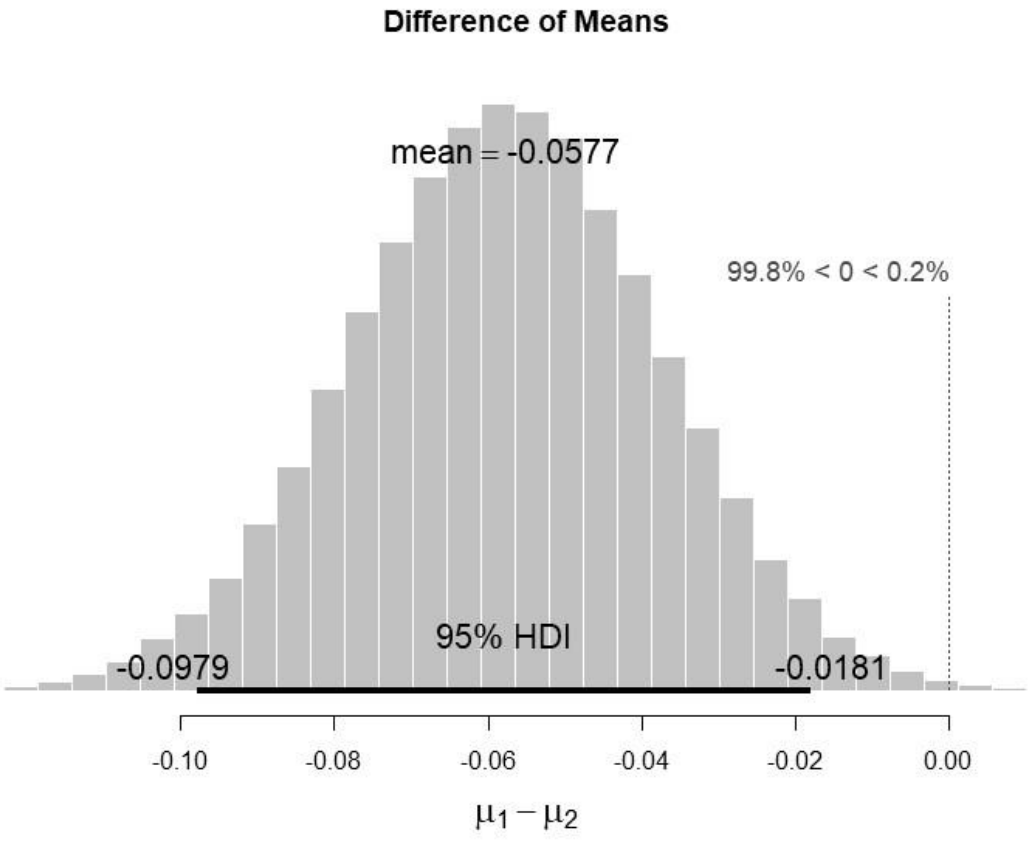

Figure 8.19: Density plot showing the difference in means between transitional pairs N -> H and E -> D. There is a 99.8% chance that the difference between group means is less than zero. This suggests that participants transition between items N and H more quickly than between items E and D .

As in Experiment 9, it is apparent that participants are responding more quickly to higher frequency transitions when transitional probability is held (almost) constant.

This is a somewhat surprising result given the results of the multi-level model, which shows participants as being slower as transitional frequency increases.

The next pair of transitions to be compared are E -> D and L -> M. These were chosen since they both have a transitional frequency of one and transitional probabilities of .80 and .25, respectively.

```
ted_lm <- BESTmcmc(ted, tlm)

plot(ted_lm)
```

Figure 8.20 shows that, when transitional frequency is held constant at the lower end of the scale, there is no meaningful difference in the time taken to transition between targets regardless of whether transitional probability is high or low. This is a departure from the findings of Experiment 9, in which higher transitional probabilities were demonstrated to be detrimental to participant performance when frequency was low.

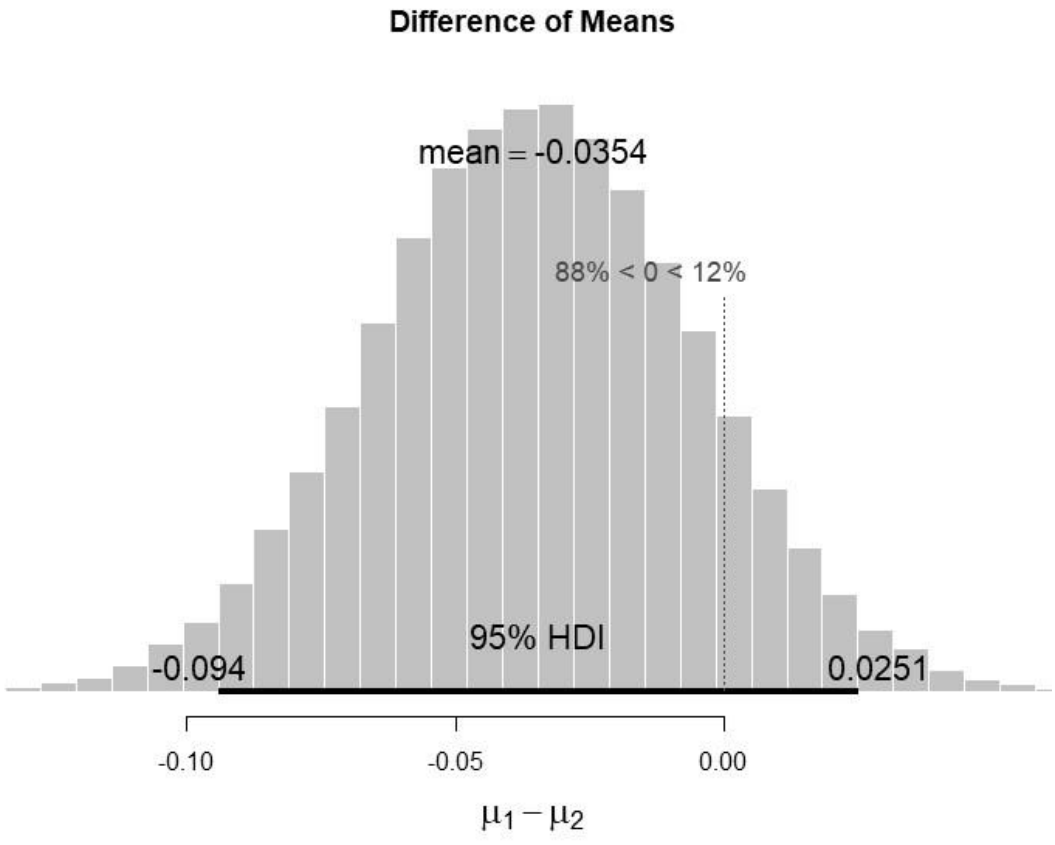

Figure 8.20: Density plot depicting the difference in group means for transitional times between targets E and D and targets L and M. There is no meaningful difference in the time taken to complete the two transitions.

The final comparison to be completed was between O -> A and L -> M. Both transitions have a probability of .25 but vary on their transitional frequency. Figure 8.21 shows that participants were 100% more likely to complete transitions from target O to target A more slowly than transitions between targets L and M. This is a highly unusual finding given the outcome of the other comparisons presented in this chapter in that, at low transitional probabilities, higher frequency transitions seem to be performed more slowly. This does, however, support the results of the multi-level model where transitional frequency is shown to be a positive predictor of response time.

```
toa_lm <- BESTmcmc(toa, tlm)

plot(toa_lm)
```

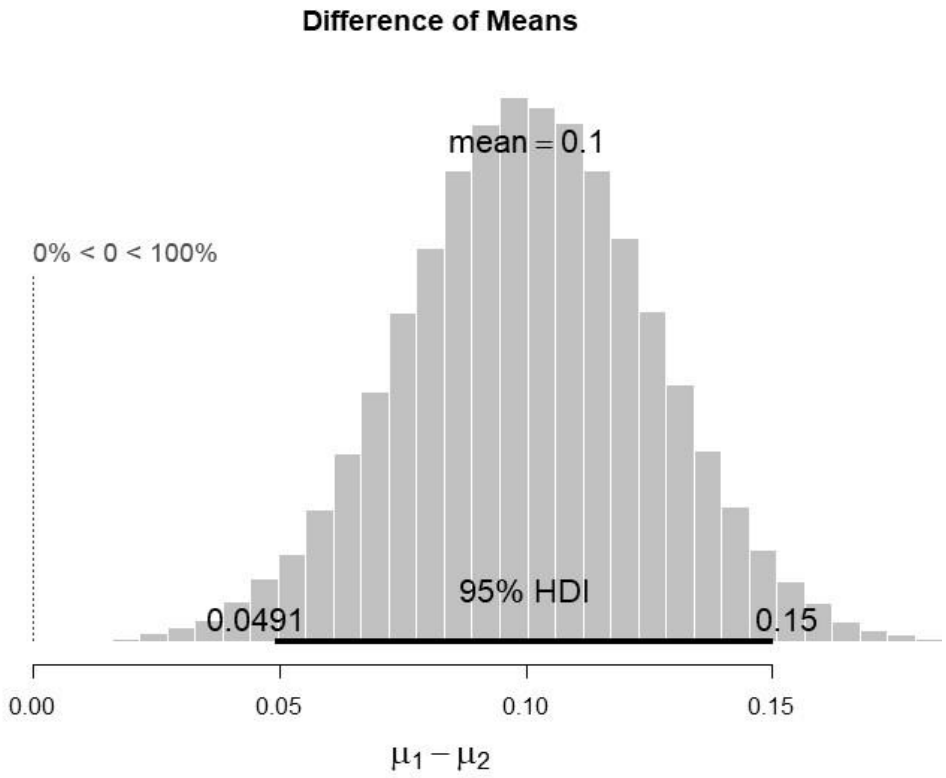

Figure 8.21: Density plot showing the difference of means between transition O -> A and transition L -> M. Transitions between the higher frequency pair are slower than those for the lower frequency transition.

Table 8.15: Summary of Bayesian equivalence tests for key comparisons from Experiment 10

Transition 1	Transition 2	M1	M2	M Diff.	SD1	SD2	SD Diff.	Effect size	% > 0	% < 0
N -> H	O -> A	6.08	6.27	-0.19	0.09	0.11	-0.02	-1.85	0.00	100.00
N -> H	E -> D	6.08	6.14	-0.06	0.09	0.10	-0.01	-0.61	0.27	99.73
E -> D	L -> M	6.14	6.18	-0.04	0.12	0.13	-0.01	-0.28	12.00	88.00
O -> A	L -> M	6.27	6.18	0.10	0.13	0.13	0.00	0.76	100.00	0.00

M1 and M2 refer to the means of transition 1 and 2, respectively; SD 1 and 2 refer to the standard deviations. M Diff. and SD Diff. refer to the difference in means and standard deviations between the two transitions

After examining the comparisons from Experiment 10, when transitional probability is high, transitions with a higher frequency are completed more quickly. Additionally, higher transitional probabilities seem to be more effective when encountered with a higher frequency. At lower frequencies, however, the effect of transitional probability appears to be negated, with the direct comparison showing no difference between transitional probabilities of .83 and .25. Perhaps more surprising is the fact that, when transitional probability is held at .25, higher frequency transitions are shown as being slower. This combination of results points towards a potential interaction between transitional probability and transitional frequency in the current experiment (see figure 8.21).

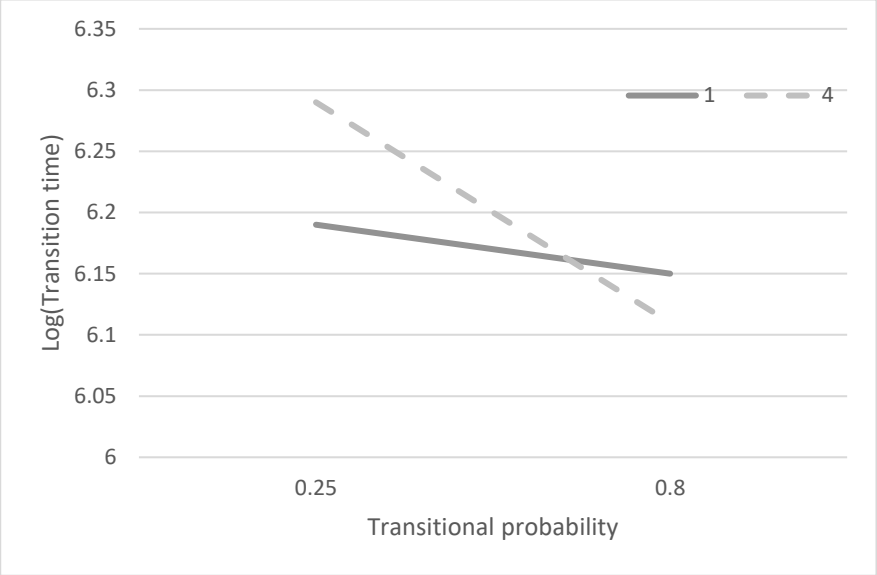

Figure 8.21: Transition times for the key comparisons in Experiment 10 plotted according to transitional probability (x-axis) and transitional frequency (blue/broken and orange/solid lines). Note that the intersection of the slopes suggests a potential interaction between the two metrics.

Based on this, a final multi-level model was compiled using the same variables as above with the addition of an interaction term. This model was then assessed using leave-one-out cross-validation and compared to the previous models for goodness of fit.

```
Model_i <- brm(log_response_time ~ age + distance + freq * tp + (1|subject_nr),
    data = df10, save_all_pars = TRUE, control = list(max_treedepth = 15))
loo(Model_i)
bayes_factor(Model_i, model_a)
bayes_factor(Model_i, model_b)
bayes_factor(Model_i, model_c)
bayes_factor(Model_i, model_d)
```

Table 8.16: Leave-one-out information criteria for Experiment 10, including the new interaction model (I)

Model	Population-level	Group-level	LOOIC (SD)
A	Age, distance	Participant	-7420.50 (129.50)
B	Transitional frequency, age, distance	Participant	-7426.50 (129.20)
C	Transitional frequency, distance, age	Participant	-7450.50 (131.20)
D	Transitional probability, transitional frequency, distance, age	Participant	-7494.80 (131.80)
I	Transitional probability, transitional frequency, distance, age	Participant	-7561.50 (133.80)

Table 8.16 shows the leave-one-out information criteria for all the models, including the interaction model and table 8.17 shows Bayes factor comparisons for the interaction model versus each of the other models.

Table 8.17: Comparative Bayes factors for the interaction model

Model	A (Baseline)	B (Transitional frequency)	C (Transitional probability)	D (Combination)
Interaction Model	> 999	> 999	>999	>999

The new model, which includes an interaction between transitional frequency and transitional probability shows the lowest information criterion and favourable Bayes factors compared to the other models. If we interpret the data based on this new model rather than the combined model (above) then we see that both the interaction between transitional frequency and transitional probability results in improved transition times as both factors increase. That is, participants perform better when transitions are both high frequency and high probability; moreover, at lower frequencies, there is no benefit to increasing transitional probabilities since there is too little exposure for participants to differentiate between them. There also appears to be detrimental effect of repeatedly exposing participants to low probability

transitions, though the reason for this is unclear. Table 8.18 provides a summary of the interaction model.

Table 8.18: Summary of the interaction model

-	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Group-level:						
Participant	0.090	0.01	0.07	0.11	300.000	1.010
Population-level:						
Intercept	5.820	0.05	5.73	5.91	141.000	1.030
Age	0.010	[0, .01]	0.01	0.01	192.000	1.020
Transitional frequency	0.040	[0, .01]	0.03	0.04	1525.000	1.000
Transitional probability	0.030	0.02	[-.01, 0]	0.07	1142.000	1.000
Transitional frequency*transitional probability	-0.060	0.01	-0.08	-0.05	1025.000	1.000
Distance	0.000	[0, .01]	[0, .01]	[0, .01]	3785.000	1.000
Family Specific:						
Sigma	0.110	[0, .01]	0.11	0.11	3406.000	1.000

Since both transitional probability and transitional frequency are continuous variables it is impractical to plot the marginal effects of the interaction in full since we would need to plot separate lines for every possible value. However, figure 8.22 shows how the effect of transitional probability varies as a function of transitional frequency; as the transitional frequency increases the effect of transitional probability becomes more pronounced. This is discussed below.

```
Conditions <- data.frame(tp = c(.25, .5, .75, 1))
Plot(marginal_effects(Model_i, effects = “freq”, conditions =
    Conditions))
```

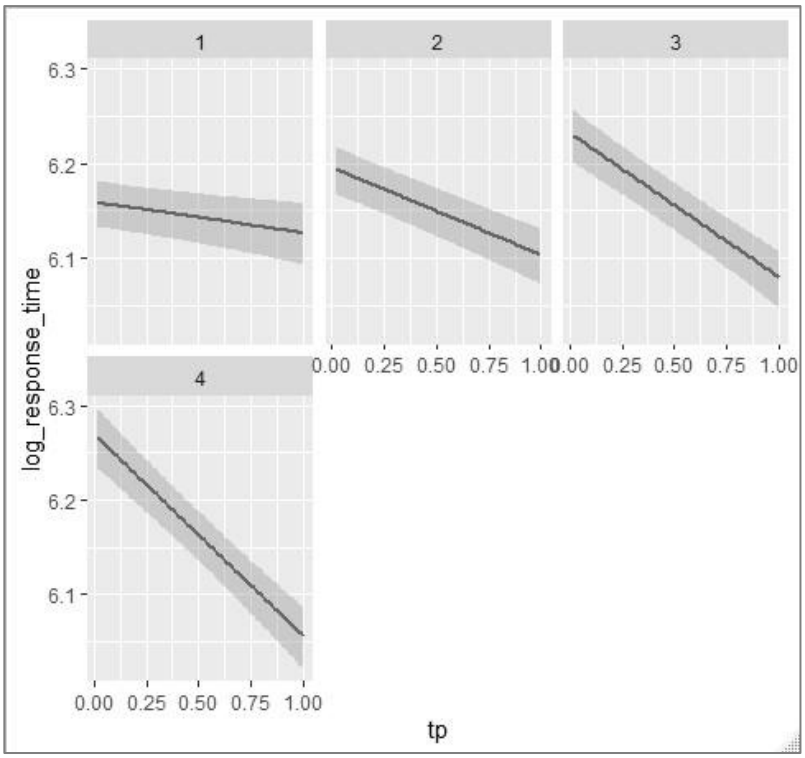

Figure 8.22: Lattice plot showing the effect of transitional probability on transition time; each plot depicts a separate level of transitional frequency.

8.11 **DISCUSSION**

The current experiment aimed to investigate the relative contributions of both transitional probability and transitional frequency to performance on a more complex sequence learning task. Data shows that the task was effective in eliciting learning from participants and that this learning was, in part, driven by both transitional probability and transitional frequency. Multi-level modelling using brms showed that transition time is primarily driven by an interaction between both transitional frequency and probability, with higher values on both variables required for better task performance.

This in no way invalidates the existing evidence for transitional probability as a robust predictor of statistical learning performance but does introduce a cautionary note: high transitional probability may only be effective with sufficient levels of exposure. Conversely, there seems to be no benefit of repeatedly exposing participants to low probability transitions, suggesting that there is a discrete point at which transitional probability becomes a useful marker of statistical regularity. This would explain why transitional probability has been shown to be a good predictor of performance in statistical learning paradigms that make use of smaller stimulus-sets - since the transitional probabilities are necessarily inflated – but not in the lexical decision experiments presented earlier in this work which use more naturalistic examples.

8.12 CHAPTER 7: REVISITED

The observation of an interaction between transitional frequency and transitional probability in this experiment led to my revisiting the meta-analysis presented in the previous chapter. Until now, there was no theoretical consideration that such an interaction could exist given that both transitional probability and transitional frequency are measuring similar things, in different ways. However, given that an interaction was observed in the sequence learning task it is possible that one may also be present in the lexical decision data. As such, I reanalysed the data from Chapter 7 with the inclusion of an interaction term.

```
dfm <- read_csv("Meta_data_all.csv")

dfm$log_bigram_freq <- log(dfm$bigram_freq + .000001)

dfm$log_trans_prob <- log(dfm$trans_prob + .000001)

dfm$log_diversity <- log(dfm$diversity + .000001)

dfm$log_response_time <- log(dfm$response_time + .000001)
```

```
Meta_interaction_model_1 <- brm(log_response_time ~ age + concreteness +
    letters + word_freq + trans_prob * bigram_freq + (1|subject) +
    (1|item), data = dfm, save_all_pars = TRUE)
Meta_interaction_model_2 <- brm(log_response_time ~ age + concreteness +
    Letters + word_freq + trans_prob * bigram_freq + (1|subject) +
    (1|item), data = dfm, save_all_pars = TRUE, chains = 3, iter = 5000)
Meta_interaction_model_3 <- brm(log_response_time ~ age + concreteness +
    letters + word_freq + trans_prob * bigram_freq + (1|subject) +
    (1|item), data = dfm, save_all_pars = TRUE, chains = 3, iter = 10000,
    warmup = 1000)
```

However, the interaction model failed to converge after 2000 iterations (model_1) and again after 5000 iterations (model_2). A final attempt at model convergence was made with 10,000 iterations (model_3); for this model, max_treedepth and adapt_delta were set as 15 and .9, respectively (defaults are set as 10 and .8 in brms). This has the effect of increasing the efficiency of the of the Monte Carlo process and reducing the size of each 'step' in the sampling chain to reduce the number of divergent transitions. However, even with these adjustments, the model still failed to converge forcing me to conclude that the interaction model is a poor fit for the meta-analysis data.

8.13 GENERAL DISCUSSION

The two experiments presented in this chapter attempted to replicate the pattern of findings reported in the earlier lexical decision tasks. Experiment 9 tasked participants with tapping on a target as it appeared in one of eight locations on a screen. Participants were then assessed on their ability to learn an underlying

sequence that they had not been made aware of. Their efficiency at learning this sequence was operationalised as the time taken to transition from one target to the next, with faster times demonstrating better learning of the transition. In Experiment 10, participants were given the same task but with sixteen rather than eight targets and a longer underlying sequence to learn. Transition times were then examined using multi-level modelling and key comparisons were made between pre-selected target pairs.

The data from Experiment 9 mirror the findings of the lexical decision experiments in suggesting that transitional frequency is the main predictor of participants' performance. Comparisons of pre-selected transitions support this conclusion and show that high frequency transitions elicit faster transition times. This implies that, when transitional frequency is high, increased transitional probability has little effect on the ability to learn the underlying sequence in the experiment. Furthermore, higher transitional probabilities seem to be having an adverse effect on performance when transitional frequency is low.

I then expanded the scope of the sequence learning task in Experiment 10 by increasing the number of potential targets as well as the length and complexity of the sequence. This resulted in a wider range of frequencies and a more comprehensive spread of transitional probabilities than in Experiment 9 and allowed me to examine whether the effects generalised to a more complex stimulus-set. The findings from Experiment 10 support those seen in Experiment 9 in showing that higher transitional frequencies are conducive to faster transition times among participants. However, the data also show an interaction between the two metrics where the effects of transitional probability are mediated by transitional frequency. When transitional frequency is low, transitional

probability is of little benefit in helping participants learn the underlying sequence; but, as transitional frequency increases the impact of transitional probability becomes more pronounced, having the most effect when frequency is at its highest.

This suggests that, contrary to my previous assertions, transitional frequency may not be a replacement metric for transitional probability but a complementary statistical measure and that participants may be attuned to both. This is evinced by the fact that, in both experiments, high transitional frequencies were only effective in reducing transition times in cases where transitional probability was also high – either 1 or .8, respectively. When transitional probability is low – at .33 in Experiment 9 and .25 in Experiment 10 – increasing the number of presentations has no benefit for participants. Additionally, increasing transitional probability only showed a meaningful effect when transitional frequency was high – held constant at four in Experiment 10 – and had no effect when help constant at either three (Experiment 9) or one (both experiments). This is shown most prominently in figure 8.22, which gives a clear summary of the interaction between the two metrics.

These findings are congruent with those of Evans et al. (2009) in which participants were presented with either twenty-one or forty-two minutes of an artificial grammar. In their study, participants who received the longer exposure were able to learn the statistical structure of the grammar where those who received less exposure were not. It was previously my assertion that, since the transitional probability remained constant across the two conditions, the increase in performance must be due to the increase in frequency. The current results support this assertion with the caveat that increased transitional frequency is only effective in conjunction with high transitional probability. This further suggests that participants may not be calculating the

probability of all transitions and are focused on learning the most predictable examples and using these to scaffold the parsing/acquisition of the grammar. After all, once you identify the inter-item transitions by their high probability, you can disregard the lower probability transitions when building a representation of the stimulus.

However, these results must once again be interpreted with caution. The experiments presented in this chapter utilise a relatively small stimulus-set which invariably leads to inflated transitional probabilities – the lowest possible being .125 in Experiment 9 and .006 in Experiment 10. Those these are lower than many of those reported in other statistical learning paradigms they are still much higher than those seen in more naturalistic datasets. In the next chapter, I will consider the results of all ten experiments in relation to one another and to the arguments presented in Chapter 2.

CHAPTER SUMMARY

In this chapter I:

- Conducted two sequence learning experiments to examine participants' ability to utilise statistical patterns to ascertain an unfamiliar sequence.
- Used cross-validation and Bayes factor comparisons to determine the most effective models at predicting the observed data.
- Directly compared pre-selected trials within each experiment to identify whether transitional frequency or transitional probability result in better sequence learning.
- Identified a potential interaction between transitional frequency and transitional probability.

9 DISCUSSION

CHAPTER OVERVIEW

In this chapter I will:

- Review the arguments for transitional probability, transitional (bigram) frequency, and (bigram) diversity.
- Provide a summary of the findings presented in each of the experimental chapters.
- Integrate the findings into a general discussion of the effects of distributional statistics on task performance.
- Discuss the implications of this work to statistical learning theory and, more specifically, statistical learning paradigms.

9.1 METRICS OF STATISTICAL LEARNING: OVERVIEW

9.1.1 Transitional probability

There is a wealth of evidence suggesting that transitional probabilities are the driving metric in statistical learning. Saffran, Aslin, and Newport (1996) showed that eight-month-old infants were capable of parsing streams of sound into nonsense words in the absence of explicit cues. Several studies have since used transitional probabilities as the measure by which they predict learning in a number of experimental paradigms. These have included both linguistic and non-linguistic stimuli with both adults and children (e.g. Aslin et al., 1998; Conway & Christianson, 2005; Daikoku et al., 2014; Frank et al., 2010; Goodman et al., 2008; Hay et al., 2011; Johnson & Tyler, 2010; Kirkham et al., 2002; Koelsh et al., 2016; Liu & Kager, 2011; Newport & Aslin, 2004; Reeder et al., 2017; Saffran, Johnson et al., 1999; Saffran, Newport, & Aslin, 1996; Saffran, Newport, Aslin, Tunick, et al., 1997; Theakston et al., 2004; Thiessen & Erickson, 2013; Thompson & Newport, 2007; Toro et al., 2005; Vouloumanos, 2008). This suggests that transitional probability is a robust indicator of statistical learning performance and has led to its acceptance as the primary metric of interest in statistical theory. In part, this is since transitional probabilities are claimed to protect the learner against the possibility of under-segmentation. They do this by adjusting the raw frequency of co-occurrence to account for the entire range of possible cooccurrence items. As such, raw co-occurrence frequency has been largely dismissed as a measure of statistical regularity in favour of transitional probability.

However, not every finding can be attributed exclusively to transitional probability. For example, in Saffran, Newport, and Aslin (1997) participants were exposed to either twenty-one or forty-two minutes of an artificial language and tested on their ability to discriminate between items from that language and novel items comprising the same phonological information arranged according to a different statistical pattern. It was found that the longer exposure time resulted in better discrimination performance, a fact that was attributed to participants having more time to encode the transitional probabilities. This effect could also be an artefact of increased frequency since only the number of presentations, not the transitional probabilities themselves, differ across the two conditions. This begs the questions as to whether more of the findings attributed to transitional probability can be explained by transitional frequency.

9.1.2 **Transitional (Bigram) frequency**

Erickson and Thiessen (2015) argue that a frequency-based mechanism is more plausible than transitional probability. In fact, computational modelling using PARSER (Perruchet & Vinter, 1998) and MOSAIC (Freudenthal et al., 2015) has demonstrated that a frequency-based system can accurately model children's speech errors.

More recently however, there have been calls to re-evaluate the prominence placed on transitional probability and consider alternative measures of statistical distribution (Slone & Johnson, 2018). Given that frequency effects are ubiquitous in studies of language acquisition (Ambridge et al., 2014) it is surprising that more has not been done to investigate raw co-occurrence frequency in statistical learning. In Chapter 2, I proposed bigram frequency as one such alternative metric. The primary argument for this is one of simplicity, which I will not repeat here except to say that a frequency-

based mechanism requires fewer cognitive resources to maintain than a probabilistic one.

It is recognised that frequency has a well-recorded effect across a number of domains including memory (Balota & Neely, 1980; MacLeod & Kampe, 1996; Hulme et al., 1997; Stretch & Wixted, 1998), reading (Dahan et al., 2001; Gerhand, & Barry, 1998; Inhoff & Rayner, 1986; Raynor & Duffy, 1986), sentence comprehension and production (Arnon & Snider, 2010; Diessel, 2007), and lexical decision performance (Grainger, 1990; Perea & Carreiras, 1998; Schilling, Rayner, & Chumbley, 1998).
There are also several experiential models of learning (Bybee, 1998; Rumelhart et al., 1986; Tomasello, 2000) which would predict stronger representations for more frequent associations. Indeed, if we consider the neural architecture required to facilitate such learning then it is not implausible to imagine discrete lexical representations with differentially weighted connections developed through their frequency of co-occurrence. The same cannot be said for a purely probabilistic representation which would require the entire experiential history to be maintained to enable online calculations of transitional probability.

9.1.3 **Bigram Diversity**

There is strong evidence to suggest that predictability is an important facet of language processing (Bates & MacWhinney, 1987; Conway et al., 2010; Glenberg & Gallese, 2012; Goldberg et al., 2005; Pickering & Garrod, 2004; 2007; Van Berkum et al., 2005). Transitional probability incorporates this predictability in a way not captured by raw co-occurrence frequency. Therefore, bigram diversity was suggested as a way of retaining the benefit of predictability without the need for the complex

online calculations associated with transitional probability. Bigram diversity is similar in concept to contextual diversity as proposed by Adelman et al. (2006).

However, the nature of bigram diversity is compatible with either a predictability effect – with lower diversity items being more predictable – and a diversity effect as predicted by Adelman et al.'s contextual diversity which would suggest that more diverse items confer a benefit on task performance. As such, the exact nature of any diversity effect was left open to exploration.

9.2 SUMMARY OF EXPERIMENTAL FINDINGS

Over the course of six experimental chapters, I presented ten experiments and a meta-analysis aimed at understanding the contributions of each of the three metrics to task performance. For each experiment, the data was analysed using Bayesian multi-level modelling and a model comparison approach was adopted using both cross-validation and Bayes factors as comparative measures. In addition, Chapter 8 also included direct comparisons of pre-selected items using Bayesian equivalence testing. The findings from each of the experimental chapters are summarised below and discussed in more detail later in this chapter.

9.2.1 Chapter 3

In Chapter 3, I presented two experiments designed to ascertain whether a lexical decision task could be used to assess whether participants were sensitive to the underlying statistical distributions present in a naturalistic stimulus-set. Faced with the prohibitive familiarisation times necessary to train participants on an artificial grammar of sufficient complexity to simulate natural language use it was decided that

an alternative approach was needed. The experimental stimuli were therefore extracted from the British National Corpus and the statistical associations between stimuli calculated. This allowed me to see if participants' task performance could be better predicted by transitional probability, frequency, or diversity. It follows that, if participants performed better for high frequency bigrams then this may be down to stronger lexical representations for those items; suggesting better encoding during the learning process.

The two experiments contrasted the effects of transitional probability with bigram frequency and bigram diversity, respectively, and were designed to provide participants with the maximal opportunity to benefit from the inherent associations between words. This was done in order to test the sensitivity of the task to the effect of the aforementioned metrics given that the paradigm has been hitherto unexplored in the context of statistical learning. The experiments were somewhat effective at detecting the effects of transitional probability and bigram frequency but showed no effect of bigram diversity; this may have been an artefact of stimuli selection which led to the inclusion of some unfamiliar items in Experiment 2.

The findings from these experiments demonstrate that learners may be sensitive to the statistical association between items in naturalistic stimulus-sets despite these being several magnitudes smaller than those traditionally seen in statistical learning research. They also show that bigram frequency has a beneficial effect on task performance in these tasks and that this supersedes any benefit of transitional probability. Furthermore, in the bigram diversity experiment, neither this nor transitional probability had any effect on task performance, suggesting no role of predictability in determining task performance.

9.2.2 **Chapter 4**

In Chapter 4, I attempted to address some methodological issues that arose in the first two experiments; the large amount of group-level variance in both experiments and the use of unusual stimuli in Experiment 2. Both experiments were therefore redone using slightly modified stimuli.

To reduce the group-level variance attributed to differences in the target items, I conducted a replication of Experiment 1 in which the target words were repeated at each level of bigram frequency (high, low, and zero). This was successful in reducing, but not eliminating, the inter-item variance and provided an unexpected result. The data showed that rather than the predicted benefit, increased transitional probability proved to be detrimental to participants' task performance. However, bigram frequency was not included in the model with the best fit to the data.

Experiment 4 was a direct replication of Experiment 2 with a slightly modified stimulus-list. The data from the repeated experiment suggest that transitional probability, when contrasted with bigram diversity, has the expected benefit of improving task performance.

The experiments presented in this and the previous chapter demonstrated to my satisfaction that the primed lexical decision task was sensitive enough to the statistical associations in the British National Corpus that conclusions could be reasonably drawn as to the effects of the different distributional statistics. However, given that these experiments were specifically designed to maximise participants' likelihood of responding to the statistical priming effect, I decided that it would be necessary to

replicate the effects using timings more typical of those found in word recognition research.

9.2.3 **Chapter 5**

Using timings more typical of previous lexical decision research, I attempted to replicate the findings presented in Chapter 3. Through these replications I showed that bigram frequency may represent a better metric of statistical learning than transitional probability, in the context of a primed lexical decision task. Furthermore, I suggested that the effects of transitional probability reported in previous research may be masking a frequency effect since, in this research, there appears to be little value in the predictability component that sets transitional probability apart from bigram frequency. Moreover, the consistent lack of effect for bigram diversity led to the conclusion that it may not be a useful metric of statistical distribution.

9.2.4 **Chapter 6**

Continuing the replications started in Chapter 5, I conducted two primed lexical decision tasks in which the target words were held constant across different levels of bigram frequency and bigram diversity (Experiments 7 & 8, respectively).

The data once again suggest that bigram frequency is a better predictor of task performance than transitional probability in these tasks. Furthermore, the data from Experiment 8 suggest that neither transitional probability nor bigram diversity influence task performance.

Taken together, these results provide more evidence that transitional probability may not be the best measure of statistical regularity as suggested by the majority of previous research.

9.2.5 **Chapter 7**

In this chapter, I performed a meta-analysis on the data from the previous eight experiments. This allowed me to draw conclusions from a larger dataset than would otherwise be possible and to incorporate bigram diversity into the models with transitional probability and bigram frequency, which was not done in previous chapters.

Following model selection using cross-validation and Bayes factors, I determined that the combined bigram frequency and transitional probability model provided the best fit to the data. This model supported the conclusions from previous chapters that bigram frequency has a beneficial influence on word recognition time in these tasks and that transitional probability has a negative effect. Bigram diversity once again showed no effect and so was dropped from remaining experiments.

However, these conclusions are based on a paradigm not previously seen in statistical learning research. As such, I decided that a more traditional approach was necessary to test these predictions.

9.2.6 **Chapter 8**

In Chapter 8, I set out to test whether the effects of transitional (bigram) frequency and transitional probability identified in the lexical decision tasks were maintained in a more traditional paradigm. I therefore used a sequence learning task to assess whether participants could become attuned to the statistical regularities of a simple pattern.

Model comparison revealed that the transitional frequency model provided the best description of the observed data and that as transitions become more frequent participants complete them more quickly. Moreover, comparisons between pre-selected transitional pairs confirmed that high frequency transitions were completed faster than those of a lower frequency. However, this increase in task performance was only meaningful in transitions that had a high transitional probability.

I then conducted a further sequence learning experiment in which I doubled the number of targets and increased the length – and therefore the complexity – of the sequence from thirty-seven to one-hundred-and-thirteen items. This resulted in a distribution of transitional probabilities similar to those seen in the British National Corpus and allowed me to test whether participants could become attuned to a more complex sequence when the only cues available were the statistical distribution of the transitions.

After selecting for the best model and conducting the pre-planned comparisons it became apparent that participant performance was being driven by an interaction between transitional frequency and transitional probability in such a way that, as transitional frequency increases the effects of transitional probability become more pronounced. Considering the interaction in the sequence learning task, I then revisited the meta-analysis data but failed to find an interaction between the two metrics in predicting lexical decision performance.

9.3 **DISCUSSION**

In this section I will draw conclusions as to the efficacy of each of the distributional statistics investigated in this work in relation to their ability to predict task performance in both the lexical decision and sequence learning tasks before discussing

some of the implications for statistical learning research in natural language. In the interest of generalisability, the term transitional frequency will be used to include bigram frequency from this point onwards; furthermore, I will be using response time to refer to both the time to complete a trial in both the lexical decision and sequence learning tasks.

Over the course of this work transitional frequency was shown to be an effective predictor of task performance in seven of the eight analyses – four lexical decision tasks, two sequence learning tasks, and the meta-analysis. In each of these cases a small but meaningful negative relationship was observed between transitional frequency and response time. This shows that transitional frequency is a reliable measure of statistical distribution in predicting both the acquisition of new information and the ability to draw on existing statistical associations to aid in a novel language task.

These findings are congruent with the claims that a frequency-based mechanism of statistical learning may be more psychologically plausible than a probabilistic one since it is more flexible in switching between linguistic units (Erickson & Thiessen, 2015) or, as demonstrated in Chapter 8, different domains. It is unsurprising that we see the benefits of frequency in these tasks given the prevalence of frequency-based effects in language tasks more generally (Ambridge et al., 2014) and these findings add to the small but important body of research investigating these effects in statistical learning specifically (e.g., Oganian et al., 2015; Schuler et al., 2017).

The overwhelming evidence (in this work) in favour of transitional frequency can be attributed to the relative simplicity involved in calculating and maintaining a frequency-based representation of the stimulus-set compared with the difficulty in maintaining and

updating a complete probabilistic representation – something that becomes more important as the set approaches naturalistic levels of complexity.

Given these findings we can attempt to answer the questions set out in Chapter 2: (1) If a simpler mechanism can facilitate effective learning, what benefit arises from the use of a more complex one? and (2) do learners require an accurate probabilistic representation of the stimulus-set to learn its inherent properties?

In relation to (1), the meta-analysis of all eight datasets shows transitional probability to have a small but meaningful detrimental effect on response time. This suggests that there is little benefit in tracking the transitional probabilities of bigrams and that doing so may introduce interference when attempting to access the associations at a later date. Which leads me to conclude, in response to (2) that learners do not require an accurate probabilistic representation of the stimuli-set in order to acquire new information. In fact, data from the sequence learning experiments presented herein suggest that transitional probability alone is insufficient to promote effective learning. This is compatible with research from domains outside of statistical learning that suggests better problem-solving performance when participants are given information in the form of frequencies rather than probabilities (Kahneman et al., 1982; Hertwig & Gigenrenzer, 1999; McDowell et al., 2018; Tversky & Kahneman, 1973).

In addition, there are numerous experiential models of learning from areas including memory (Balota & Neely, 1980; Hulme et al., 1997; MacLeod & Kampe, 1996; Stretch & Wixted, 1998), reading (Dahan et al., 2001; Gerhand & Barry, 1998; Inhoff & Rayner, 1986; Raynor & Duffy, 1986), and sentence comprehension and production (Arnon & Snider, 2010; Diessel, 2007) that suggest better performance for higher

frequency items. The same pattern of effects can also be seen in word recognition and naming studies (Grainger, 1990; Perea & Carreiras, 1998; Schilling et al., 1998) where individual word frequency is considered to be a major predictor of task performance.

It has also been demonstrated that children with atypical language development require more exposure – and therefore higher frequency – in order to learn the statistical properties of the stimulus-set and that typically developing learners also perform better under these conditions (Evans et al., 2009). Since transitional probability is constant across different levels of exposure, it follows that frequency is the driving force behind these improvements. Otherwise we would not expect to see any effect for typically developing participants who were able to acquire the regularities at the shorter exposure times given that, once you know that A precedes B in 100% of cases, there is little benefit in repeated presentations. The issue becomes less clear when considering the observed interaction between transitional frequency and transitional probability in Experiment 10; this implies that there is at least some benefit of transitional probability beyond that provided by transitional frequency – though only for higher frequency items – in the acquisition of new information. I am reticent, however, to draw more than tentative conclusions from this finding given that the interaction is not apparent in the lexical decision data nor in the other sequence learning task.

In fact, the overall lack of a consistent effect for transitional probability is perhaps the most surprising outcome of the data presented herein. I have repeatedly highlighted the prevalence of transitional probability as the preferred metric in the statistical learning literature and multitudinous studies have demonstrated its relationship with performance on a variety of different tasks. Why then do we see no effect in the current studies?

Could it be that transitional probability is masking a frequency-based effect in many of these tasks? I have already detailed studies from Saffran, Aslin, and Newport (1997) and Evans et al. (2009) where improvements in learning performance after greater periods of exposure have been attributed to increased opportunity to become attuned to the transitional probabilities of the stimulus-set. It is not too difficult to imagine that the effect may have more to do with increased frequency than better attunement. Similarly, Koelsh et al. (2016) describe exposing participants to low, intermediate, and high probability events in which the third item of a triplet varied as the first two remained constant. The rate of occurrence in each trial was ten percent for the low probability trials, thirty percent for the intermediate, and sixty percent for the high probability trials. On examination, we can see that this arrangement of stimuli means that the high probability trials occur six times more frequently than the low probability ones meaning that any effect of probability could also be attributed to frequency.

Simply put, in tasks where transitional probability is the chosen metric of statistical distribution, it is necessary to disambiguate any effect of frequency if reliable conclusions are to be drawn. The current work accounts for the effects of both transitional probability and frequency and finds that, when transitional frequency is included in statistical models of task performance, transitional probability no longer elicits the predicted effects. This lends further credence to the suggestion that transitional probability may be masking a frequency-based mechanism of learning. If this is the case, then we must consider whether the predictability component of transitional probability is providing any benefit beyond that obtained from raw cooccurrence frequency.

In the current work, the predictability of a stimulus-set was also captured by bigram diversity. However, in a suite of lexical decision tasks and a meta-analysis bigram diversity was shown to be a poor predictor of task performance – returning null results on each occasion. We also see that the only experiments where transitional probability is shown to have a beneficial effect on response time are those where it is contrasted with bigram diversity and not transitional frequency. This suggests that, when frequency is not accounted for in the model, transitional probability is reflecting the frequency effect rather than one of predictability. If transitional probability were truly capturing predictability, then we would expect to see a benefit alongside (or instead of) that of frequency in the lexical decision experiments. In the absence of any evident effect, we must conclude that the predictability component of transitional probability does nothing to aid task performance beyond that which can be explained by a frequency-based mechanism.

The current study used an innovative approach to investigate a commonly accepted phenomenon - that humans are capable of tracking distributional information within the environment. A lexical decision task was used to assess previously learnt associations. This allowed for an examination of naturalistic distributions without engaging in a lengthy familiarisation process with participants. This procedure highlighted several things. Firstly, individuals are capable of accessing previously learnt statistical relationships and making predictions based on these prior associations. Furthermore, this demonstrates the persistent nature of these associations, some of which may not have been encountered for extended periods prior to testing or may only be encountered extremely infrequently. Crucially, the current study also demonstrates the applicability of statistical learning theories to

large, complex stimulus-sets. The associations presented herein were extracted from the BNC and constitute a far richer example of language than would have been possible with artificial grammars.

The current work demonstrates that individuals can use transitional frequency to respond to statistical primes in a lexical decision task and that this constitutes a better predictor of task performance than transitional probability - at least when accessing previously learnt associations. It is suggested that, although transitional probabilities provide a more complete distributional representation of the stimulus-set, the benefit gained from such a representation does not justify the additional computational costs. This provides a measure of support for the psychological plausibility of a frequency-based mechanism of learning, as suggested by Erikson and Thiessen (2015).

Conclusions

- Transitional frequency represents a better metric of statistical distribution for predicting task performance in a primed lexical decision task
- Bigram diversity does not constitute an effective measure of statistical distribution
- Transitional probability may not be as effective a predictor of task performance as previously suggested but may have some value in the acquisition of new information
- Future studies should attempt to disambiguate the effects of transitional probability and frequency
- Statistical learning can be applied to naturalistic datasets, but caution is advised when attempting to generalise from artificial grammars, particularly when referencing the effects of transitional probability

Additional package citations

A number of R packages were integral to the production of this thesis but were uncited in the main text due to stylistic or ease-of-reading concerns; the following packages were used but not cited: readr (Wickham, Hester, & Francois, 2017); formatR (Xie, 2017); Rcpp (Eddelbuettel & Francois, 2011; 2013; 2017); ggplot2 (Wickham, 2016); flextable (Gohel, 2019a); officer (2019b).

References

Adelman, J. S. (2011). Letters in time and retinotopic space. *Psychological Review, 118,* 580582.

Adelman, J. S., & Brown, G. D. (2008). Modelling lexical decision: The form of frequency and diversity effects. *Psychological Review, 115*, 214.

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17,* 814-823.

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of child language, 42,* 239-273.

Anderson, J. L., Morgan, J. L., & White, K. S. (2003). A statistical basis for speech sound discrimination. *Language and Speech, 46*, 155-182.

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

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321-324.

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of Word Segmentation in Fluent Maternal Speech to Infants. *In Signal to Syntax* (pp. 117–134).

Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology, 46*, 81–86.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133,* 283-316.

Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental psychology: Human Learning and Memory, 6*, 576.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. *Mechanisms of Language Acquisition*, 157-193.

Bernardo, J. M., & Smith, A. F. (2009). Bayesian Theory (Vol. 405). John Wiley & Sons.

Bishop, K., Symons, N., Fry, S., Laurie, H., 2 Entertain (Firm), British Broadcasting Corporation., BBC Worldwide Ltd., ... Warner Home Video (Firm). (2007). A bit of Fry & Laurie: Season three. England: BBC Video

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18,* 355-387.

Bock, K., & Loebell, H. (1990). Framing sentences. *Cognition, 35*, 1-39.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science, 27*, 45-50.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977-990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904-911.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80,* 1-28.

Bybee, J. (1998). The emergent lexicon. *In Chicago Linguistic Society* (Vol. 34, No. 2, pp. 42135).

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., …, Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*

Carr, D., ported by Lewin-Koh, N., Maechler, M., and contains copies of lattice functions written by Sarkar, D. (2019). hexbin: Hexagonal binning routines. R Package version 1.27.3. https://CRAN.R-project.org/package=hexbin

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition, 87,* B69-B77.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2010). A Vowel Is a Vowel: Generalizing Newly Learned Phonotactic Constraints to New Contexts. *Journal of Experimental Psychology: Learning Memory and Cognition, 36*, 821.

Christiansen, M. H., & Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in Cognitive Sciences, 7*, 300-307.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335-359.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. *Perception and Production of Fluent Speech,* 133-163.

Coltheart, M. (1978). Lexical access in simple reading tasks. *Strategies of Information Processing,* 151-216.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100,* 589–608.

Coltheart, M., & Leahy, J. (1996). Assessment of lexical and nonlexical reading abilities in children: Some normative data. *Australian Journal of Psychology, 48,* 136–140.

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

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 24.

Conway, C. M., Pisoni, D. B., Anaya, E. M., Karpicke, J., & Henning, S. C. (2011). Implicit sequence learning in deaf children with cochlear implants. *Developmental Science, 14,* 69-82.

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1119.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42,* 317367.

Daikoku, T., Yatomi, Y., & Yumoto, M. (2015). Statistical learning of music-and language-like sequences and tolerance for spectral shifts. *Neurobiology of Learning and Memory, 118*, 8-19.

de Groot, A. M. (1989). Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 824.

de Groot, A. M., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. Language learning, 50, 1-56.

Dell, G.S., Reed, K.D., Adams, D.R., & Meyer, A.S. (2000). Speech errors, phonotactic constraints and implicit learning: A study of the role of experience in language

production. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 1355-1367.

Diependaele, K., Brysbaert, M., & Neri, P. (2012). How noisy is lexical decision? Frontiers in Psychology, 3, 348.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. Frontiers in psychology, 5, 781.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology, 25*, 108-127.

Dijkstra, T., Hilberink-Schulpen, B., & Van Heuven, W. J. (2010). Repetition and masked form priming within and between languages using word and nonword neighbors. *Bilingualism: Language and Cognition, 13,* 341-357.

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory, 6*, 125-155.

Durrant, S. J., Taylor, C., Cairney, S., & Lewis, P. A. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia, 49*, 1322-1331.

Eddelbuettel, D., & Francois, R. (2011). Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software, 40*, 1-18.

Eddelbuettel, D. (2013) Seamless R and C++ Integration with Rcpp. Springer, New York. ISBN 978-1-4614-6867-7.

Eddelbuettel, D., & Balamuta, J.J. (2017). Extending R with C++: A Brief Introduction to Rcpp. PeerJ Preprints 5:e3188v1. https://doi.org/10.7287/peerj.preprints.3188v1.

Ellis, N. C. (2015). Implicit and explicit learning: Their dynamic interface and complexity. In P. Rebuschat (Ed.), *Implicit and Explicit Learning of Languages* (pp. 3-23). Amsterdam: John Benjamins.

Ellis, N. C., Frey, E., & Jalkanen, I. (2008). The Psycholinguistic Reality of Collocation and Semantic Prosody (1): Lexical Access. In U. Römer & R. Schulze (Eds.), *Exploring the Lexis-Grammar Interface*. Amsterdam: John Benjamins.

Ellis, N. C., & Hooper, A. M. (2001). Why learning to read is easier in Welsh than in English: Orthographic transparency effects evinced with frequency-matched tests. *Applied Psycholinguistics, 22,* 571-599.

Endress, A. D., & Langus, A. (2017). Transitional probabilities count more than frequency, but might not be used for memorization. *Cognitive psychology, 92*, 37-64.

Endress, A. D., & Mehler, J. (2009a). Primitive computations in speech processing. *The Quarterly Journal of Experimental Psychology, 62,* 2187-2209.

Endress, A. D., & Mehler, J. (2009b). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language, 60,* 351-367.

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review, 37*, 66-108.

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 52,* 321-335.

Fernald, A. (1985). Four-Month-Old Infants Prefer to Listen to Motherese. *Infant Behavior and Development,* 181-195.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Meot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42,* 488-496.

Ferré, P., Guasch, M., García-Chico, T., & Sánchez-Casas, R. (2015). Are there qualitative differences in the representation of abstract and concrete words? Within-language and cross-language evidence from the semantic priming paradigm. *The Quarterly Journal of Experimental Psychology, 68,* 2402-2418.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 458.

Fiser, J. (2009). Perceptual learning and representational learning in humans and animals. *Learning & Behavior, 37,* 141-153.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition, 117*, 107-125.

Freudenthal, D., Pine, J. M., Jones, G., & Gobet, F. (2015). Simulating the cross-linguistic pattern of Optional Infinitive errors in children's declaratives and Wh-questions. *Cognition, 143,* 61-76.

Frost, R. L., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of nonadjacent dependencies from continuous speech. *Cognition, 147*, 70-74.

Gebhart, A. L., Newport, E. L., & Aslin, R. N. (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic Bulletin & Review, 16,* 486-490.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15,* 15931623.

Gerhand, S., & Barry, C. (1998). Word frequency effects in oral reading are not merely age of-acquisition effects in disguise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 267.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32*, 249 – 268.

Glenberg, A. M., & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex, 48,* 905-922.

Gohel, D. (2019a). flextable: Functions for tabular reporting. R package version 0.5.4. https://CRAN.R-project.org/package=flextable

Gohel, D. (2019b). officer: Manipulation of Microsoft Word and PowerPoint documents. R package 0.304. https://CRAN.R-project.org/package=officer

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2005). The role of prediction in construction-learning. *Journal of Child Language, 32*, 407-426.

Goldrick, M. (2004). Phonological features and phonotactic constraints in speech production. *Journal of Memory and Language, 51,* 586-603.

Goldrick, M., & Larson, M. (2008). Phonotactic probability influences speech production. *Cognition, 107,* 1155-1164.

Gomez, R. L. (2002). Variability and detection of invariant structure*. Psychological Science, 13,* 431-436.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70,* 109-135.

Gómez, R. L., & Lakusta, L. (2004). A first step in form-based category abstraction by 12month-old infants. *Developmental Science, 7*, 567-580.

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy, 7,* 183-206.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language, 35,* 515-531.

Goswami, U., Gomert, J. E., & de Barrera, L. F. (1998). Children's orthographic representations and linguistic transparency: Nonsense word reading in English, French, and Spanish. *Applied Psycholinguistics, 19*, 19–52.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming*. Journal of memory and language, 29*, 228-244.

Grunow, H., Spaulding, T. J., Gómez, R. L., & Plante, E. (2006). The effects of variation on learning word order rules by adults with and without language-based learning disabilities. *Journal of Communication Disorders, 39,* 158-170.

Harris, M., Barrett, M., Jones, D., & Brookes, S. (1988). Linguistic input and early word meaning. Journal of Child Language, 15, 77-94.

Hasher, L., & Chromiak, W. (1977). The processing of frequency information: An automatic mechanism? *Journal of Verbal Learning and Verbal Behavior, 16*, 173–184.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition, 78,* B53-B64.

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology, 63,* 93-106.

Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12,* 275305.

Hoekstra, R., Monden, R., van Ravenzwaaij, D., & Wagenmakers, E. J. (2018). Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. *PloS one, 13*, e0195474.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15,* 13511381.

Höhle, B., Bijeljac-Babic, R., Herold, B., Weissenborn, J., & Nazzi, T. (2009). Language specific prosodic preferences during the first half year of life: Evidence from German and French infants. *Infant Behavior and Development, 32*, 262-274.

Houx, P. J., Jolles, J., & Vreeling, F. W. (1993). Stroop interference: aging effects assessed with the Stroop Color-Word Test. *Experimental Aging Research, 19*, 209-224.

Hsu, H. J., & Bishop, D. V. M. (2014). Sequence-specific procedural learning deficits in children with specific language impairment. *Developmental Science, 17,* 352-365.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D., Martin, S., & Stuart, G. (1997). Word frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1217.

Hurtado, N., Marchman, V. A., & Fernald, A. (2008). Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in Spanish-learning children. *Developmental Science, 11*, F31-F39.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., ... & Buchanan, E. (2013). The semantic priming project. Behavior Research Methods, 45, 1099-1114.

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

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2017). Testing statistical learning implicitly: a novel chunk-based measure of statistical learning. *In the*

39th Annual Conference of the Cognitive Science Society (CogSci 2017) (pp. 564-569). Cognitive Science Society.

Jain, A. K., Duin, R. P., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*, 4-37.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44,* 548-567.

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science, 13*, 339-345.

Jones, G. (2016). The influence of children's exposure to language from two to six years: the case of nonword repetition. *Cognition, 153*, 79-88.

Jones, G., & Macken, B. (2018). Long-term associative learning predicts verbal short-term memory performance. *Memory & Cognition, 46*, 216-229.

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive psychology, 98*, 1-21.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants′ detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29*, 1-23.

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (Eds.). (1982). Judgment under uncertainty: Heuristics and biases. Cambridge university press.

Kanske, P., & Kotz, S. A. (2007). Concreteness in emotional words: ERP evidence from a hemifield study*. Brain research, 1148*, 138-148.

Kapatsinski, V., & Radicke, J. (2008). Frequency and the emergence of prefabs: Evidence from monitoring. In R. Corrigan, E. A. Moravcsik, H. Ouali & K. M. Wheatley (Eds.), *Formulaic language. Volume 2: Acquisition, loss, psychological reality, functional explanation* (pp. 499-522). Amsterdam: John Benjamins.

Keuleers, E., Diependaele, K. & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology, 174*.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44*, 287-304.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one, 7*, e36399.

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: is it long-term and implicit? *Neuroscience letters, 461*, 145-149.

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

Koelsh, S., Busch, T., Jentschke, S., & Rohrmeier, M. (2016). Under the hood of statistical learning: A statistical MMN reflects the magnitude of transitional probabilities in auditory sequences. *Scientific Reports, 6,* 19741

Kooijman, V., Hagoort, P., & Cutler, A. (2009). Prosodic structure in early word segmentation: ERP evidence from Dutch ten-month-olds. *Infancy, 14*, 591-612.

Kruschke, J. K., & Meredith, M. (2018). BEST: Bayesian estimation supersedes the t-test. Rpackage Version 0.2, 2.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience, 5*, 831.

Kusunose, Y., Hino, Y., & Lupker, S. J. (2016). Masked semantic priming effects from the prime's orthographic neighbours. *Journal of Cognitive Psychology, 28,* 275-296.

Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German-English comparison. *Cognition, 63,* 315–334.

Leech, G., Rayson, P., & Wilson, A. (2001). Word frequencies in written and spoken English: based on the British National Corpus. London: Longman.

Lester, N., Feldman, L., & del Prado Martín, F. M. (2017). You can take a noun out of syntax...: Syntactic similarity effects in lexical priming. In *CogSci.*

Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition, 122*, 241-246.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology, 52*, 362-375.

Liu, L., & Kager, R. (2011). How do statistical learning and perceptual reorganization alter Dutch infant's perception to lexical tones? *In Proceedings of the 17th International Congress of Phonetic Sciences* (Vol. 17, pp. 1270-1273).

Lüdecke, D. (2018a). _sjPlot: Data Visualization for Statistics in Social Science_. R package version 2.5.0, https://CRAN.R-project.org/package=sjPlot

Lüdecke, D. (2018b). strengejacke: Load Packages Associated with Strenge Jacke!. R package version 0.3.0. https://github.com/strengejacke/strengejacke

MacLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of experimental psychology: Learning, memory, and cognition, 22,* 132.

Madan, C. R., Shafer, A. T., Chan, M., & Singhal, A. (2016). Shock and awe: Distinct effects of taboo words on lexical decision and free recall, *The Quarterly Journal of Experimental Psychology,* 1-18

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101-B111.

McCall, R. B., & Kagan, J. (1970). Individual differences in the infant's distribution of attention to stimulus discrepancy. *Developmental Psychology*, 2, 90

McDowell, M., Galesic, M., & Gigerenzer, G. (2018). Natural frequencies do foster public understanding of medical tests: Comment on Pighin, Gonzalez, Savadori, and Girotto (2016). *Medical Decision Making, 38,* 390-399.

Miller, G. A., & Selfridge, J. A. Verbal context and the recall of meaningful material. *The American Journal of Psychology, 63,* 176-185

Milne, A. E., Petkov, C. I., & Wilson, B. (2017). Auditory and visual sequence learning in humans and monkeys using an artificial grammar learning paradigm. *Neuroscience, 389,* 104-117.

Misyak, J. B., & Christiansen, M. H. (2012). Statistical Learning and Language: An Individual Differences Study. *Language Learning, 62,* 302-331

Monroy, C. D., Gerson, S. A., & Hunnius, S. (2017). Toddlers' action prediction: Statistical learning of continuous action sequences. *Journal of Experimental Child Psychology, 157,* 14-28.

Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 116.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111,* 721–756

Nemko, B. (1984). Context versus isolation: Another look at beginning readers. *Reading Research Quarterly*, 461-467.

New, B. (2006). Re-examining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13*, 45-52.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Re-examining word length effects in visual word recognition: New evidence from the English lexicon project. *Psychonomic Bulletin & Review, 13*, 45–52.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of nonadjacent dependencies. *Cognitive Psychology, 48,* 127-162.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature, 417*, 611.

Öney, B., & Durgunoğlu, A. Y. (1997). Beginning to read in Turkish: A phonologically transparent orthography. *Applied Psycholinguistics, 18,* 1–15.

Onishi, K. H., Chambers, K. E., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition, 83,* B13-23.

O'Regan, J. K., & Jacobs, A. M. (1992). Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception & Performance, 18,* 185-197

Oganian, Y., Conrad, M., Aryani, A., Spalek, K., & Heekeren, H. R. (2015). Activation patterns throughout the word processing network of l1-dominant bilinguals reflect language similarity and language decisions. *Journal of Cognitive Neuroscience, 27,* 2197-2214.

Ortells, J. J., Kiefer, M., Castillo, A., Megías, M., & Morillas, A. (2016). The semantic origin of unconscious priming: Behavioral and event-related potential evidence during category congruency priming from strongly and weakly related masked words. *Cognition, 146,* 143-157.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8month-old infants. *Child Development, 80,* 674-685.

Perea, M., & Carreiras, M. (1998). Effects of syllable frequency and syllable neighborhood frequency in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 24, 134.*

Perea, M., & Gomez, P. (2012). Subtle increases in inter-letter spacing facilitate the encoding of words during normal reading. *PLoS One, 7,* e47568.

Perea, M., Marcet, A., Vergara-Martínez, M., & Gomez, P. (2016). On the Dissociation of Word/Nonword Repetition Effects in Lexical Decision: An Evidence Accumulation Account. *Frontiers in Psychology, 7,* 215.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences, 10,* 233–238.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39,* 246-263.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27,* 169-190.

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

Pickering, M. J., & Ferreira, V. S. (2008). Structural Priming: A Critical Review. *Psychological Bulletin, 134,* 427.

Plante, E., Gomez, R., & Gerken, L. A. (2002). Sensitivity to word order cues by normal and language/learning disabled adults. *Journal of Communication Disorders, 35,* 453-462.

Quine, W. V. O., Word and Object [1960]. New edition, with a foreword by Patricia Churchland, Cambridge, Massachusetts: MIT Press, 2015.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Raftery, A. E. (1995). Bayes factors and BIC: Comment on "A critique of the Bayesian information criterion for model selection". *Sociological Methods & Research, 27,* 411427.

Raftery, A. E. (1998). Bayes factors and BIC: Comment on Weakliem. Technical Report 347, Department of Statistics, University of Washington.

Raftery, A. E., & Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association, 98,* 931-938.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology Section A, 55,* 1339-1362.

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

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2017). Distributional learning of subcategories in an artificial grammar: Category generalization and subcategory restrictions. *Journal of Memory and Language, 97*, 17-29.

Reines, M. F., & Prinz, J. (2009). Reviving Whorf: The Return of Linguistic Relativity. *Philosophy Compass.*, *4*, 1022-1032.

Richardson, J., Harris, L., Plante, E., & Gerken, L. (2006). Subcategory learning in normal and language learning-disabled adults: How much information do they need? *Journal of Speech, Language, and Hearing Research, 49, I1257-1266..*

Richardson, P., & Voss, J. F. (1960). Replication report: Verbal context and the recall of meaningful material. *Journal of Experimental Psychology, 60,* 417-418.

Riches, N.G., Tomasello, M., & Conti-Ramsden, G. (2005). Verb learning in children with SLI: Frequency and spacing effects. *Journal of Speech, Language, and Hearing Research, 48,* 1397-1411.

Romberg, A. R., & Saffran, J. R. (2013). Expectancy learning from probabilistic input by infants. *Frontiers in Psychology, 3*, 610.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review, 25,* 102-113.

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill*. Journal of child language, 35,* 185-205.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1,* 45-76.

Saffran, J. R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition, 81*, 149-169.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. Current directions in psychological science, 12(4), 110-114.

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

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*, 27-52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*, 606-621.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*, 101-105.

Saffran, J. R., Werker, J. F., & Werner, L. A. The Infant's Auditory World: Hearing, Speech, and the Beginnings of Language. Handbook of Child Psychology.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: multilevel statistical learning by 12-month-old infants. *Infancy, 4,* 273-284.

Schilling, H. E., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition, 26,* 1270-1281.

Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., … Larmarange, J. (2018). GGally: Extension to 'ggplot2'. R package version 1.4.0. https://CRAN.Rproject.org/package=GGally

Schuler, K. D., Reeder, P. A., Newport, E. L., & Aslin, R. N. (2017). The effect of Zipfian frequency variations on category formation in adult artificial language learning. *Language Learning and Development, 13,* 357-374.

Schur A.I., Tappert C.C. (2017) Speed and Accuracy Improvements in Visual Pattern Recognition Tasks by Employing Human Assistance. In: Nunes I. (eds) Advances in Human Factors and System Interactions. Advances in Intelligent Systems and Computing, vol 497. Springer, Cham

Schwartz, R. G., & Terrell, B. Y. (1983). The role of input frequency in lexical acquisition. *Journal of Child Language, 10,* 57-64.

Seidl, A., Cristià, A., Bernard, A., & Onishi, K. H. (2009). Allophonic and Phonemic Contrasts in Infants' Learning of Sound Patterns. *Language Learning and Development*, *5*, 191-202.

Shapiro, B. J. (1969). The subjective estimate of relative word frequency. *Journal of Verbal Learning and Verbal Behavior, 8*, 248–251.

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition, 177*, 198-213.

Slone, L. K., & Johnson, S. P. (2018). When learning goes beyond statistics: infants represent visual sequences in terms of chunks. *Cognition, 178*, 92-102.

Stan Development Team (2016). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.13.1. http://mc-stan.org/.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1379.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition, 108*, 850-855.

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2004). Semantic generality, input frequency and the acquisition of syntax. *Journal of Child Language, 31*, 61-99.

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium.

Thiessen, E. D., & Erickson, L. C. (2013). Discovering words in fluent speech: The contribution of two kinds of statistical information. *Frontiers in Psychology, 3,* 590

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental Psychology, 39*, 706.

Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics, 11,* 61-82.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development, 3,* 1-42.

Tomblin, J. B., Mainela-Arnold, E., & Zhang, X. (2007). Procedural Learning in Adolescents with and without Specific Language Impairment. *Language Learning and Development., 3,* 269-293.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97,* B25-B34.

Toro, J. M., & Trobalón, J. B. (2005). Statistical computations over a speech stream in a rodent*. Perception & Psychophysics, 67*, 867-875.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*, 552.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207-232.

Underwood, B. J. (1971). Recognition memory. In H. H. Kendler & J. T. Spence (Eds.), Essays in Neo-Behaviorism (pp. 313–335). New York: Appleton-Century-Crofts.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 443.

Van Heugten, M., & Johnson, E. K. (2010). Linking infants' distributional learning abilities to natural language acquisition. *Journal of Memory and Language, 63,* 197-209.

Van Heugten, M., & Shi, R. (2009). French-learning toddlers use gender information on determiners during word recognition. *Developmental Science, 12,* 419-425.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*, 1176-1190.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and computing, 27, 1413-1432.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition, 107*, 729-742.

Warker, J. A., Xu, Y., Dell, G. S., & Fisher, C. (2009). Speech errors reflect the phonotactic constraints in recently spoken syllables, but not in recently heard syllables. *Cognition, 112*, 81-96.

Whorf, B. L. (1956). Language, thought, and reality: selected writings. Technology Press books in the social sciences.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York

Wickham, H., Hester, J., & Francois, R. (2017). readr: Read Rectangular Text Data. R package version 1.1.1. https://CRAN.R-project.org/package=readr

Widdowson, H. G. (1989). Knowledge of language and ability for use. *Applied linguistics, 10*, 128-137.

Wingfield, A., Lindfield, K. C., & Goodglass, H. (2000). Effects of age and hearing sensitivity on the use of prosodic information in spoken word recognition. *Journal of Speech, Language, and Hearing Research, 43*, 915-925.

Xie, Y. (2017). formatR: Format R Code Automatically. R package version 1.5. https://CRAN.R-project.org/package=formatR

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory & Language, 60,* 502-529.

Yap, M. J., Balota, D. A., & Tan, S. E. (2013). Additive and interactive effects in semantic priming: Isolating lexical and decision processes in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 140-158.

Yap, M. J., Hutchison, K. A., & Tan, L. C. (2016). Individual differences in semantic priming performance: insights from the Semantic Priming Project. *Big Data in Cognitive Science: From Methods to Insights,* 203-226.

Zhang, Q., Guo, C. Y., Ding, J. H., & Wang, Z. Y. (2006). Concreteness effects in the processing of Chinese words. *Brain and Language, 96*, 59-68.

List of appendices:

1. Word list (Bigram frequency; Experiments 1, 3, 5, & 7)
2. Word list (Experiment 2)
3. Word list (Experiment 4, 6, & 8)

Appendix 1: Word list (Bigram frequency; Experiments 1, 3, 5, & 7)

Bigram frequency	Prime	Target	Target concreteness	Prime diversity	Target Letters	Transitional probability	Target frequency
0	berate	involved	2.03	0	8	0	19964
0	ski	aircraft	4.4	17	8	0	6203
0	chilli	call	4	5	4	0	19484
0	oval	hipster	2.5	9	7	0	19060
0	lifer	hugs	4.14	0	4	0	103
0	tides	mauve	4	4	5	0	222
0	reaches	timers	4.69	16	6	0	51
0	way	agree	2.31	263	5	0	8181
0	fart	course	3.82	0	6	0	19694
0	year	hundred	3.2	343	7	0	19109
0	recede	socks	4.91	0	5	0	991
0	gimlet	coding	3.03	0	6	0	494
0	snuffles	model	4.53	0	5	0	13335
0	meet	gone	2.04	112	4	0	19548
0	drubs	nudge	4.47	0	5	0	153
0	stippled	trade	3.08	0	5	0	19981
0	have	sihgt	0	0	5	0	0
0	faith	rink	4.56	49	4	0	141
0	trillion	droop	3.68	0	5	0	77
0	briskly	allow	2.41	8	5	0	11469
0	cycle	language	2.35	44	8	0	18778
0	rethinks	scaly	4.22	0	5	0	75
0	geese	wits	1.76	4	4	0	400
0	oaken	whose	1.68	0	5	0	19834
0	abase	number	3.3	0	6	0	49385
0	vexes	street	4.75	0	6	0	19614
0	systems	short	3.61	154	5	0	18652
0	gunboat	found	2.53	0	5	0	48923
0	building	food	4.8	177	4	0	18992
0	winds	agreed	1.93	27	6	0	14692
0	snuffles	moedl	0	0	5	0	0
0	bile	ptah	0	0	4	0	0
0	secret	whsoe	0	0	5	0	0
0	their	fodo	0	0	4	0	0
0	chilli	acll	0	0	4	0	0
0	eighth	atcion	0	0	6	0	0
0	recent	adrdess	0	0	7	0	0
0	gimlet	coidng	0	0	6	0	0
0	acid	adedd	0	0	5	0	0
0	chow	trdae	0	0	5	0	0
0	rugby	stgae	0	0	5	0	0
0	lifer	hgus	0	0	4	0	0

0	barbed	folor	0	0	5	0	0
0	quickly	allwo	0	0	5	0	0
0	darn	agered	0	0	6	0	0
0	fleet	adivce	0	0	6	0	0
0	way	agere	0	0	5	0	0
0	year	hunderd	0	0	7	0	0
0	ski	aicrraft	0	0	8	0	0
0	fart	ocurse	0	0	6	0	0
0	funny	adlut	0	0	5	0	0
10	time	across	3.07	569	6	6.45E-05	25203
10	people	achieve	2.29	679	7	8.05E-05	6768
10	many	active	3.32	975	6	0.000112	7290
10	before	actual	2.22	517	6	0.000113	6851
10	want	advice	2.73	139	6	0.000174	10437
10	really	able	2.38	505	4	0.000208	30410
10	local	access	2.71	568	6	0.000216	10940
10	always	accept	3.03	567	6	0.000216	9807
10	case	above	3.33	237	5	0.000231	25748
11	give	advance	2.57	248	7	0.000244	5040
10	less	afraid	2.7	479	6	0.00028	5967
10	large	adult	4.4	468	5	0.000296	5078
10	interest	account	3.08	149	7	0.000362	15891
10	quickly	added	2.74	129	5	0.000807	15375
10	secret	address	3.89	69	7	0.001735	7036
461	good	practice	2.52	850	8	0.005706	17114
10	hostile	action	2.86	19	6	0.006068	22099
2083	their	children	4.89	3442	8	0.007983	46608
4525	have	done	2	1900	4	0.009553	35473
10	enforced	absence	2.31	11	7	0.010091	5780
311	recent	times	2.07	203	5	0.019612	29910
5568	other	hand	4.72	1749	4	0.039173	35352
10	rustic	style	2.67	1	5	0.041152	10725
1769	second	half	3.27	498	4	0.042674	29782
198	funny	thing	3.17	62	5	0.044098	35211
381	fifty	five	3.87	78	4	0.044395	40739
4530	make	sure	1.73	376	4	0.056995	24595
13	glide	path	4.41	1	4	0.059091	6251
14	canned	food	4.8	3	4	0.065116	18992
573	cash	flow	3.72	134	4	0.066721	5244
13	coiled	spring	3.89	1	6	0.06701	5983
512	daily	post	4.3	74	4	0.067112	9339
551	shot	dead	4.07	74	4	0.06722	12494
243	rapid	growth	2.89	53	6	0.067406	12982
13	pelvic	floor	4.8	4	5	0.068421	11556
508	credit	card	4.9	109	4	0.068881	5739
535	bloody	hell	2.41	75	4	0.073977	5315

12	hoax	call	4	2	4	0.075472	19484
10	craggy	face	4.87	1	4	0.082645	34978
437	acid	rain	4.97	70	4	0.087963	6517
466	index	closed	3.37	52	6	0.100172	9877
17	larval	stage	4.64	3	5	0.116438	16565
8465	last	night	4.52	473	5	0.117273	36529
1815	award	title	3.32	50	5	0.120087	9790
10	slush	fund	3.18	1	4	0.121951	6407
428	rugby	union	3.38	39	5	0.124238	17607
599	inner	city	4.79	65	4	0.133408	23247
305	fleet	street	4.75	25	6	0.135737	19614
187	eighth	army	4.7	21	4	0.146322	11441
12	darn	sight	3.21	1	5	0.153846	6712
11	puck	fair	2.39	1	4	0.183333	9210
234	bile	acid	4.25	17	4	0.197802	4968
14	chow	test	3.93	1	4	0.208955	13701
255	toxic	waste	3.24	20	5	0.209016	6762
3624	date	award	4.14	92	5	0.222947	15114
2745	wide	range	3.22	115	5	0.22628	20427
452	ozone	layer	3.52	14	5	0.351751	2543
1872	armed	forces	2.69	40	6	0.390163	11775
10	markup	language	2.35	1	8	0.454545	18778
207	barbed	wire	4.72	1	4	0.713793	2269
0	berate	ivnolved	0	0	8	0	0
0	large	wtis	0	0	4	0	0
0	give	langugae	0	0	8	0	0
0	larval	ofod	0	0	4	0	0
0	reaches	timesr	0	0	6	0	0
0	fifty	afarid	0	0	6	0	0
0	tides	muave	0	0	5	0	0
0	want	scayl	0	0	5	0	0
0	canned	aggs	0	0	4	0	0
0	ozone	acheive	0	0	7	0	0
0	recede	scoks	0	0	5	0	0
0	pelvic	strete	0	0	6	0	0
0	toxic	tset	0	0	4	0	0
0	rapid	sprnig	0	0	6	0	0
0	hostile	rnik	0	0	4	0	0
0	people	numebr	0	0	6	0	0
0	drubs	nugde	0	0	5	0	0
0	glide	dorop	0	0	5	0	0
0	meet	ogne	0	0	4	0	0
0	good	adavnce	0	0	7	0	0
0	less	fuond	0	0	5	0	0
0	coiled	shrot	0	0	5	0	0
0	oval	hispter	0	0	7	0	0

0	glabrous	eeys	0	0	4	0	0
0	panics	gruop	0	0	5	0	0
0	loony	numebr	0	0	6	0	0
0	hodiernal	aronud	0	0	6	0	0
0	beady	insdie	0	0	6	0	0
0	sweeper	gian	0	0	4	0	0
0	plate	cgae	0	0	4	0	0
0	cell	swtich	0	0	6	0	0
0	shady	depe	0	0	4	0	0
0	research	alnog	0	0	5	0	0
0	abaya	pian	0	0	4	0	0
0	they	baet	0	0	4	0	0
0	polyp	egiht	0	0	5	0	0
0	inunct	rdue	0	0	4	0	0
0	should	baer	0	0	4	0	0
0	blotter	sesne	0	0	5	0	0
0	benthos	letf	0	0	4	0	0
0	your	anwser	0	0	6	0	0
0	fipple	ssytem	0	0	6	0	0
0	will	appael	0	0	6	0	0
0	cadged	takn	0	0	4	0	0
0	little	bbay	0	0	4	0	0
0	this	anicent	0	0	7	0	0
0	gilded	objetc	0	0	6	0	0
0	even	aomng	0	0	5	0	0
0	downright	nealry	0	0	6	0	0
0	volt	deifned	0	0	7	0	0
0	behave	suop	0	0	4	0	0
0	that	palce	0	0	5	0	0
0	deedy	meetr	0	0	5	0	0
0	canorous	corenr	0	0	6	0	0
0	first	anunal	0	0	6	0	0
0	revolve	lakc	0	0	4	0	0
0	lentil	alnoe	0	0	5	0	0
0	septic	ekep	0	0	4	0	0
0	snare	ahppen	0	0	6	0	0
0	strict	campiagn	0	0	8	0	0
0	with	nubmer	0	0	6	0	0
0	certain	amuont	0	0	6	0	0
0	shitty	durm	0	0	4	0	0
0	would	appaer	0	0	6	0	0
0	miaow	semll	0	0	5	0	0
0	musty	mahcine	0	0	7	0	0
0	effable	chesee	0	0	6	0	0
0	curd	failrue	0	0	7	0	0
0	crotch	sewnig	0	0	6	0	0

0	omophagy	sihg	0	0	4	0	0
0	total	figrue	0	0	6	0	0
0	research	swtich	0	0	6	0	0
0	sickle	sttae	0	0	5	0	0
0	clearly	graet	0	0	5	0	0
0	will	pltae	0	0	5	0	0
0	just	deisgn	0	0	6	0	0
0	could	culb	0	0	4	0	0
0	common	gruond	0	0	6	0	0
0	whacking	pnoy	0	0	4	0	0
0	worldly	piont	0	0	5	0	0
0	marbled	tcuk	0	0	4	0	0
0	dimmer	digets	0	0	6	0	0
0	uniped	dsah	0	0	4	0	0
0	support	frie	0	0	4	0	0
0	runny	scorll	0	0	6	0	0
0	ghostly	lmip	0	0	4	0	0
0	logomachy	soto	0	0	4	0	0
0	very	braed	0	0	5	0	0
0	nacarat	salst	0	0	5	0	0
0	from	godl	0	0	4	0	0
0	musket	leis	0	0	4	0	0
0	crusty	haed	0	0	4	0	0
0	zoolatry	buisness	0	0	8	0	0
0	carat	gruop	0	0	5	0	0
0	labarum	arae	0	0	4	0	0
0	thurifer	blie	0	0	4	0	0
0	martlet	perss	0	0	5	0	0
0	ratite	cirsp	0	0	5	0	0
0	heart	behnid	0	0	6	0	0
0	lagging	yrads	0	0	5	0	0
0	must	efefct	0	0	6	0	0
0	cadence	nihgt	0	0	5	0	0
0	croquet	pbu	0	0	3	0	0
0	living	clel	0	0	4	0	0
0	sinker	haet	0	0	4	0	0
0	panurgic	satb	0	0	4	0	0
0	jumentous	tiol	0	0	4	0	0
0	wanker	aawy	0	0	4	0	0
0	hallowed	agll	0	0	4	0	0
0	screwy	kist	0	0	4	0	0
0	solander	darw	0	0	4	0	0
0	about	nsoe	0	0	4	0	0
0	spent	godos	0	0	5	0	0
0	slug	gags	0	3	4	0	89
0	bloody	hlel	0	0	4	0	0

0	before	grwoth	0	0	6	0	0
0	vexes	lanugage	0	0	8	0	0
0	markup	wrie	0	0	4	0	0
0	geese	stlye	0	0	5	0	0
0	last	nihgt	0	0	5	0	0
0	second	hlaf	0	0	4	0	0
0	cycle	fnud	0	0	4	0	0
0	really	rian	0	0	4	0	0
0	date	awrad	0	0	5	0	0
0	inner	ctiy	0	0	4	0	0
0	hoax	laeyr	0	0	5	0	0
0	trillion	accuont	0	0	7	0	0
0	oaken	fiar	0	0	4	0	0
0	rethinks	acecpt	0	0	6	0	0
0	building	abvoe	0	0	5	0	0
0	rustic	thnig	0	0	5	0	0
0	make	srue	0	0	4	0	0
0	award	tilte	0	0	5	0	0
0	enforced	fvie	0	0	4	0	0
0	other	hnad	0	0	4	0	0
0	puck	tiems	0	0	5	0	0
0	always	steret	0	0	6	0	0
0	armed	focres	0	0	6	0	0
0	daily	psot	0	0	4	0	0
0	interest	aicd	0	0	4	0	0
0	slug	acecss	0	0	6	0	0
0	credit	crad	0	0	4	0	0
0	stippled	fcae	0	0	4	0	0
0	shot	daed	0	0	4	0	0
0	slush	pratcice	0	0	8	0	0
0	winds	acitve	0	0	6	0	0
0	wide	ragne	0	0	5	0	0
0	time	amry	0	0	4	0	0
0	gunboat	absnece	0	0	7	0	0
0	systems	acutal	0	0	6	0	0
0	craggy	watse	0	0	5	0	0
0	local	chidlren	0	0	8	0	0
0	faith	acorss	0	0	6	0	0
0	index	clsoed	0	0	6	0	0
0	case	unoin	0	0	5	0	0
0	cash	folw	0	0	4	0	0
0	many	dnoe	0	0	4	0	0
0	briskly	albe	0	0	4	0	0
0	abase	clal	0	0	4	0	0

Appendix 2: Word list (Experiment 2)

Bigram frequency	Prime	Target	Target concreteness	Target frequency	diversity	Target letters	Transitional probability
0	have	sihgt	0	0	0	0	0
0	snuffles	moedl	0	0	0	0	0
0	bile	ptah	0	0	0	0	0
0	secret	whsoe	0	0	0	0	0
0	their	fodo	0	0	0	0	0
0	chilli	acll	0	0	0	0	0
0	eighth	atcion	0	0	0	0	0
0	recent	adrdess	0	0	0	0	0
0	gimlet	coidng	0	0	0	0	0
0	acid	adedd	0	0	0	0	0
0	chow	trdae	0	0	0	0	0
0	rugby	stgae	0	0	0	0	0
0	lifer	hgus	0	0	0	0	0
0	barbed	folor	0	0	0	0	0
0	quickly	allwo	0	0	0	0	0
0	darn	agered	0	0	0	0	0
0	fleet	adivce	0	0	0	0	0
0	way	agere	0	0	0	0	0
0	year	hunderd	0	0	0	0	0
0	ski	aicrraft	0	0	0	0	0
0	fart	ocurse	0	0	0	0	0
0	funny	adlut	0	0	0	0	0
0	berate	ivnolved	0	0	0	0	0
0	large	wtis	0	0	0	0	0
0	give	langugae	0	0	0	0	0
0	larval	ofod	0	0	0	0	0
0	reaches	timesr	0	0	0	0	0
0	fifty	afarid	0	0	0	0	0
0	tides	muave	0	0	0	0	0
0	want	scayl	0	0	0	0	0
0	canned	aggs	0	0	0	0	0
0	ozone	acheive	0	0	0	0	0
0	recede	scoks	0	0	0	0	0
0	pelvic	strete	0	0	0	0	0
0	toxic	tset	0	0	0	0	0
0	rapid	sprnig	0	0	0	0	0
0	hostile	rnik	0	0	0	0	0
0	people	numebr	0	0	0	0	0
0	drubs	nugde	0	0	0	0	0
0	glide	dorop	0	0	0	0	0
0	meet	ogne	0	0	0	0	0
0	good	adavnce	0	0	0	0	0
0	less	fuond	0	0	0	0	0

0	coiled	shrot	0	0	0	0	0
0	oval	hispter	0	0	0	0	0
11	volt	meter	4.7	487	1	5	0.085938
0	glabrous	eeys	0	0	0	0	0
0	panics	gruop	0	0	0	0	0
0	behave	pub	4.71	3821	0	3	0
21	beady	eyes	4.85	29706	2	4	0.375
0	loony	numebr	0	0	0	0	0
0	hodiernal	aronud	0	0	0	0	0
0	beady	insdie	0	0	0	0	0
0	abaya	digest	3.07	475	0	6	0
0	sweeper	gian	0	0	0	0	0
12	gilded	cage	5	1021	2	4	0.041522
1071	would	appear	3.13	10914	1036	6	0.004197
0	plate	cgae	0	0	0	0	0
0	miaow	point	3.39	40274	0	5	0
0	cell	swtich	0	0	0	0	0
35	with	number	3.3	49385	4777	6	5.37E-05
0	shady	depe	0	0	0	0	0
0	research	alnog	0	0	0	0	0
19	septic	tank	4.8	3324	2	4	0.208791
768	certain	amount	2.74	15429	300	6	0.035505
10	curd	cheese	4.7	2589	1	6	0.097087
375	came	along	2.14	19335	171	5	0.00795
72	revolve	around	1.96	45286	1	6	0.566929
0	abaya	pian	0	0	0	0	0
0	they	baet	0	0	0	0	0
0	polyp	egiht	0	0	0	0	0
12	downright	rude	2.52	985	2	4	0.043011
0	inunct	rdue	0	0	0	0	0
0	should	baer	0	0	0	0	0
10	snare	drum	4.96	985	1	4	0.104167
0	glabrous	tuck	3.86	468	0	4	0
0	canorous	pony	4.9	710	0	4	0
600	were	almost	1.66	31588	2752	6	0.001859
12	sweeper	system	2.94	44674	1	6	0.078947
367	that	place	3.48	48651	5217	5	0.000329
0	blotter	sesne	0	0	0	0	0
0	benthos	letf	0	0	0	0	0
0	your	anwser	0	0	0	0	0
0	fipple	ssytem	0	0	0	0	0
0	will	appael	0	0	0	0	0
0	cadged	takn	0	0	0	0	0
0	hodiernal	lies	3.11	5268	0	4	0
0	little	bbay	0	0	0	0	0
0	this	anicent	0	0	0	0	0

0	gilded	objetc	0	0	0	0	0
0	even	aomng	0	0	0	0	0
0	downright	nealry	0	0	0	0	0
117	will	appeal	1.73	11002	1128	6	0.00046
0	fipple	state	3.52	39112	0	5	0
0	volt	deifned	0	0	0	0	0
0	behave	suop	0	0	0	0	0
0	that	palce	0	0	0	0	0
0	inunct	gall	2.6	1150	0	4	0
0	deedy	meetr	0	0	0	0	0
34	loony	left	3.7	47089	1	4	0.225166
186	even	among	2.38	22864	541	5	0.002581
0	canorous	corenr	0	0	0	0	0
0	first	anunal	0	0	0	0	0
0	revolve	lakc	0	0	0	0	0
0	lentil	alnoe	0	0	0	0	0
0	septic	ekep	0	0	0	0	0
16	musty	smell	3.7	3755	2	5	0.113475
0	benthos	head	4.75	37906	0	4	0
73	little	baby	5	9070	828	4	0.001518
47	common	object	3.66	6325	269	6	0.002575
0	snare	ahppen	0	0	0	0	0
0	blotter	group	4.12	41547	0	5	0
0	panics	yards	4.82	3678	0	5	0
0	strict	campiagn	0	0	0	0	0
0	with	nubmer	0	0	0	0	0
0	certain	amuont	0	0	0	0	0
0	cadged	limp	4.15	516	0	4	0
0	shitty	durm	0	0	0	0	0
0	would	appaer	0	0	0	0	0
10	polyp	group	4.12	41547	1	5	0.113636
84	this	ancient	2.04	5083	2909	7	0.000181
0	miaow	semll	0	0	0	0	0
0	deedy	scroll	4.11	214	0	6	0
67	first	annual	1.78	8154	1261	6	0.000564
0	musty	mahcine	0	0	0	0	0
0	effable	heat	3.79	5957	0	4	0
0	effable	chesee	0	0	0	0	0
64	strict	sense	2.61	21935	38	5	0.030933
226	they	beat	3.97	5675	1616	4	0.000521
0	shitty	night	4.52	36529	0	5	0
12	shady	corner	4.61	7500	1	6	0.043478
125	should	bear	4.88	5799	564	4	0.001124
10	lentil	soup	4.72	1353	1	4	0.3125
0	curd	failrue	0	0	0	0	0
256	your	answer	2.89	14421	1871	6	0.001851

0	crotch	sewnig	0	0	0	0	0
0	omophagy	sihg	0	0	0	0	0
0	total	figrue	0	0	0	0	0
11	hallowed	ground	4.77	16200	1	6	0.083333
48	support	machine	4.25	8938	208	7	0.001609
80	total	lack	2.04	10068	261	4	0.004554
0	research	swtich	0	0	0	0	0
0	sickle	sttae	0	0	0	0	0
49	lagging	behind	3.48	23698	1	6	0.494949
0	clearly	graet	0	0	0	0	0
0	zoolatry	business	3.28	35758	0	8	0
0	panurgic	stab	4.07	428	0	4	0
0	will	pltae	0	0	0	0	0
0	just	deisgn	0	0	0	0	0
0	could	culb	0	0	0	0	0
0	common	gruond	0	0	0	0	0
11	ghostly	figure	3.63	17613	1	6	0.042308
0	whacking	pnoy	0	0	0	0	0
0	ratite	crisp	3.69	798	0	5	0
0	worldly	piont	0	0	0	0	0
0	marbled	tcuk	0	0	0	0	0
0	labarum	area	3.72	35144	0	4	0
11	dimmer	switch	4.07	3316	2	6	0.150685
19	worldly	goods	4.26	10142	1	5	0.076305
0	dimmer	digets	0	0	0	0	0
12	whacking	great	1.81	45217	1	5	0.27907
11	croquet	club	3.78	16465	1	4	0.076923
0	uniped	dash	3.39	758	0	4	0
0	uniped	dsah	0	0	0	0	0
363	clearly	defined	2.07	5898	215	7	0.02365
0	jumentous	toil	2.67	182	0	4	0
239	living	alone	2.86	13265	154	5	0.01493
0	support	frie	0	0	0	0	0
0	runny	scorll	0	0	0	0	0
230	must	keep	2.37	27813	448	4	0.003169
0	ghostly	lmip	0	0	0	0	0
0	logomachy	soto	0	0	0	0	0
0	very	braed	0	0	0	0	0
0	solander	draw	3.97	7398	0	4	0
0	logomachy	soot	4.61	196	0	4	0
0	nacarat	salst	0	0	0	0	0
22	from	number	3.3	49385	3081	6	5.32E-05
0	thurifer	bile	4.46	1183	0	4	0
383	about	eight	4.04	17309	1217	5	0.001943
71	carat	gold	4.81	7792	1	4	0.496503
0	from	godl	0	0	0	0	0

0	musket	leis	0	0	0	0	0
0	crusty	haed	0	0	0	0	0
107	heart	failure	2.08	7763	120	7	0.007792
0	zoolatry	buisness	0	0	0	0	0
0	martlet	press	3.9	13115	0	5	0
0	omophagy	sigh	3.89	1171	0	4	0
0	carat	gruop	0	0	0	0	0
0	labarum	arae	0	0	0	0	0
0	thurifer	blie	0	0	0	0	0
33	runny	nose	4.89	4337	1	4	0.358696
0	crotch	sewing	4.4	606	0	6	0
0	martlet	perss	0	0	0	0	0
37	crusty	bread	4.92	3770	1	5	0.253425
0	ratite	cirsp	0	0	0	0	0
350	could	happen	1.78	8760	790	6	0.002078
0	heart	behnid	0	0	0	0	0
146	very	deep	3.38	10700	987	4	0.001186
190	just	inside	3.67	14309	895	6	0.001467
10	cadence	design	3.27	12939	1	6	0.192308
0	lagging	yrads	0	0	0	0	0
0	must	efefct	0	0	0	0	0
32	sinker	plate	4.77	4096	1	5	0.470588
0	cadence	nihgt	0	0	0	0	0
84	sickle	cell	4.44	5518	1	4	0.509091
0	croquet	pbu	0	0	0	0	0
0	living	clel	0	0	0	0	0
0	sinker	haet	0	0	0	0	0
0	panurgic	satb	0	0	0	0	0
0	jumentous	tiol	0	0	0	0	0
0	wanker	aawy	0	0	0	0	0
0	hallowed	agll	0	0	0	0	0
0	screwy	kist	0	0	0	0	0
10	marbled	effect	1.8	23361	1	6	0.091743
11	musket	fire	4.68	14104	1	4	0.183333
0	nacarat	salts	4.89	406	0	5	0
50	spent	nearly	1.89	11494	100	6	0.004243
0	wanker	away	2.23	38747	0	4	0
37	research	campaign	3	9518	284	8	0.001371
0	screwy	kits	4.47	431	0	4	0
0	solander	darw	0	0	0	0	0
0	about	nsoe	0	0	0	0	0
151	will	gain	2.24	5218	1128	4	0.000593
0	spent	godos	0	0	0	0	0
0	stippled	fcae	0	0	0	0	0
0	inner	ctiy	0	0	0	0	0
0	building	abvoe	0	0	0	0	0

0	rethinks	acecpt	0	0	0	0	0
0	credit	crad	0	0	0	0	0
0	date	awrad	0	0	0	0	0
0	bloody	hlel	0	0	0	0	0
0	gunboat	absnece	0	0	0	0	0
0	local	chidlren	0	0	0	0	0
0	vexes	lanugage	0	0	0	0	0
0	daily	psot	0	0	0	0	0
0	award	tilte	0	0	0	0	0
0	slug	acecss	0	0	0	0	0
0	wide	ragne	0	0	0	0	0
0	cycle	fnud	0	0	0	0	0
0	winds	acitve	0	0	0	0	0
0	briskly	albe	0	0	0	0	0
0	abase	clal	0	0	0	0	0
0	other	hnad	0	0	0	0	0
0	slush	pratcice	0	0	0	0	0
0	faith	acorss	0	0	0	0	0
0	cash	folw	0	0	0	0	0
0	oaken	fiar	0	0	0	0	0
0	systems	acutal	0	0	0	0	0
0	interest	aicd	0	0	0	0	0
0	enforced	fvie	0	0	0	0	0
0	many	dnoe	0	0	0	0	0
0	armed	focres	0	0	0	0	0
0	markup	wrie	0	0	0	0	0
0	really	rian	0	0	0	0	0
0	trillion	accuont	0	0	0	0	0
0	case	unoin	0	0	0	0	0
0	hoax	laeyr	0	0	0	0	0
0	before	grwoth	0	0	0	0	0
0	rustic	thnig	0	0	0	0	0
0	puck	tiems	0	0	0	0	0
0	last	nihgt	0	0	0	0	0
0	always	steret	0	0	0	0	0
0	index	clsoed	0	0	0	0	0
0	shot	daed	0	0	0	0	0
0	make	srue	0	0	0	0	0
0	craggy	watse	0	0	0	0	0
0	geese	stlye	0	0	0	0	0
0	time	amry	0	0	0	0	0
0	second	hlaf	0	0	0	0	0

Appendix 3: Word lists (Experiments 4, 6, & 8)

Bigram frequency	Prime	Target	Target concreteness	Target frequency	diversity	Target letters	Transitional probability
383	about	eight	4.04	17309	493	5	0.001943
0	above	sturse	0	0	116	0	0.000466
0	abyss	cuzzed	0	0	2	0	0.08
0	account	musts	0	0	76	0	0.001573
0	acronym	easy	2.07	14774	3	4	0.159664
0	addict	gouls	0	0	2	0	0.068966
0	almost	grefs	0	0	182	0	0.00114
0	assemblage	please	1.64	14351	1412	6	4.64E-05
0	backlog	swinds	0	0	1	0	0.574257
85	back	pain	3.5	7338	259	4	0.000831
21	beady	eyes	4.85	29706	2	4	0.375
0	binge	libes	0	0	2	0	0.090278
0	blip	sunes	0	0	1	0	0.185185
0	bottom	theep	0	0	63	0	0.00861
0	bounty	blull	0	0	2	0	0.06701
0	breeder	yerp	0	0	1	0	0.106061
10	cadence	design	3.27	12939	1	6	0.192308
375	came	along	2.14	19335	143	5	0.00795
71	carat	gold	4.81	7792	1	4	0.496503
0	carnage	snerf	0	0	3	0	0.055249
768	certain	amount	2.74	15429	117	6	0.035505
0	chapter	swach	0	0	56	0	0.001074
0	cheese	clett	0	0	25	0	0.00618
0	chevron	sound	3.7	14542	36	5	0.003159
363	clearly	defined	2.07	5898	128	7	0.02365
0	clink	chims	0	0	17	0	0.013541
0	column	brounced	0	0	25	0	0.006683
0	come	greeds	0	0	12	0	0.038314
0	comely	answer	2.89	14421	12	6	0.038314
47	common	object	3.66	6325	155	6	0.002575
0	conflate	art	4.17	15587	22	3	0.014515
0	consul	crynch	0	0	3	0	0.047847
0	contour	kneant	0	0	2	0	0.085
350	could	happen	1.78	8760	617	6	0.002078
0	course	flates	0	0	188	0	0.00066
11	croquet	club	3.78	16465	1	4	0.076923
37	crusty	bread	4.92	3770	1	5	0.253425
0	culprit	spralf	0	0	1	0	0.096447
10	curd	cheese	4.7	2589	1	6	0.097087
0	days	cret	0	0	129	0	0.000786
0	deal	shreths	0	0	73	0	0.0009
0	design	shroft	0	0	86	0	0.00085
0	detritus	works	3.79	14528	1	5	0.908213

0	dimmer	spalms	0	0	3	0	0.107692
11	dimmer	switch	4.07	3316	1	6	0.150685
0	dissemble	fire	4.68	14104	1	4	0.211538
0	dope	spuits	0	0	340	0	0.000737
12	downright	rude	2.52	985	1	4	0.043011
0	elixir	comes	2.27	15968	12	5	0.012308
3	entire	squad	3.65	1095	1	5	0.120567
0	equals	fenth	0	0	51	0	0.00172
0	errand	jeight	0	0	2	0	0.07483
186	even	among	2.38	22864	372	5	0.002581
0	exceed	dwic	0	0	3	0	0.031175
0	exclaim	forward	2.66	15205	2	7	0.206422
0	fell	biewed	0	0	77	0	0.00327
0	feud	thrieled	0	0	2	0	0.173913
67	first	annual	1.78	8154	753	6	0.000564
0	fodder	bleuth	0	0	2	0	0.082192
0	force	ouse	0	0	117	0	0.000698
0	form	keiled	0	0	125	0	0.000464
0	fray	cloist	0	0	2	0	0.054726
0	friend	slaids	0	0	88	0	0.001779
0	frill	strisped	0	0	1	0	0.129412
11	ghostly	figure	3.63	17613	1	6	0.042308
12	gilded	cage	5	1021	2	4	0.041522
0	giver	freins	0	0	156	0	0.000332
0	glimmer	ghond	0	0	2	0	0.044983
0	graft	phlug	0	0	1	0	0.092593
0	grimace	scrcair	0	0	1	0	0.154839
0	habitat	bed	5	15896	17	3	0.008355
11	hallowed	ground	4.77	16200	1	6	0.083333
107	heart	failure	2.08	7763	86	7	0.007792
0	helix	clearly	2.04	15349	699	7	2.05E-05
0	hoard	thriff	0	0	1	0	0.275132
0	hoary	despite	1.33	14592	5	7	0.041667
0	holds	deeled	0	0	87	0	0.000735
0	hundred	splurb	0	0	72	0	0.001151
0	imbue	force	3	15752	41	5	0.002404
0	income	clerb	0	0	86	0	0.002739
0	jeering	mitched	0	0	2	0	0.1375
0	jink	natural	1.85	14315	1	7	0.092593
0	jolt	zamn	0	0	3	0	0.0625
0	jumbo	truts	0	0	1	0	0.223881
0	just	drothed	0	0	587	0	0.000355
190	just	inside	3.67	14309	587	6	0.001467
0	kilo	toosed	0	0	1	0	0.156863
0	know	gailed	0	0	205	0	0.000341
49	lagging	behind	3.48	23698	1	6	0.494949

0	latvia	dranns	0	0	70	0	0.002705
0	leakage	orm	0	0	3	0	0.247664
10	lentil	soup	4.72	1353	1	4	0.3125
0	ley	poor	2.7	15125	1	4	0.163934
73	little	baby	5	9070	588	4	0.001518
239	living	alone	2.86	13265	112	5	0.01493
34	loony	left	3.7	47089	1	4	0.225166
0	lymph	throached	0	0	1	0	0.205405
10	marbled	effect	1.8	23361	1	6	0.091743
0	mark	chaph	0	0	18	0	0.006859
0	marked	flince	0	0	45	0	0.002248
0	median	blurled	0	0	10	0	0.033659
0	meeting	jows	0	0	114	0	0.000625
0	member	churke	0	0	46	0	0.002058
0	morning	rhast	0	0	109	0	0.000473
11	musket	fire	4.68	14104	1	4	0.183333
0	must	blult	0	0	362	0	0.000441
230	must	keep	2.37	27813	362	4	0.003169
16	musty	smell	3.7	3755	2	5	0.113475
0	name	scrawks	0	0	126	0	0.001749
0	nemesis	ways	2	14932	56	4	0.002304
0	news	twurk	0	0	818	0	0.000335
0	number	scrobes	0	0	95	0	0.000223
0	oaf	offer	2.23	15873	8	5	0.021614
0	only	frawns	0	0	728	0	0.004146
0	optic	hond	0	0	3	0	0.073733
0	opulent	recent	2.5	15858	46	6	0.005346
0	orate	red	4.24	15136	2048	3	0.008033
0	outflow	qwouse	0	0	2	0	0.440415
0	part	pofts	0	0	94	0	0.000277
0	party	gwanc	0	0	189	0	0.000521
0	patriot	knenched	0	0	2	0	0.078947
0	period	flonned	0	0	107	0	0.000494
0	pike	ghowse	0	0	2	0	0.029268
0	place	thwissed	0	0	179	0	0.000822
0	plan	drurze	0	0	96	0	0.001336
0	plate	dwists	0	0	29	0	0.003174
0	point	thrimbs	0	0	152	0	0.00725
10	polyp	group	4.12	41547	1	5	0.113636
0	port	wofts	0	0	29	0	0.00503
0	pounds	fusk	0	0	61	0	0.001602
0	present	wushed	0	0	185	0	0.001568
0	probe	gwoints	0	0	16	0	0.02771
0	proton	grorgues	0	0	2	0	0.141463
0	putter	zez	0	0	1	0	0.075472
0	quad	account	3.08	15891	1	7	0.235294

0	quip	phrinks	0	0	1	0	0.2
0	range	cuilts	0	0	55	0	0.000587
0	rant	suill	0	0	70	0	0.008465
0	rasp	phres	0	0	1	0	0.382716
0	rate	flapsed	0	0	97	0	0.001417
0	reach	gwerge	0	0	57	0	0.001477
0	rebate	stoos	0	0	4	0	0.045249
0	reggae	grulps	0	0	2	0	0.103448
37	research	campaign	3	9518	122	8	0.001371
0	rest	shebb	0	0	74	0	0.001543
72	revolve	around	1.96	45286	1	6	0.566929
0	ribbed	final	2.67	15648	7	5	0.063025
0	rise	gwodd	0	0	54	0	0.001891
0	rubric	pruns	0	0	1	0	0.457627
33	runny	nose	4.89	4337	1	4	0.358696
0	sable	field	4.26	15298	7	5	0.021142
0	saccade	girl	4.85	15762	7	4	0.021142
0	saline	thalse	0	0	5	0	0.032
0	screen	blowns	0	0	46	0	0.002802
0	second	swarp	0	0	355	0	0.000651
0	section	veek	0	0	69	0	0.001075
0	seeker	swowd	0	0	1	0	0.126582
19	septic	tank	4.8	3324	1	4	0.208791
0	shading	smurds	0	0	2	0	0.081761
12	shady	corner	4.61	7500	1	6	0.043478
125	should	bear	4.88	5799	436	4	0.001124
84	sickle	cell	4.44	5518	1	4	0.509091
32	sinker	plate	4.77	4096	1	5	0.470588
0	sinner	spriege	0	0	1	0	0.07483
0	size	blorked	0	0	59	0	0.002598
0	sleeper	chault	0	0	1	0	0.068493
10	snare	drum	4.96	985	1	4	0.104167
50	spent	nearly	1.89	11494	80	6	0.004243
0	story	clealed	0	0	85	0	0.002413
64	strict	sense	2.61	21935	18	5	0.030933
0	stun	flugged	0	0	3	0	0.121528
0	subject	keaked	0	0	99	0	0.000934
0	sunday	grourn	0	0	87	0	0.000868
48	support	machine	4.25	8938	144	7	0.001609
0	sweden	croiced	0	0	1	0	0.089286
12	sweeper	system	2.94	44674	1	6	0.078947
0	talisman	project	3.62	15215	20	7	0.014293
0	tape	cleeced	0	0	49	0	0.002665
0	taxes	lead	4.1	14555	9	4	0.026087
0	tenet	sprerfs	0	0	1	0	0.6
367	that	place	3.48	48651	2074	5	0.000329

0	then	cack	0	0	690	0	0.002475
0	theses	blole	0	0	395	0	0.000112
226	they	beat	3.97	5675	970	4	0.000521
0	thing	prused	0	0	132	0	0.000767
84	this	ancient	2.04	5083	1588	7	0.000181
0	this	foaths	0	0	1588	0	3.02E-05
0	thud	fobed	0	0	2	0	0.235023
0	timbre	rherked	0	0	2	0	0.87931
0	time	frackt	0	0	356	0	0.000116
0	today	plym	0	0	110	0	0.000531
0	toggle	scunged	0	0	3591	0	8.79E-05
80	total	lack	2.04	10068	152	4	0.004554
0	trade	thwogs	0	0	110	0	0.001151
0	trellis	front	3.77	15106	63	5	0.002028
0	type	prench	0	0	60	0	0.001386
0	typhoid	hoursed	0	0	2	0	0.573034
0	typing	drounced	0	0	2	0	0.573034
0	union	spoot	0	0	107	0	0.000966
0	verse	crolt	0	0	20	0	0.006944
146	very	deep	3.38	10700	542	4	0.001186
0	view	ghelved	0	0	109	0	0.015115
11	volt	meter	4.7	487	1	5	0.085938
0	vortex	fute	0	0	1	0	0.121339
0	warren	kept	2.79	14306	144	4	0.000681
0	weapon	franced	0	0	23	0	0.021191
0	weekday	slinked	0	0	135	0	0.00031
0	week	wef	0	0	135	0	0.008535
600	were	almost	1.66	31588	1170	6	0.001859
0	were	gleld	0	0	1170	0	0.00031
12	whacking	great	1.81	45217	1	5	0.27907
0	which	bown	0	0	878	0	7.53E-05
0	while	blypts	0	0	206	0	0.000194
0	wicket	march	4.03	15997	10	5	0.010949
0	wigan	clulched	0	0	94	0	0.000585
117	will	appeal	1.73	11002	870	6	0.00046
151	will	gain	2.24	5218	870	4	0.000593
35	with	number	3.3	49385	1743	6	5.37E-05
0	wool	clarge	0	0	21	0	0.009534
19	worldly	goods	4.26	10142	1	5	0.076305
1071	would	appear	3.13	10914	814	6	0.004197
0	xylophone	green	4.07	14637	6	5	0.266366
0	year	snarfed	0	0	226	0	0.000244
0	yonder	month	4.2	15011	1	5	0.092593
256	your	answer	2.89	14421	1068	6	0.001851