Chapter 1

Introduction

Living in a world in which the Internet allows us to communicate with people in different countries in an instant, and where the increasing convenience of travelling to and living in countries different to our home country is easier than ever, it is no surprise that bilingualism has become the rule instead of being the exception. Bilingualism here refers to the ability to produce meaningful utterances in two or more languages (Grosjean, 2010). It has been estimated that more than half of the world's population is bilingual (Grosjean, 2010). For instance, 66% of Europeans aged between 25 and 64 years speak at least two languages and, as reported in 2011, within the 28 member states of the European Union, 83% of all students in primary and lower secondary school as well as 94% of all students in upper secondary school were learning English as a second language (European Comission, 2013). While it was initially thought that being bilingual had negative consequences (i.e., when learning to use either of the two or more languages properly), nowadays it is widely agreed that being bilingual is actually beneficial. Not only does speaking two languages allow the bilingual to communicate effectively in two or more languages, reaching a greater audience and being able to obtain information from more sources, it has also been shown that bilingualism has a positive impact on cognitive functioning (for a review see Bialystok, 2011).

The question of how bilinguals are able to speak two or more languages effortlessly has received an increasing amount of attention in the last decade. Previously, the topics of bilingualism and speech production had received little attention within the field of psy-

cholinguistics which mainly focused on monolingual language processing. Therefore, "our knowledge of the mechanisms and representations involved in bilingual language production is limited" (Costa & Santesteban, 2006, p. 115). However, theoretical frameworks to investigate bilingual language production have now been established which allow for the study of these mechanisms and representations more closely (Costa & Santesteban, 2006).

Given every monolingual has the potential to become bilingual, it can be argued that any model of language representation and production should be able to incorporate the bilingual case (De Bot, 1992). Also, most research into bilingual word production uses what is known about monolingual word production as a scaffold (e.g., Costa & Caramazza, 1999; Finkbeiner, Almedeida, Janssen, & Caramazza, 2006; La Heij, 2005). Consequently, theories and evidence regarding both monolingual and bilingual word production will be reviewed below.

1.1 Words in the Mental Lexicon

When considering the storage of words in the mind, the complexity of the term 'word' itself needs to first be considered. Everyone 'knows' what a word is but, when trying to give a definition of 'a word', the complexity of this concept becomes manifest (see Kersten, 2009). Foremost, there are a number of features of a word that need to be taken into account. A word has a meaning. It belongs to a grammatical class and it has a written and a spoken form that might change depending on the context in which it is used. To represent a word in the mind for correct use, these different features have to be dealt with. And indeed these features (with the exception of meaning) have been theorised to make up our internal representations of words (Aitchison, 2003).

The mental lexicon is defined as the place in the mind where humans store their internal word representations (Aitchison, 2003). These internal word representations are also known as lexical entries. Lexical entries consist of word-related information such as the semantic, syntactic, morphologic, orthographic, and phonological properties of a word that are required for a content word to represent a concept within a language and

for function words to do their job (Aitchison, 2003). This word-related information is often grouped into two entities: the lemma and the lexeme (see Aitchison, 2003; De Bot, 1992; Kersten, 2009; Levelt, 1999; Levelt et al., 1999; Vigliocco, Antonini, & Garret, 1997). The lemma contains syntactic and certain semantic information and links to a word's concept. The lemma is a commonly accepted concept among researchers (see Levelt, 1999). The lexeme is said to contain morphological information and information regarding the orthographic and phonological form of a word. In contrast to the lemma, the lexeme is a construct that is less agreed upon, with some researchers doubting its existence as a separate entity (see Caramazza & Miozzo, 1997). Together, the lemma and the lexeme make up a lexical entry and are stored in the mental lexicon.

Early evidence for a representational distinction between different features of a word came from the analysis of speech errors back in the 1970s (e.g., Cutler & Fay, 1977; Fromkin, 1973). Speech errors show systematic patterns that highlight the different features of a word (for examples see Fromkin, 1973). First, there are semantic errors in which, for instance, the opposite of a word is used (e.g., "This room is too damn hot – cold."). Second, parts of the sentence can be mixed up (e.g., "a laboratory in our own computer – a computer in our own laboratory"), whereby, however, only words of the same class are exchanged (e.g., noun for a noun or a verb for a verb). Third, there are errors wherein morphological aspects, such as affixes or suffixes, are exchanged (e.g., "infinity clauses – infinitive clauses"). Fourth, there are errors of form (e.g., "no strings attrached – no strings attached").

This evidence for the different features found in spontaneous speech was then later complemented with evidence from experimental paradigms. Lupker (1979) conducted the first picture-word interference study in which participants had to name pictures that were accompanied by a written distractor. He found that when a picture was accompanied by a semantically related distractor (e.g., cat – DOG), it was named more slowly than when it was accompanied by an unrelated (e.g., jug – DOG) or no distractor. This was interpreted as being caused by an overlap in semantic features between the distractor and the target word. A couple of years later, stimulus onset asynchrony (SOA) was introduced to the task which saw a variation of the onset of both stimuli (e.g., the distractor could be

presented milliseconds before or after the target; Glaser & Düngelhoff, 1984). This way, it was possible to investigate whether and at which time different features were processed. For instance, when a semantically related distractor was presented 150ms before the target it had an impact on target naming, while a phonologically related distractor had no impact presented this early. On the other hand, when a phonological distractor was presented up to 150ms after the target, it did influence target naming. Given these findings it was concluded that semantic features of a word were processed relatively early in the production process, while phonological features were processed later (Schriefers, Meyer, & Levelt, 1990). Evidence for the grouping of features into a lemma and a lexeme originates from the tip of the tongue (TOT) phenomenon. The TOT phenomenon describes the state in which someone wants to say a certain word but cannot manage to do so. When experiencing a TOT state, a person usually feels that they know exactly what they want to say, but simply cannot find the right word (Brown, 1991). Despite the apparent lack of information about the word's form, the person is able to define the word they want to produce and, more importantly, give syntactic information, such as the grammatical class of the word or a noun's gender (e.g., Vigliocco et al., 1997), pointing to a separation between semantic/syntactic and form information.

1.2 The Bilingual Mental Lexicon

A quite obvious question regarding the mental lexicon in the bilingual case is the question of whether the bilingual has two mental lexicons, one for each language, or one mental lexicon housing both languages. Back in the 1950s, Weinreich (1953) proposed a number of different solutions regarding this question which still bears a lot of similarity to today's views on this issue. Weinreich assumed that there were separate representations for semantic and lexical information and that the way in which these two kinds of representations are connected between languages depends on how the languages were acquired. As illustrated in Figure 1.1, in a compound organisation, semantic information is shared between the languages, but lexical information is stored separately. This organisation is proposed to be the result of a simultaneous acquisition and use of the two languages. In

a coordinate organisation, semantic and lexical representations exist separately for each language. This type of organisation is most likely achieved when L2 and L1 are acquired separately, such as when one language is spoken at home, while the other language is spoken in school. In a subordinate organisation, the L2 is connected to its meaning via its L1 equivalents. This organisation is proposed to be the result of L2 acquisition close to the L1 (e.g., when the L2 is acquired as a secondary language in school). An important assumption made by Weinreich is that an individual's organisation should be able to change depending on their language experience.

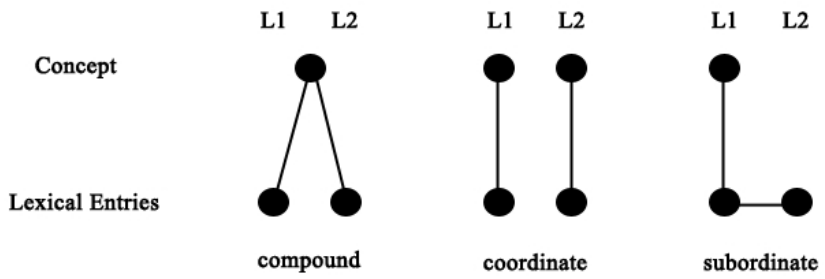

Figure 1.1: Weinreich's (1953) types of bilingual organisations (adapted from De Bot & Verspoor, 2005)

This proposal was adopted in the Revised Hierarchical Model (RHM; Kroll, Van Hell, Tokowicz, & Green, 2010), which proposes that late bilinguals first organise their L2 words around translation equivalents and not their concept (as in Weinreich, 1953, subordinate organisation). As the speaker becomes more proficient, an independent connection between the L2 and the concept is achieved (as in the compound system). Evidence for differences in organisation between bilinguals comes from the finding that when judging Spanish-Catalan translation pairs, high proficient bilinguals show interference when a target is not the translation equivalent, but semantically related to the equivalent (e.g., perro [dog] – hiena [hyena]) while low proficient bilinguals showed interference when a target was only form related (e.g., silla [chair] – cadena [chain; similar to cadira - chair]; Ferré, Sánchez-Casas, & Guash, 2006).

It has been argued, however, that both the compound and coordinate organisation make it hard to explain the cross-language phenomena of code-switching (e.g., Poulisse & Bongearts, 1994) and facilitation of cognate naming (e.g., Costa, Colomé, & Caramazza, 2000), therefore four different types of organisation have more recently been hypothesised for the bilingual mental lexicon (Hulstijn, 1997; Paradis, 2004). The extended system hypothesis refers to one storage system for both languages in which the two languages are only distinguished by language tags. According to the dual system hypothesis, both languages are stored separately. This is the least plausible proposal given code-switching and cognate effects. The tripartite system hypothesis describes the mental lexicon as including one storage for words from both languages, preferably similar words such as cognates, and also two separate storages for words that are language specific. Last, according to the subsystem hypothesis, words from both languages are contained in the same storage, and the distinction between words from different languages is achieved by connections between the words in the combined storage, which allow for subsystems to form. This last hypothesis is the most accepted so far (see De Bot, 2004).

1.3 Approaches to Representation and Processing

In the past there have been two fundamentally different approaches underlying theories for how language is represented and processed: the classical (see Fodor, 2008; Marr, 1982; Turing, 1950) and the connectionist (see Rumelhart, 1989; Rumelhart & McClelland, 1986; Smolensky, 1988) approaches. Nowadays most models are neither strictly classical, nor do they fully take a connectionist stance. However, they still rely on many of the assumptions provided by these approaches (see Levelt, 1999, for a review). Therefore the critical features of the classical and the connectionist approach to representation and processing are highlighted below.

1.3.1 Representation

In a classical architecture, information is represented by persistent symbols that carry an inherent meaning. Symbols are an abstract representation of information. They each

achieve this function through the possession of particular sets of formal properties which correspond to specific sets of processes (see Fodor, 2008; Marr, 1982; Turing, 1950). In contrast, in a connectionist architecture information is represented by activation patterns in a network of identical nodes. Meaning is purely based on reference through the association of different patterns of activation, and eventually with specific sensory profiles that reflect states of the world (see Rumelhart, 1989; Rumelhart & McClelland, 1986; Smolensky, 1988). Most current models of word production, for instance, assume a network structure, as proposed by the connectionist approach. However, some models also assume that nodes in these networks are symbolic in nature; nodes represent linguistic entities (see Levelt, 1999, for a review). Therefore, the purpose of this investigation is to look at how information is represented, and to look at the constraints on interaction between different clusters of information.

1.3.2 Processing

In a classical architecture, processing can be described in a rule-like manner. Mental operations interact with mental forms (symbols) to generate a priori products. Many mental operations and the symbols they operate on are domain specific; each mental domain has its own rules and symbols. This leads to discrete processing: processing takes places in stages, whereby each stage is dependent on the output of the previous stage (see Fodor, 2008; Levelt et al., 1999; Marr, 1982; Turing, 1950). In contrast, in a connectionist architecture involving parallel distributed processing, processing is activation based; representations are retrieved through activation. Activation spreads freely through the network. Different levels of processing can interact with each other, while each area of processing is happening (see Dell, 1986; Rumelhart, 1989; Rumelhart & McClelland, 1986; Smolensky, 1988). These two different approaches to processing are present in current models of word production (see Levelt, 1999, for a review).

1.4 Organisation within the Mental Lexicon

Unlike in a dictionary, entries in the mental lexicon are not sorted alphabetically. An alphabetical organisation would not allow humans to retrieve words as quickly and efficient as they do when they speak. Specifically, words at the end of the alphabet would be retrieved much slower than words at the beginning of the alphabet, which is not the case (e.g., 'zebra' is on average retrieved faster than 'abbey'; see Balota et al., 2007). Instead, entries in the mental lexicon seem to be organised on the basis of certain lexical characteristics not just initial sound (e.g., alphabetic). The two characteristics that influence retrieval time in word production, as an indicator of lexical organisation, are frequency and neighbourhood density (Aitchison, 2003).

Word frequency refers to how frequently a word occurs, that is how often a word is used. Research has found that high frequency words (i.e., words that are used often), are named faster in picture naming tasks than low frequency words (i.e., words that are used less often; Griffin & Bock, 1998; Jescheniak & Levelt, 1994). Assuming activation-based processing, this finding has been explained in terms of a lowered activation threshold for high frequency words compared to low frequency words which allows these words to be accessed faster (Jescheniak & Levelt, 1994). Alternatively, in a rule-based architecture, this finding might be explained in terms of high frequency words being sorted onto the top of a representation 'stack' because they are retrieved more often than low frequency words and go back onto the stack's top when retrieved. As a result high frequency words are then retrieved faster from the top the next time they are accessed (e.g., Forster, 1976).

Neighbourhood density refers to word clusterings with respect to form similarity. A word neighbour was originally defined as a word that differs from another word by one letter (e.g., horse – HOUSE; Coltheart, Davelaar, Johnasson, & Besner, 1977). Words with many neighbours have a high neighbourhood density as many of their neighbours are also neighbours among themselves, while words with few neighbours have a low neighbourhood density. Research has found that words from high-density neighbourhoods are named faster in a picture naming task than words from low-density neighbourhoods. Additionally, words from high-density neighbourhoods are also found to be less error prone than words

from low-density neighbourhoods (Vitevitch, 2002). In an activation-based architecture, target words from high-density neighbourhoods could be named faster and more reliably because their neighbours are also activated due to their similarity and, as a result, boost the target word with their activation (Vitevitch, 2002). Alternatively, in a rule-based architecture, words in high-density neighbourhoods could have a more efficient coding to set them apart from all the similar representations among them and, as a result, they could be retrieved faster (compare proposal in Forster & Taft, 1994).

Additionally, in contrast to frequency, neighbourhood density can be calculated across languages which could reveal more details regarding the relationships between the two lexicons of a bilingual. In the domain of word recognition, Van Heuven, Dijkstra, and Grainger (1998) were the first to investigate the effects of first language (L1) neighbourhood size of second language (L2) words. They found that in the absence of any L1 cues, L1 neighbourhood size impacted on the recognition of L2 words. This finding was interpreted as evidence for the existence of at least links between word forms in both languages and the activation of both languages in the context of otherwise monolingual recognition. Given that there do not seem to be more studies using cross-language neighbourhoods as a tool to investigate organisation of the bilingual mental lexicon, in Chapter 3 this thesis seeks to replicate Van Heuven et al.'s findings looking at form-related organisation in late bilingual word production using cross-language neighbours.

1.5 Models of Language Production

After looking at how words might be represented and organised, the next question is how these lexical entries are accessed and processed during word production. There are currently two prominent models of monolingual language production which are often used as a base for modelling bilingual language production (e.g., Costa, 2005; Finkbeiner, Gollan, & Caramazza, 2006; La Heij, 2005): A discrete network model (Levelt et al., 1999) and a cascading, interactive network model (Dell, 1986).

Levelt et al. (1999) proposed a discrete network model of word production. In this model (see Figure 1.2), word production happens as a result of a series of discrete stages

(e.g., conceptual preparation, lexical selection etc.).

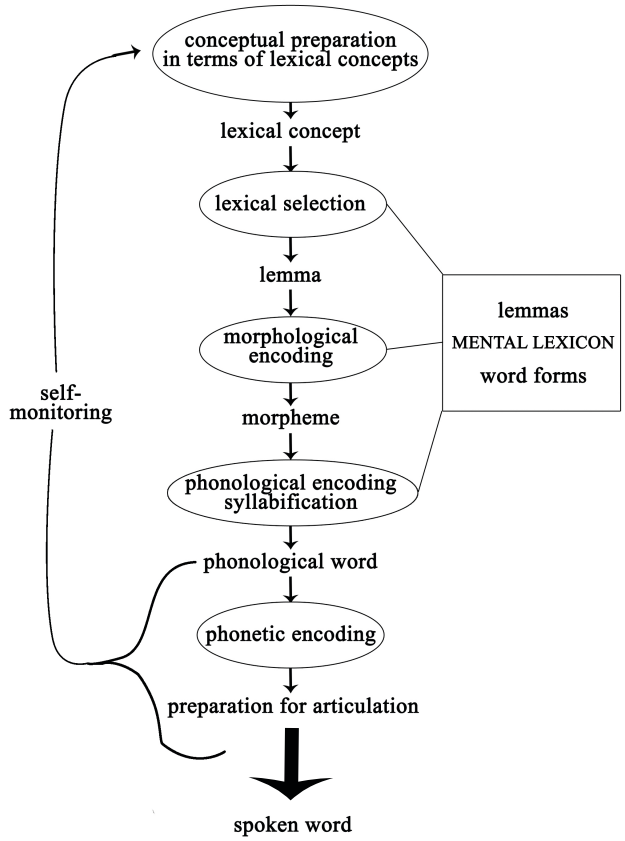

Figure 1.2: Levelt et al.'s (1999) word production model (adapted from Levelt et al., 1999).

Each stage is dependent upon the output of the previous stage (e.g., the output of the concept preparation stage is a lexical concept which is needed in the lexical selection stage to find and to select the corresponding lemma). This means that a stage only starts operating after the previous stage has provided an output to be processed further. Within each stage, there can be spreading of activation. For example, if the picture of a 'horse' is to be named, the lexical concept of a horse will activate the 'horse' lemma at the lexical selection stage, but possibly also the 'donkey,' 'cow' and 'zebra' lemmas might be activated due to semantic similarity.

Lexical selection, selecting a lexical entry given a concept, is modelled as being competitive in this model. The lemmas 'horse,' 'donkey,' 'cow' and 'zebra' all compete for selection. In the end, the lemma with the highest activation, 'horse,' wins. The time and ease it takes to retrieve a certain lemma is dependent on the 'Luce-ratio' (Luce, 1959, as cited in Levelt et al., 1999) which describes a lemma's activation, relative to the activation of all other active lemmas at that moment.

Once the 'horse' lemma is selected, it activates its corresponding '/hɔːs/' lexeme. The activation of 'donkey,' 'cow' and 'zebra' lemmas is not thought to spread to their corresponding lexemes as they were not selected as the output in the previous stage. Furthermore, the model does not postulate feedback between levels and therefore processing at the lexeme level will never affect processing at the lemma level. Instead, Levelt et al. (1999) argue that there is a self-monitoring mechanism which could correct mistakes before a word that was not intended is articulated (see Figure 1.2).

In the interactive activation network model as proposed originally by Dell (1986), activation can spread through the network freely. In this model (see Figure 1.3), after a lemma is selected, non-selected alternatives can still activate phonological forms (i.e., the lexeme; note that in Dell's model the lemma is connected to phonological segments, not a phonological word form).

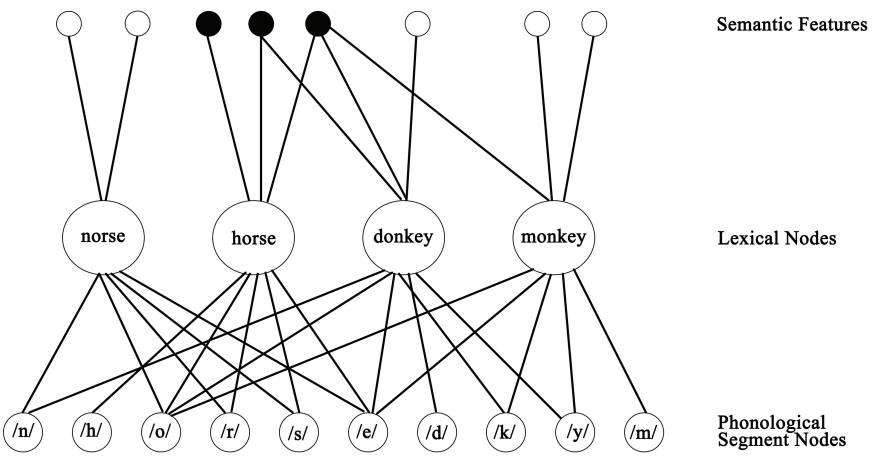

Figure 1.3: Dell's (1986) word production model (adapted from Dell & O'Seaghdha, 1992).

Using the earlier example, if the picture of a 'horse' is to be named and 'donkey,' 'cow' and 'zebra' receive activation because of semantic similarities then their activation can flow through to the phonological level. Additionally, activation can feed back from the phonological level via the lemma to the semantic level, theoretically allowing the phonological segments of the word 'horse' to activate the concept of 'norse', because of their overlap. However, the target word will be the overall output due to its receiving the greatest activation and thus ultimately winning over its competitors.

Given the different assumptions of the two models (discrete processing vs. interactive processing), the models face different challenges. Models that allow activation to spread freely are at odds with the fact that speech production is mostly accurate. On the other hand, the simple retrieval of lexemes given a target lemma as proposed by Levelt et al. (1999) cannot explain the observed preponderance of form plus meaning-based errors (e.g., "curtain – cushion") because in their model there is no feedback from the form to the concept level.

1.6 Models of Bilingual Language Production

Models of monolingual language production are usually used as a scaffold to model bilingual language production (e.g., Costa, 2005; Finkbeiner, Gollan, & Caramazza, 2006; La Heij, 2005). There are two associated challenges any model of bilingual language production faces: How is it possible that bilingual speakers can use their languages in a seemingly monolingual way while they are also able to switch and mix between languages at the same time?

In the past, there have been two major proposals as to how words were accessed in the bilingual mental lexicon for word production, each satisfying one of the challenges but not both, at least not without further modification (see Kroll & Sunderman, 2003). The first proposal assumes language-specific lexical access in a separate lexicon, thus restricting lexical access to the target language which is stored separately from the non-target language. This approach can explain why bilinguals are able to use one language without any interference from the other; however, it falls short when attempting to explain

instances in which both languages are used, for example, when code-switching erroneously or on purpose (e.g., Poulisse, 2000; Poulisse & Bongearts, 1994). On the other hand, the second proposal assumes language non-selective access in an integrated lexicon, and so does not restrict lexical access to the target language, which is stored along with the non-target language. This approach can explain instances in which both languages are used, but it needs extra mechanisms to explain how languages are kept apart in such an integrated non-restrictive setting (e.g., Green, 1998). Later, however, the proposal was modified to acknowledge that representation and access are in fact independent of each other. This allows for the possibility that there is language non-specific parallel access of two separate stores or language selective access of an integrated store (see Sunderman & Kroll, 2006).

To test these different conceptualisations of bilingual lexical access, most studies of bilingual language production have focused on whether the non-target language is activated and interacts with the target language during production. Most of these studies have used either a picture-word interference paradigm (e.g., Costa & Caramazza, 1999; Costa, Colomé, Gómez, & Sebastián-Gallés, 2003; Costa, Miozzo, & Caramazza, 1999; Hermans, Bongaerts, De Bot, & Schreuder, 1998) or a picture-picture interference paradigm (Colomé & Miozzo, 2010). For example, Colomé and Miozzo (2010) used a picture-picture interference paradigm in which Catalan-Spanish bilinguals had to name pictures in Catalan (e.g., ARMILLA [vest]), which were accompanied by a distractor picture with a Spanish name that was either phonologically related (e.g., ardilla [squirell]) or unrelated (e.g., pico [beak]) to the target. It was found that Catalan pictures were named faster when the distractor picture was phonologically related in Spanish (e.g., ardilla – ARMILLA). The results were interpreted as evidence for cross-language activation and interaction in a seemingly monolingual task; even when a bilingual uses only one of his or her languages, the other language is nonetheless active and can influence the production of the target language.

Although previous studies have come to similar conclusions, as highlighted by Colomé and Miozzo (2010), it is possible that the non-target language was activated as a result of the experimental paradigm (i.e., using auditory distracters sounding very similar to the

translation equivalent of the target) and results need to be interpreted with caution if cognates were used (for a discussion on why cognates might be problematic, see Colomé & Miozzo, 2010; see also Costa, La Heij, & Navarrete, 2006). Nonetheless, the general conclusion is that bilingual lexical access in production is characterised by the activation of both languages, even the non-target one, and that both languages can interact during the production process.

Given the assumption that the organisation in the bilingual mental lexicon might change over time (e.g., Hulstijn, 1997; Kroll et al., 2010; Paradis, 2004; Weinreich, 1953), the question remains as to how this impacts on the process of lexical access in bilinguals. It has indeed been argued that the way a target language is ultimately selected and the way a target word is not only influenced by other words in the target, but also in the non-target language, is very much dependent on the organisation of the mental lexicon which is itself dependent on features such as proficiency, experience, context of acquisition and context of use (see Kroll, Bobb, & Wodniecka, 2006). Both experiments in Chapter 3 and Chapter 4 will look at this issue more closely.

1.7 Bilingual Differences

Studying bilinguals as a group is complicated because bilinguals are very heterogeneous. A common distinction that is made within the bilingual community is between early and late bilinguals. Early and late bilinguals differ with respect to their age of L2 acquisition. Bilinguals that have acquired their L2 during their childhood years are called early bilinguals, while late bilinguals are bilinguals that have acquired their L2 around or after puberty (see Piske, Flege, MacKay, & Meador, 2002). Age of acquisition is apparently a good predictor of L2 proficiency, with early bilinguals usually being more proficient than late bilinguals (Genesee, 2009), but it is not the only predictor (e.g., Birdsong & Paik, 2008).

Two fundamental differences between early and late bilinguals are the extent of the influence of the L1, and the fact that some aspects of a L2 are difficult to acquire after a certain age. Research has shown that age of acquisition is a solid predictor for the

acquisition of certain language aspects such as correct vowel pronunciation (Piske et al., 2002) or correct use of grammatical aspects such as grammatical gender (Blom, Polisenská, & Weerman, 2008). Additionally, the more mature L1 has a greater influence on L2 acquisition in late bilinguals where it serves as a blueprint for additional acquisition more often than in early bilinguals (see Jiang, 2000; Singleton, 1999).

The L1 influence has already been emphasised in hypotheses regarding the bilingual mental lexicon and should be kept in mind when studying bilinguals. If early and late bilinguals show, to some extent, different organisation and processing of their two languages, they should be examined separately. In this thesis, the goal is to characterise in more detail late bilinguals' changes over time in the representation and processing of a L2 in the face of an existing L1.

1.8 Summary and Aims

Over the last few decades it has been established that bilingualism is rather the rule than the exception, sparking interest in bilingual word production. Since every monolingual has the potential to become a bilingual, monolingual models of word representation and production such as the models by Levelt et al. (1999) and Dell (1986) have been adapted to the bilingual case. There are a number of possibilities regarding the representation and processing of two instead of one language, and therefore a number of different models have been proposed. The development and processing of L2 representations might be different for early and late bilinguals and may also change over time as proficiency increases. The aim of this thesis is to investigate the development and processing of L2 representations in late bilinguals in more detail. Specifically, this thesis aims to investigate whether late bilinguals store their two languages in functionally separate or shared lexicons, how processing for production is characterised, and whether late bilinguals also show an advantage with regards to cognitive control.

Chapter 2 sets the stage with an exploration of proficiency tests with the aim of finding a test that reliably divides a group of late bilinguals into low and high proficient bilinguals in order to later investigate more closely the development of L2 representations for late

bilinguals.

Chapter 3 presents an investigation into the development of L2 lexical representations for low and high proficient late bilinguals. A method devised to study the development of orthographic form representations in monolingual word recognition is used to examine whether L2 lexical entries are functionally connected to or separated from L1 lexical entries.

Lastly, Chapter 4 examines whether lexical selection in late bilinguals is competitive or non-competitive and whether late bilinguals show an advantage in general and language-specific cognitive inhibition as a result of having to keep their languages apart. Chapter 4 also seeks to confirm the findings from Chapter 3.

Chapter 2

Describing the Late Bilingual Sample

2.1 Introduction

Before the development, organisation and processing of L2 lexical representations in late bilinguals can be investigated there are two aspects that need to be examined in more detail in order to do so. The first aspect concerns specific word knowledge and the second concerns the proficiency of the participants.

2.1.1 Specific Word Knowledge

The experiments in Chapter 3 and Chapter 4 rely on cross-language neighbourhood statistics for a set of English words. Both chapters examine how these English words are organised and processed given their neighbourhood density in the L2 (English) and the L1 (German or Icelandic) of the bilingual participants. In order to study how these words are organised and processed, it was important to first make sure the participants had already acquired them.

A simple task investigating word knowledge is the lexical decision task in which participants have to decide whether a shown string of letters is a word or a non-word. If the bilingual participants know the words used throughout experiments in this thesis, they should be able to recognise these words correctly in a lexical decision task. In fact, they should be able to be as accurate as monolingual participants. Distinguishing between words and non-words on the other hand should be more difficult for bilingual participants

(see Lemhöfer & Dijkstra, 2004, for possible explanations). So, while they should recognise the words as accurately as monolinguals, in accordance with previous findings (Lemhöfer & Dijkstra, 2004; Lemhöfer & Radach, 2009), bilinguals should classify non-words less accurately than monolinguals.

2.1.2 Proficiency

As highlighted in the introduction, there are a few differences between bilinguals thought to impact L2 organisation and processing. One factor described in the introduction is the age of L2 acquisition; classifying bilinguals as early or as late bilinguals. Another factor that appears to be related to L2 organisation and processing is proficiency (see Costa & Santesteban, 2004b; Potter, So, Von Eckhard, & Feldman, 1984). Therefore, it was important to measure proficiency of the late bilinguals participating in this study. This allowed for the examination of the relationship between proficiency and the organisation and processing of L2 representation; that is, what organisation/processing is associated with better proficiency, and ensures comparability of the results from this thesis to other studies in the future.

Unfortunately, proficiency is a problematic construct because there is no agreement on a working definition of proficiency (see Llurda, 2000). In fact, many studies looking at differences between bilinguals with regards to their proficiency do not define what constitutes proficiency in their study (i.e., Costa & Santesteban, 2004b; Lemhöfer & Broersma, 2011; Van Heuven et al., 1998).

In an attempt to give a broad definition, Thomas (1994, p. 330) defined proficiency as "a person's overall ability to perform in a L2". On the other hand, Stern (1985, as cited in Llurda, 2000), provides a more comprehensive definition of the term and acknowledges the progressive character of proficiency (e.g., progression towards a near native use of a L2). He states that "proficiency is the actual performance of a given learner or a group of learners" and adds that proficiency involves first, "the intuitive mastery of the forms of the languages"; second, "the intuitive mastery of the linguistic, cognitive, affective and socio-cultural meaning expressed by language forms"; third, "the capacity to use the language with maximum attention to communication and minimum attention to form"; and fourth,

"the creativity of language use" (Stern, 1985, as cited in Llurda, 2000, p. 89).

In subsequent chapters, studies will investigate the development of lexical representations, that is, how and how well acquired words are stored. This relates to the first aspect above: the mastery of form. Also, this research looks at word production processes, that is, how and how well acquired words are used. This relates to Stern's third aspect in how far is there attention to communication and not to form or how automatically is language used. Therefore, Stern's (1985, as cited in Llurda, 2000) definition fits well with the aims of inquiry of this thesis.

The objective of this chapter, then, is to find a solid measure of these two aspects. Unfortunately, not only do many studies not provide a working definition of proficiency, there is also no agreement on how to measure proficiency or aspects of proficiency as highlighted above (see Hulstijn, 2012; Lemhöfer & Broersma, 2011; Thomas, 1994). Arguably as a result of not providing a definition, many studies rely on non-standardised measures of proficiency such as participants' self-ratings or assessment (see Hulstijn, 2012; Lemhöfer & Broersma, 2011, for reviews). This makes it difficult to compare results across studies and also leads to the question of how valid self-ratings and/or assessment are in assessing proficiency. In the past, it has been shown that self-assessment of L2 reading ability was not a good predictor of actual L2 reading performance (Brantmeier, 2006), while others have shown that overall self-ratings could predict overall L2 proficiency (Lemhöfer & Broersma, 2011).

While relying on self-ratings or self-assessments for proficiency is not ideal, the fact that these measures are widely used is, however, not surprising given the fact that most current standardised proficiency tests are not made for psycholinguistic research. Existing tests such as the Test of English as a Foreign Language (TOEFL; 2003), the International English Language Testing System (IELTS) or the Quick Placement Test (QPT; 2001) were created to assess English proficiency for the purpose of gaining university entry in an English speaking country or placing students in classes at the appropriate level. These tests are exhaustive, expensive and, with exception of the QPT, not available to be conducted by researchers. Additionally, these tests take at least two hours to complete, which is not ideal in a research situation.

As a result, Lemhöfer and Broersma (2011) developed the Lexical Test for Advanced Learners of English (LexTALE), a test of vocabulary knowledge that is also a fairly good predictor of overall proficiency (r =.63). The LexTALE was specifically designed for psycholinguistic research. The researchers took into account that there is a need for a simple, inexpensive and short but effective test of proficiency. The LexTALE is a 60-item, untimed lexical decision task that takes most participants around five minutes to complete. It is available for free on the Internet. The main aim of this test is to measure vocabulary knowledge in late bilinguals. This was tested against translation performance of normed items that varied in difficulty, resulting in a high correlation between those two measures, r =.75. Additionally, LexTALE scores also correlated highly (r =.63) with performance on the QPT, a test of general proficiency (Lemhöfer & Broersma, 2011).

While the LexTALE should be a good indicator of the first proficiency aspect highlighted above, the mastery of forms, as it is a test for vocabulary knowledge, it is not clear to what extent it can also provide a good measure for the second aspect, how well the acquired words are used in communication. The LexTALE was able to predict general proficiency fairly well (Lemhöfer & Broersma, 2011), but it does not explicitly tap into how intuitively and automatically the second language is used. This is an important aspect within this thesis because the following experiments examine how and how well L2 word forms are represented in late bilinguals and how fast and automatic word forms are retrieved for production. Thus, it was decided to examine how well the LexTALE converges with a reliable measure of automatic word form retrieval.

The Test of Reading Word Efficiency (TOWRE; Torgensen, Wagner, & Rashotte, 1997) is a timed test of vocabulary knowledge and phonemic awareness. Phonemic awareness is the ability to recognise and manipulate phonemes, the smallest units of speech. Originally designed to measure these two aspects to describe reading ability in children, the TOWRE can also be used to measure these aspects in adult (non-native) English speakers (Torgensen et al., 1997). Within the TOWRE participants have to read aloud as many words or non-words as they can within 45 seconds. The timed nature of the task allows it to measure how fast and automatic a participant can retrieve words or sound out non-words. It is a standardised, easy to use test that has demonstrated good internal

consistency, reliability and validity in the past (all α's and rs $> .8$; Tanna, 2009; Torgensen et al., 1997). While the construct of vocabulary knowledge mostly relates to the first proficiency aspect within this thesis, just like the LexTALE, the construct of phonemic awareness relates to the second aspect. In the past, phonemic awareness has been linked to oral proficiency in bilinguals, with more proficient bilinguals showing greater phonemic awareness using their L2 (Sparks, Patton, Ganschow, Humbach, & Javorsky, 2006). Also, phonemic awareness has been linked to rapid automatised naming, the ability to retrieve and name stimuli, with greater phonemic awareness being linked to faster naming (Furnes & Samuelsson, 2011). Therefore the TOWRE seems to be an appropriate measure to compare with the LexTALE.

If the LexTALE converges well with automatic retrieval and oral proficiency, its scores should correlate significantly with the part of the TOWRE that measures phonemic awareness (i.e., phonemic decoding). Additionally, the part of the TOWRE measuring vocabulary size (i.e., word reading) should also correlate significantly with scores on the LexTALE, especially since both set out to measure vocabulary knowledge.

Additionally, self-ratings were also collected. If self-ratings are after all a good predictor of second language proficiency, they should also correlate significantly with scores on the LexTALE and the TOWRE.

If the LexTALE proves to be a good measure of the two aspects of proficiency important within this thesis, then the participants' proficiency could be described in terms of LexTALE scores.

2.2 Experiment 1

2.2.1 Method

Participants

German participants Thirty-three native speakers of German (14 female) participated in this study. The participants' ages ranged between 23 and 31 years with a mean age of 26.1 ($SD = 1.78$) years. All of the German participants were currently students at a

German University and were monolingual prior to English lessons at school at a mean age of 10.6 ($SD = 0.59$), which is typical for a group like this. English is a mandatory subject in all German schools starting at grade 5, when students are 10 or 11 years old. All participants were used to reading English texts and engaging with English media frequently (at least once a week) during their studies. Seventy-two percent of the participants indicated that they spoke a third language (French or Spanish), which they had acquired in secondary school. However, they also indicated that they were not very proficient in that third language and rarely used it (not more than once or twice a year).

Icelandic participants Along with the German Participants, 15 native speakers of Icelandic (10 female) participated in this study. The participants' ages ranged between 21 and 44 years with a mean age of 31.4 ($SD = 7.87$) years. All of the Icelandic participants were currently students at the University of Akureyri, Iceland where data collection took place. Similar to the German participants, the Icelandic participants were monolingual prior to English lessons at school at a mean age of 10.6 ($SD = 1.88$). English is also a mandatory subject in all Icelandic schools starting at grade 5 when students are aged between 10 and 11. All Icelandic participants were also used to reading English texts and frequently engaging with English media (at least once a week) during their studies and were used to watching English movies and television shows with Icelandic subtitles at home, as foreign movies are not dubbed in Iceland. Seventy-three percent of the participants indicated that they spoke a third language (Danish or German), which they had acquired in secondary school. However, like the German participants, the native Icelandic speakers also indicated that they were not very proficient in their third language and rarely used it (not more than once or twice a year).

English participants Fifteen native speakers of Australian English (8 female) also participated in this study. The participants' ages ranged between 19 and 46 years with a mean age of 29.4 ($SD = 9.32$) years. They were all current students at the University of New England, Australia, and had grown up monolingually. Four participants indicated that they had studied a second language (French or Japanese) at university, however, they also indicated not being proficient in that language beyond the basics and to have never

used it outside of their studies. All 63 participants in this study had normal or corrected to normal vision and participated voluntarily, with no incentives, with the option to withdraw at any time. All experiments within this thesis were approved by the Human Research Ethics Committee of the University of New England (HE10/190).

Materials

Demographic questionnaire All participants filled out a demographic questionnaire created for this study. The questionnaire consisted of general questions concerning the participant's age, gender, handedness, occupation, and language specific questions. For the German and Icelandic participants, language specific questions included a self-rating on L2 proficiency (e.g., How proficient do you consider yourself in English?), rated on a 5-point Likert scale ranging from 1 (native like), to 5 (not proficient at all). For the English participants, the language specific questions consisted of questions regarding possible second or third languages and their proficiency in, and frequency of, use of those languages. The questionnaire was presented in the participant's L1 (German, Icelandic or English, respectively) and can be found in Appendix A.

LexTALE The LexTALE is a standardised test of vocabulary knowledge and general second language proficiency developed by Lemhöfer and Broersma (2011). Its main target test-takers are university students who learned English as a second language in secondary school roughly at the age of 10 or 11. It consists of 60 items in total: 40 English words and 20 non-words. The unequal proportion of words and non-words is as a result of the fact that many of the used words have a low frequency and therefore might cause the impression that more non-words are presented as participants are not expected to be familiar with many of the words.

The items come from four difficulty categories in order to maximise discrimination between good and poor performances on the test. The items consist of four to 11 letter words, with a mean length of 7.3 and a frequency range of between one and 26 in a million (mean frequency = 6.4 in a million) according to the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995, as cited in Lemhöfer & Broersma, 2011). The non-words

are all orthographically legal, following English phonotactic rules. The LexTALE is scored as the percentage of items correctly identified, corrected for the unequal proportion of words and non-words. That is, the percentage of words and non-words identified as correct is averaged across to accommodate the tendency to over-classify items as words or non-words. Correct scores tend to range from around 58.75% in the 10th percentile to 91.25% in the 90th percentile. A higher percentage correct indicates greater vocabulary knowledge and proficiency in English. The percentages on the LexTALE can be translated to the proficiency levels provided by the Common European Framework (CEF; Council of Europe, 2013).

The reliability of the LexTALE ranges from .81 in Dutch-English bilinguals to a lower reliability of .67 in Korean-English speakers. As for validity, the LexTALE showed a large correlation of .75 with performance on a vocabulary knowledge task (i.e., translation task using normed items) and also a large correlation of .63 with the performance on a general proficiency task, the QTP (Lemhöfer & Broersma, 2011).

TOWRE The TOWRE (Torgensen et al., 1997) is a test of printed word and non-word reading skills. It was originally designed to test reading ability in children and adolescents aged six to 24 years. The TOWRE consists of two parts: Sight word efficiency (SWE), measuring vocabulary knowledge; and phonemic decoding efficiency (PDE), measuring phonemic awareness. SWE is measured through 104 regular words increasing in length from two to 11 letters and thereby in difficulty. In order to read as many words as possible in 45 seconds, the words have to be read as a whole (sight words), which assumes knowledge of those words in printed and spoken form. That way, performance on SWE is an indicator of vocabulary knowledge as well (Torgensen et al., 1997). PDE is measured through 104 phonotactically legal non-words that increase in length from two to 10 letters and thereby in difficulty. In order to achieve a high score on this part, the participant must have phonemic awareness, that is, knowledge of phonemes that are usually represented by combinations of letters in existing English words. On both parts, a raw score is obtained by counting the words and non-words read correctly within 45 seconds. These scores can then be converted to z-scores or grades, based on norms for children, adolescents and

young adults between six and 24 years of age. As the participants here were often older than 24, the raw scores were used as an indication of relative SWE and PDE with a higher score indicating better vocabulary knowledge and phonemic awareness, respectively. The TOWRE shows high internal consistency ranging from .86 to .98, and high test-retest reliability ranging between .82 and .97. Both the SWE and the PDE subtest also show high validity, with .85 and .89, respectively (Tanna, 2009; Torgensen et al., 1997).

Lexical decision task The lexical decision task was carried out to assess whether participants were familiar with the words used throughout the experiments in this thesis. The stimuli used in the lexical decision task consisted of 80 words and 80 non-words, each composed of four to six letters with a mean length of 4.9 (SD = .79) letters. The 80 words had mean frequency of 17.22 in a million occurrences (cf. Balota et al., 2007). More information regarding the selection of these 80 words can be found in Chapter 3, Experiment 3. Along with those 80 words, 80 non-words were created using the English Lexicon Project (Balota et al., 2007) non-word generator. Each of these 80 non-words was matched with one of the 80 target words with regards to first letter, number of letters, number of syllables and number of phonemes. Each non-word was pronounceable following English phonotactic rules. Both the 80 words and the 80 non-words can be found on the word list in Appendix B in Table B.2. In addition, 20 matching practice items were also included in the experiment but not in the analysis.

Apparatus

Participants were seated and tested individually approximately 50cm in front of a PC or a Macintosh computer running Windows XP with a 14-inch monitor connected. The DMDX display software developed by Forster and Forster (2003) was used for stimulus presentation and data collection. The stimulus presentation was synchronised to the 85 Hz screen with a refresh rate of 11.76 ms. Further a Logitech(R) Precision(TM) Gamepad was used for the Lexical Decision task.

Procedure

Before being seated in front of the computer, the participants were asked to fill out the demographic questionnaire taking as much time as they needed. The testing then began with the two paper-based parts of the TOWRE, SWE and PDE, followed by the LexTALE and the lexical decision task presented on the computer. The German and Icelandic participants completed these tasks and the tasks in the following experiments in one session with the opportunity to have a quick break in between tasks if needed.

TOWRE SWE and PDE The participants were informed that they would see some lists of words which they had to read as fast as they could. They were then shown a list of practice words which they had to then read as fast as they could to ensure they understood what they had to do. If they wanted, they could use their finger to help them keep the place. Participants were also told that they should skip a word if they did not know how to read it. After the practice list, they were instructed to read the longer 104 item list until they would be told to stop (after 45 seconds). As the participants were reading, words read correctly were counted until 45 seconds were up. The same was then repeated to for the non-words, with the instruction to read them as they would sound if they were real English words.

LexTALE For the LexTALE, the participants were then seated in front of the computer, wearing headphones to minimise interruptions due to surrounding noise. The LexTALE script (Lemhöfer & Broersma, 2011) was run using PRAAT (Broersma, 2001) and the participants were instructed to follow the instructions on the screen. The instructions informed them that they would see a string of letters, one at a time, which was either an English word using British spelling rules or a non-word. They would then have to say whether it was a word or not by clicking either the 'yes' (it is a word) or 'no' (it is not a word) button underneath the string using a computer mouse. There was no time limit and the program would tell them when they were finished.

Lexical decision task After the LexTALE the Lexical Decision task was started in DMDX and once again participants were told that they had to follow the instructions on

the screen. The instruction screen told the participants to decide, as quickly as possible while making minimal errors, whether a presented string of letters was a word or a non-word. The instruction was to press a green marked button on the Gamepad for words and a pink marked button for non-words.

After 20 practice trials, a total of 160 test trials were completed by each participant. Targets were presented in a random order. Each trial started with the presentation of a target in uppercase letters to which the participant had to respond. The target was presented for a maximum of 1500 ms. Feedback was given regarding the correctness and time of the response given. Response times were recorded in milliseconds.

2.2.2 Results and Discussion

Lexical Decision Task: Distinguishing Native from L2 Speakers?

The first aim was to confirm that the participants indeed knew the words used throughout the experiments. A lexical decision task was used to examine word recognition in the three language groups. Table 2.1 shows the descriptive statistics for word and non-word error percentage and reaction time (RT) across the three language groups.

Table 2.1: Mean Word and Non-word Error Rates (%) and RTs (ms), their Associated Standard Deviations (in parentheses) and their Ranges in the German, Icelandic and English Groups

Lexical Decision		German (n=33)	Icelandic (n=15)	English (n=15)
Word Error (%)				
	M (SD)	5.98 (3.41)	2.33 (1.88)	2.91 (3.12)
	Min	1.25	0.00	0.00
	Max	12.5	6.25	10.0
Non-word Error (%)				
	M (SD)	10.7 (5.95)	10.5 (8.08)	5.70 (6.07)
	Min	1.30	0.00	0.00
	Max	23.8	31.3	23.8
Word RT (ms)				
	M (SD)	556 (139)	603 (75)	520 (76)
	Min	400	502	431
	Max	864	761	684
Non-word RT (ms)				
	M (SD)	592 (130)	715 (100)	568 (78)
	Min	440	562	452
	Max	870	943	739

To analyse group differences with regards to error percentages, a two-way mixed

ANOVA was used to examine the following factors: Status (Word vs. Non-word) and Language Group (German vs. Icelandic vs. English). There was a significant interaction between Status and Language Group: $F(2, 60) = 3.56$, $MSE = 15.9$, $p < .05$, showing a medium effect, $\eta^2 = .062$. As shown in Figure 2.1, pairwise comparisons looking at words only revealed that the German participants made significantly more errors than both participants in the Icelandic ($M_{dif} = 3.64\%$, $p < .01$) and the English ($M_{dif} = 3.07\%$, $p < .01$) groups, while there was no difference between the English and Icelandic groups ($M_{dif} = 0.58\%$, $p = .543$).

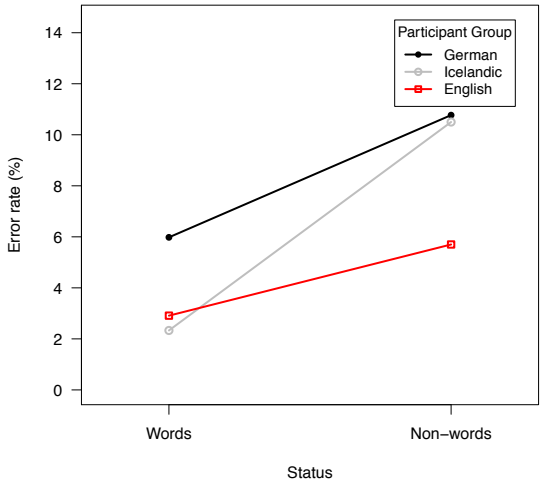

Figure 2.1: Interaction between status and language group for error rates.

Looking at the error rates for the non-words, there was no significant difference between the German and the Icelandic groups ($M_{dif} = 0.25\%$, $p = .904$). Compared to the English group, however, both the German ($M_{dif} = 5.07\%$, $p < .05$) and the Icelandic groups ($M_{dif} = 4.83\%$, $p < .05$) made significantly more errors.

From the analysis it seems as if the German participants did not know the words as well as the Icelandic participants and, more importantly, as well as the English participants. However, the difference between the groups and the overall mean error of 5.8% in the German group is relatively small. It corresponds to four to five words classified wrongly out of 80 words and does not seem to be of too much concern. With regards to the non-words, both German and Icelandic participants classified more non-words incorrectly as

words than the English participants. This difference was expected, as it has been shown in the past that bilinguals have difficulty with rejecting non-words in their L2, as illustrated by higher error rates and longer RTs (Lemhöfer & Radach, 2009). In light of this, it is interesting to note that the Icelandic participants made fewer errors for words than the Germans, but as many errors for the non-words, illustrating the difficulty of non-word rejection.

Looking at the RTs, another two-way mixed ANOVA analysing Status and Language group was conducted. Again, the interaction between Status and Language Group was significant, $F(2, 60) = 13.7$, $MSE = 1120$, $p < .001$, showing a small effect, $\eta^2 = .031$. Post-hoc comparisons revealed that, for words, the small differences in RTs between the three groups illustrated in Figure 2.2 were not significant, all $p < .05$.

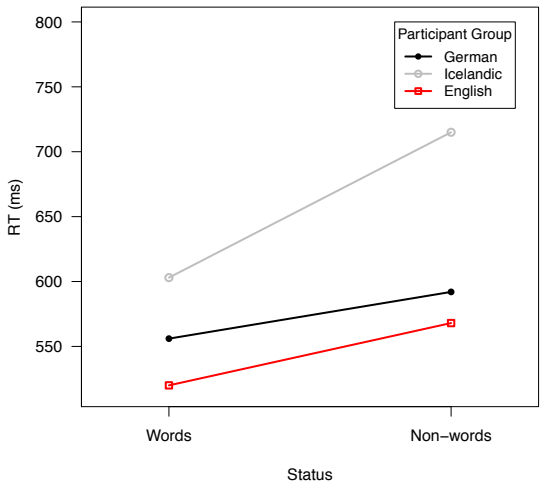

Figure 2.2: Interaction between status and language group for RTs.

However, looking at non-word RTs, the Icelandic participants responded significantly slower than English ($M_{dif} = 83.1$ms, $p < .01$) or German ($M_{dif} = 46.5$ms, $p <. 01$) participants. The difference between English and German participants was not significant ($M_{dif} = 36.5$ms, $p = .437$). Again, given previous results (Lemhöfer & Radach, 2009), longer RTs for non-words were expected in both bilingual groups compared to the monolingual group. However, this time the German participants named non-words as quickly

as the English participants. It could be the case that the German participants prioritised speed over accuracy, as indicated by native-like RTs, but significant higher error rates for both words and non-words. In contrast, the Icelandic participants only showed significantly more errors and higher RTs for non-words as expected for a late bilingual group like this (Lemhöfer & Radach, 2009). In light of the native-like recognition of words in the Icelandic group and the rather small amount of errors in the German group with regards to the words (possibly as a result of speed-accuracy-trade-off) it can be concluded that the bilinguals were familiar enough with these words for them to be included in the experimental tasks to follow.

Proficiency Measures

The LexTALE, the TOWRE (SWE and PDE), and self-ratings were used to measure L2 (English) proficiency in the German and the Icelandic groups. The aim was to assess whether the LexTALE would be a good measure of proficiency given the aspects important to the definition of proficiency within this thesis: vocabulary size and oral proficiency as indicated by phonemic awareness. Descriptive statistics and the mean differences between the groups for all the proficiency measures can be found in Table 2.2

Table 2.2: Mean Proficiency Scores of German and Icelandic Participants

Task	German (n=33)			Icelandic (n=15)			
	M (SD)	Min.	Max.	M (SD)	Min.	Max.	M Difference
LexTALE	74.6 (8.96)	57	88	78.1 (8.22)	62	92	3.48
TOWRE							
SWE	81.5 (5.70)	64	94	78.2 (9.20)	54	93	3.32
PDE	50.6 (3.62)	44	59	52.6 (4.27)	46	58	1.97
Self-Rating	2.40 (0.61)	2	4	2.60 (0.74)	2	4	0.20

Combining the German and Icelandic proficiency groups Unfortunately, one problem common to studies including bilingual participants are low sample sizes (e.g., Bialystok, Craik, Klein, & Visnawanathan, 2004; Costa, Colomé, & Caramazza, 2000; Cruz Martin, Macizo, & Bajo, 2010; Lemhöfer & Radach, 2009). While the German participant group just exceeds the recommended minimum sample size of 30 (e.g., Field, Miles, & Field, 2012, p. 43), the Icelandic group only included 15 participants. Using

G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009), it was determined that with the current sample size of $n = 15$ in the Icelandic sample and an $\alpha = .05$, the power in subsequent correlational analyses lies at .50 when aiming to detect large effects (r =.50), which gives only a 50% chance to reject a false null hypothesis. In order to detect a large effect (r =.50) with a recommended power of .80 (Cohen, 1992), at least 29 participants would be required. It was therefore considered best to combine the two language groups for further analyses.

The original rationale behind using German and Icelandic participants was to seek comparable results in two distinct, but similar bilingual groups to confirm our hypotheses. From a linguistic viewpoint German and Icelandic both belong to the Germanic language family, like English, and are equally related to the English language. Also, both languages have a shallow orthographic depth (relatively consistent phoneme-grapheme correspondence) compared to English, which has a deep orthographic depth (less consistent phoneme -grapheme correspondence; Seymour, Aro, & Erskine, 2003). Therefore German and Icelandic speakers start learning English under similar L1 conditions and are good participants to compare across.

On the other hand, measures of L2 proficiency should be able to generalise proficiency from L2 speakers with different kinds of language backgrounds. For example, (Lemhöfer & Broersma, 2011) used both Dutch-English and Korean-English bilinguals to validate the LexTALE. Thus, the different language background of the German and Icelandic speakers should not matter in this case, and the need for adequate power does prevail. After all, combining the two samples ($N = 48$) would provide a power of .96 for large effects.

While the L1 background should not matter for measuring L2 proficiency, the two samples should be comparable in terms of other participant characteristics (e.g., age, education level, etc.), as well as overall test scores to avoid the inclusion of confounding variables. Thus, group differences on important participant characteristics including proficiency measures were examined. With regards to their characteristics, two t-tests were used to compare age and age of L2 acquisition. There was a significant difference with regards to age; the Icelandic participants were on average 4.6 years older than the German participants, $t(14.6) = 2.24$, $p < .05$. The groups did not differ with regards to age of L2

acquisition, $t < 1$. As covered in the Method section, participants in both groups were undergraduate university students who grew up monolingually before learning English as a mandatory subject in school around the same age (10.6 years) and were roughly similarly exposed to English. With regards to the proficiency measures, and as expected given the participants' history with their L2 English, four independent t-tests revealed no significant differences between the groups regarding their LexTALE scores, $t(46) = 1.27$, $p = .39$, $\eta^2 = .03$, TOWRE SWE scores, $t(46) = -1.41$, $p = .16$, $\eta^2 = .04$, TOWRE PDE scores, $t(46) = 1.67$, $p = .10$, $\eta^2 = .05$, or self-ratings ($t < 1$). With regards to their characteristics, the two groups seem very comparable. As highlighted above, German and Icelandic participants differed with regards to their performance on the lexical decision task. However, as these differences seem to be the result of a speed-accuracy-trade-off in the German group and as both groups performed similarly on non-word recognition – typical for bilingual participants – this performance difference should not be an issue. Therefore, it was decided to combine the two groups to evaluate the proficiency measures closer.

Examination of the proficiency measures The relationship between the four proficiency measures (LexTALE, TOWRE SWE, TOWRE PDE, and Self-Rating) were analysed using six Pearson's product-movement correlations. Self-ratings were reversed, so that the higher the rating, the more proficient the participant. The results can be found in Table 2.3.

Table 2.3: Correlation Matrix for the Performance on the Proficiency Measures in Both Second Language Groups Combined (n=48)

	TOWRE SWE	TOWRE PDE	Self-rating
1) LexTALE	.376*	.468**	.535**
2) TOWRE SWE		.483**	.317*
3) TOWRE PDE			.324*

Note. *$p < .05$. **$p < .01$.

As expected, and shown in Table 2.3, the higher the LexTALE score, the higher the TOWRE SWE and PDE scores, as seen in significant medium correlations between the variables; LexTALE scores could explain 13% of the variance in the TOWRE SWE scores and 22% of the variance in the TOWRE PDE scores. Both the LexTALE and the TOWRE

SWE rely on vocabulary knowledge, so even though the LexTALE is not speeded while the SWE is, the shared variance is not surprising. The finding confirms the hypothesis that given their reliance on the same construct, vocabulary knowledge, the LexTALE and the TOWRE SWE should show a significant relationship.

On the other hand, the PDE relies on phonemic awareness which is not tested in the LexTALE. Also, the PDE is speeded while the LexTALE is not. Nonetheless the LexTALE and the PDE share 22% variance. This could highlight the fact that the LexTALE is good at predicting other aspects of proficiency, such as automatic retrieval and oral proficiency as indicated by better phonemic awareness (see Sparks et al., 2006). This would confirm the hypothesis that the LexTALE is a good all-round proficiency measure as needed for this thesis.

Interestingly, the LexTALE, and the TOWRE SWE and PDE all correlated with the self-ratings significantly. The higher the self-rating, the higher the LexTALE score, with a significant positive large correlation: 28% of the variance in the LexTALE scores was explained by self-ratings. The correlations between self-ratings and the TOWRE scores were also significant and positive, but only medium in size: a total of 10% of the variance in the TOWRE SWE and in the TOWRE PDE scores could be explained by self-ratings. This is in line with previous findings that self-ratings can predict proficiency (see Lemhöfer & Broersma, 2011).

Proficiency in this Sample From the correlations above it appears that the LexTALE is an appropriate proficiency measure for this study. In the combined sample, the mean LexTALE score was 75.7% (SD = 8.79), which translates to the upper intermediate language user level on the CEF (Council of Europe, 2013). According to the CEF, an upper intermediate user "can use English effectively, with some fluency, in a range of contexts" (Council of Europe, 2013, p. 2). The lowest score was 57%, which describes a lower intermediate user who "can communicate essential points and ideas in familiar contexts" (Council of Europe, 2013, p. 2). The highest score was 92.6%, which describes an upper proficient user, who "can use English very fluently, precisely and sensitively in most contexts" (Council of Europe, 2013, p. 2).

As one aim is to compare results within experiments between lower and higher proficient bilinguals to examine the impact of proficiency, it was further decided to split the sample into two groups. A median split was performed at a LexTALE score of 75.2%, creating two groups with each 24 participants. The low proficiency group contains lower and upper intermediate users, while the high proficiency group contains the highest scoring upper intermediate users as well as all the upper proficient users.

2.3 Summary

In this chapter the familiarity of the participants with the set of words used throughout the experiments in this thesis was assessed. This was important because RT and accuracy with respect to these words is analysed later on in terms of the influence of their word neighbourhood. If a word is not known, it is not part of a neighbourhood yet and its RT or accuracy does not say much about neighbourhood influence (see Vitevitch, 2002). Fortunately, the bilingual participants seemed to be as familiar with the words used as monolingual speakers as illustrated in low error rates on lexical decision. Only the German participants made significantly more errors than the Icelandic and English participants, however, their overall error rate only translated to around five words out of 80. Furthermore, they also made more mistakes classifying words as words which emphasises the possibility of a speed-accuracy-trade-off. Therefore it is assumed that the bilinguals know the used words well enough for them to be used in later experiments.

Second, this chapter aimed to investigate whether the LexTALE would be a good measure of proficiency in this thesis even though it does not explicitly measure automaticity of retrieval or oral proficiency. The LexTALE correlated significantly with phonemic awareness which is a good predictor of rapid automatised naming (Furnes & Samuelsson, 2011) and oral proficiency (Sparks et al., 2006). Additionally, the LexTALE also correlated significantly with a measure of vocabulary knowledge, confirming its validity as a test of vocabulary knowledge. Also, it was found that self-ratings are a good predictor of proficiency. In the past, self-ratings and self-assessment have been shown to predict (e.g., Lemhöfer & Broersma, 2011) and to not predict (e.g., Brantmeier, 2006) proficiency

and performance. The results in this study support the former findings. Given that the LexTALE did correlate significantly with both of the important aspects of proficiency (i.e., vocabulary knowledge and oral proficiency), this test was selected as the primary proficiency measure throughout this thesis. This was also done on the basis that this is a simple, quick and, most of all, standardised test for late bilinguals, fitting the needs of psycholinguistic research.

In order to evaluate the LexTALE, both the German and Icelandic data were combined. This was done to increase power after the power in the Icelandic group comprising only 15 participants was found to be very low. Given that a proficiency test should be able to generalise across language backgrounds, and given the fact that the German and the Icelandic were very similar with regards to their English background, combining the two samples was deemed appropriate.

Subsequently, the overall sample was characterised in terms of its proficiency using the LexTALE, showing that, on average, the participants were upper intermediate users of English according to the CEF (Council of Europe, 2013). The sample was then divided into a high and a low proficient group for further comparison in the upcoming experiments.

In sum, in this chapter the notion that most psycholinguistic studies looking at proficiency as a factor do not define or measure proficiency accordingly was considered. Consequently, a definition of proficiency and a solid measure of proficiency, the LexTALE, were sought. As a result, within this thesis, participants can be compared confidently. Also, the results regarding proficiency can be placed into the context of other studies, providing a similar proficiency classification.

Chapter 3

Representation of L2 Phonological Word Forms

3.1 Introduction

The aim of this study was to investigate how L2 lexical entries are represented within the developing lexicon of late bilinguals. Specifically, the aim was to determine the extent to which the representation of L1 phonological word forms influences the means by which L2 phonological word forms are represented. A masked form priming picture naming task was used for this purpose.

3.1.1 Cross-Language Influence in Word Production

Many studies of bilingual language production have investigated whether the non-target language is activated and interacts with the target language during production. The bulk of the evidence points towards the simultaneous activation of the bilinguals' two languages in production tasks. Most of these studies used either a picture-word interference paradigm (e.g., Costa & Caramazza, 1999; Costa et al., 2003, 1999; Hermans et al., 1998) or a picture-picture interference paradigm (Colomé & Miozzo, 2010). For example, Colomé and Miozzo (2010) used a picture-picture interference paradigm in which Catalan-Spanish bilinguals had to name pictures in Catalan (e.g., ARMILLA [vest]), which were accompanied by a distractor picture that can take a Spanish name that happens to be

phonologically similar (e.g., ardilla [squirrel]) or unrelated (e.g., pico [beak]) to the target. It was found that Catalan pictures were named faster when the distractor picture was phonologically related in Spanish (e.g., ardilla – ARMILLA). These and similar results using non-target language in seemingly monolingual production tasks have been interpreted as evidence for cross-language activation and interaction. Even when a bilingual uses only one of his or her languages, the other language can nonetheless influence the speed with which words are produced in the target language. Colomé and Miozzo (2010) suggest that the non-target language may have been activated as a result of the experimental paradigm (e.g., using auditory distractors sounding very similar to the translation equivalent of the target), and results need to be interpreted with caution, most patently if cognates were used. The problem with using non-target language distractors that are in any way related to the target word is that while these distractors might indeed cause cross-language activation, they may not reflect the natural processing of words. So although it seems that a non-target language can influence target language production, this might not be happening in the absence of any distractors or in the absence of any resemblance of the word to be said to the non-target language.

This problem can be compounded if cognates are present in the experimental paradigm. Cognates are translation equivalents that are phonologically similar; for example, dog in German Hund, /hʊnt/ and in Dutch hond, /hɔnt/. It may be that cognates are represented differently than non-cognates. Cognates share their morphological stem, they are learned faster and easier, and they are used more often (their morphological stem will be activated both when using the German and the Dutch word for dog), so their representations tend to be more robust (see Colomé & Miozzo, 2010; see also Costa et al., 2006). Therefore care should be taken when performance on cognates is compared with performance on non-cognates.

When investigating whether lexical entries in two languages are functionally separated during language production, it is thus essential to ensure that no obvious cross-language cues are used, such as in the case of cognates. One possibility is to look at the effect of neighbourhood in the domain of word recognition. Van Heuven et al. (1998) investigated the influence of the L1 neighbours of L2 items in a lexical decision task using late bilin-

guals. They found that in a task completely conducted in the participants' L2, not only an L2 word's L2 neighbours, but also its L1 neighbours influenced decision times, thus highlighting the influence of lexical items in the non-target language. This points towards lexicons being at least functionally connected during word recognition. In production, on the other hand, the questions about the methodology raised above mean we do not yet have clear evidence about possible functional connections between L1 and L2 lexical entries and for whether bilinguals have one or two lexicons. Therefore, the objective of this study was to investigate the development of L2 lexical phonological forms with respect to their L1 forms in bilinguals. More specifically, we want to know whether L2 lexical phonological forms are represented and organised in line with L1 lexical phonological forms or are independent of the L1 forms. It is highly likely that language representation and processing differs depending on the manner and timing of acquisition of the two languages (Jiang, 2000) and, for this reason, this study will focus on late bilinguals only.

3.1.2 The Development of Lexical Entries

The experiments below employ a masked form priming paradigm with picture naming rather than the standard word naming or lexical decision since the focus of this study is on lexical access for production purposes. Masked form priming typically consists of a target (e.g., GOAT) preceded by a word or non-word prime which is either related (e.g., geat - GOAT) or unrelated (e.g., gemp - GOAT) in form to the target, and which is presented so quickly (around 50ms) that it is not processed consciously. Targets preceded by a related prime tend to be responded to faster than those preceded by an unrelated prime (i.e., positive form priming). Form priming tends to be more effective when non-words are used as form primes and when the prime is masked rather than overt. Masked form priming has been used to investigate monolingual written word recognition in adults (Forster & Davis, 1991; Forster & Taft, 1994) and the development of word recognition in children (Castles, Davis, Cavalot, & Forster, 2007). One prominent finding using this paradigm in both lexical decision and word naming tasks is the finding that target form priming depends on the target's similarity to other words in the lexicon; that is, the neighbourhood density of the target. The more neighbours a target has, the less likely

it is to be primed positively by a non-word similar in form (Castles et al., 2007; Forster & Davis, 1991; Forster & Taft, 1994; Kinoshita & Norris, 2012). The finding that words with many neighbours (i.e., high-density (HD) neighbourhoods) are not as susceptible to form priming as those with fewer neighbours, may be the result of an adaptation in a growing vocabulary, allowing the system to discriminate between very similar entries more efficiently (Castles et al., 2007; Forster & Taft, 1994). One means by which this could be achieved is that entries from HD neighbourhoods are recoded in terms of their sub-syllabic units into more fine-grained representations requiring the input to retrieve them to be more specific (e.g., 'geat' can retrieve the low-density (LD) word GOAT, but 'freg' cannot retrieve the HD word FROG). This is known as the lexical tuning hypothesis (LTH; cf. Castles et al., 2007, p. 167). According to the LTH, no form priming effects are found for HD words because since the target would have been recoded into a more fine-grained form, the prime is less likely to match the target enough to facilitate its retrieval. Whereas the LTH was proposed for visual word recognition, in the area of auditory word recognition the lexical restructuring model (LRM; Walley, Metsala, & Garlock, 2003) proposes that lexical representations become more segmentally structured as the vocabulary grows, with greater restructuring taking place in HD neighbourhoods to allow better discrimination between these items (Walley et al., 2003). Evidence for this comes from studies that link increases in vocabulary size with increased phonemic awareness on the part of children for HD words but not LD words (De Cara & Goswami, 2003; Storkel, 2002). In light of the effects neighbourhood density has on the way words are represented, we propose to investigate how lexical entries are structured in late bilinguals' mental lexicons by looking at neighbourhood density changes in their developing lexicons.

According to the LTH and the LRM, recoding/restructuring of lexical entries happens as a response to a growing lexicon/vocabulary. For example, contrary to what has been found in adults, children in grade 3 showed form priming effects for HD words in a lexical decision task. However, by grade 5, the same children did not show form priming effects for HD words anymore. The researchers argued that as these children became better readers and acquired more orthographic representations, lexical entries were recoded to allow efficient discrimination between them as predicted by the LTH. As a result, in

grade 5, HD entries were not primed anymore (Castles et al., 2007). With respect to late bilinguals, the question is what this development would imply for L2 vocabulary given bilinguals already possess a mature L1 system when they start to acquire their L2. Will the presence of the L1 influence the representation of the L2 lexical entries? After all, when a late bilingual starts to acquire L2 entries there is already an L1 lexicon that could be used as a reference (see Jiang, 2000; Singleton, 1999).

3.1.3 The Development of the Current Study

To answer whether the presence of the L1 influences the representation of L2 lexical entries from the perspective of production mechanisms, the masked form prime naming paradigm of Forster and Davis (1991) was adapted into a production task using picture targets. Although form priming effects have only been investigated in the domain of visual word recognition (Castles et al., 2007; Forster & Davis, 1991; Forster & Taft, 1994), given that restructuring is found in phonological neighbourhoods as well as orthographic ones (De Cara & Goswami, 2003; Storkel, 2002), similar effects would be expected in a picture naming task. Word production and recognition, while arguably involving some separate processes, are thought to use the same lexicon (see Levelt et al., 1999, for argument and evidence) or, if there are different lexicons, these at least show similar organisations (Costa & Santesteban, 2004a). If word recognition and production use the same lexicon or involve similar ones, the LTH and the LRM would predict similar effects in a production paradigm as have been reported for word recognition paradigms. This in turn means that studying the variables influencing these effects could aid in understanding how L2 lexical entries in late bilinguals are stored initially and how or whether this changes as the language develops. Why would the organisation of L2 words be influenced by the organisation of L1 words? Given that the L1 system has already matured in late bilinguals, it is possible that in the early stages of L2 acquisition, L2 words are integrated in the already existing system; L2 words forms could simply be treated as new L1 word forms (see Jiang, 2000, for a similar proposal). If so, the LTH would predict that at least the L2 items that have many neighbours in the L1 are recoded accordingly. In that case, L2 items that have many neighbours in the L1, regardless of their L2 neighbourhood density, should

show no form priming. However, in time, as more L2 words are acquired, neighbourhoods will start to develop with respect to these new words and the L2 neighbourhood size should then also influence recoding. At this point, priming effects will either depend on the L2 neighbourhood density or a combination of the L1 and L2 neighbourhood density, depending on the integration of the two lexicons as proficiency increases. To track this development, both high and low proficient late bilinguals were observed.

Until recently, studies investigating orthographic neighbourhoods (Castles et al., 2007; Forster & Davis, 1991; Forster & Taft, 1994; Van Heuven et al., 1998) have all employed Coltheart's N (Coltheart et al., 1977) to calculate neighbourhood densities. In production, however, the relevant form is phonology. Phonological neighbourhoods should then be defined as consisting of neighbours that differ by one phoneme (e.g., back – BIKE) through substitution, addition, or deletion. While it would have been feasible to use this neighbourhood definition to calculate cross-language neighbourhoods for this study, Yarkoni, Balota, and Yap (2008) have developed an even more effective measure, the PLD20. This new measure is based on the Levensthein distance, a measure of string similarity commonly used in computer science. In contrast to Coltheart's N, the PLD20 does not only consider direct neighbours (by substitution, addition, or deletion of a single morpheme), but also takes into account 'indirect' neighbours that differ by two or more transformations (e.g., mouse - HORSE). In effect, it provides the mean number of transformation for the 20 closest neighbours of a word. The measure is thereby more sensitive to variations in neighbourhood size, especially in LD neighbourhoods and neighbourhoods of words with many letters (Yarkoni et al., 2008). Since the PLD20 provides the mean number of transformations for the 20 closest neighbours of a word, the lower a word's PLD20 is, the more phonological neighbours it has. This is in contrast to Coltheart's N where a small N translates to a word having only a few neighbours. By using a neighbourhood density measure that is more sensitive to variations in neighbourhood size in LD neighbourhoods, since secondary or tertiary neighbours are also considered, some form priming effect might be found for HD items in this study. That is, some items that are classified as HD items in this study would have actually been LD items in previous studies because previous studies did not include secondary or tertiary neighbours (Castles et al.,

2007; Forster & Taft, 1994). Given the differences in the definition of neighbourhood density and in the measurement of neighbourhood density, it remains to be seen what constitutes LD or HD neighbourhoods according to the LTH.

The study presented here consisted of two experiments. The aim of the first experiment was to establish whether the form priming pattern across neighbourhoods reported in earlier recognition studies could be replicated in a picture naming task. For this, monolingual speakers of Australian English were used. The experiment was conducted to validate the method and to provide baseline monolingual data. The aim of the second experiment was to investigate form priming patterns in late bilinguals. Specifically, the second experiment sought to investigate whether form priming patterns emerged as a function of the L1 or the L2 neighbourhood of L2 items. Manipulating not only the L2 neighbourhood of L2 words, but also the cross-language L1 neighbourhood, allows for the observation of whether the expected form priming pattern is influenced by L1 or L2 organisation as an indication of functionally connected or separated L1 and L2 lexicons.

3.2 Experiment 2

The first point that needs to be addressed is whether neighbourhood effects in form priming can be detected for phonological neighbourhoods in a task that can be considered more representative of production, namely picture naming. Experiment 2 looks at whether the form priming patterns across neighbourhoods can be replicated in both word naming and picture naming tasks when using a measure of word similarity, the PLD20, based on phonology rather than orthography. Given the evidence provided earlier, parallel form priming effects were expected for both word and picture naming tasks, with significant priming effects for items in LD neighbourhoods and reduced or no priming for items in HD neighbourhoods.

3.2.1 Method

Participants

Fifty-nine students (49 female) of the University of New England with ages ranging from 18 to 59 years ($M = 28.5$, $SD = 11.55$) participated in this experiment. All participants were native speakers of Australian English. Twenty-one participants indicated that they spoke another language at a very basic level and not frequently (less than once a year). All participants had normal or corrected-to-normal vision. The participants received no monetary reward.

Materials

Eighty pictures from the International Picture Naming Database and their corresponding word names (available at http://crl.ucsd.edu/experiments/ipnp/) were used in this task (see Appendix B, Table B.1). Only four to six letter mono-morphemic words with 2 to 6 phonemes, and a maximum of two syllables were included. Only low frequency words tend to reliably show form priming (Andrews, 1989). Therefore only words with a frequency less than 40000 based on Lund and Burgess (1996) Hyperspace Analogue to Language (HAL) frequency norms were considered. A median split was carried out to divide the sample of words into two phonological neighbourhood density groups, based on a PLD20 of 1.50.The result was 40 targets belonging to the LD (LD words/pictures) with higher PLD20 values and 40 targets belonging to the HD group (HD words/pictures) with lower PLD20 values. An overview of the mean lexical characteristics for both LD and HD items can be found in Table 3.1.

As shown in Table 3.1, there are significant differences between items in the HD and the LD condition with regards to the number of letters, phonemes and syllables. Previous studies have shown that number of letters and phonemes do not significantly determine picture naming time (Alario et al., 2004; Barry, Morrison, & Ellis, 1997). Additionally Alario et al. (2004) found that while the number of syllables did influence naming times, this difference was only significant between words with one or two syllables and words with three or more syllables. Therefore these differences were not considered to be significant

in the context of this study.

Table 3.1: Means and Standard Deviations (in parentheses) for Word Frequency, Number of Letters, Phonemes and Syllables, and the PLD20 for LD and HD Items

	LD	HD	Difference
HAL frequency	6736 (6331)	8926 (8810)	2189 (7702)
No. of letters	5.60 (0.59)	4.32 (0.47)	1.28* (0.71)
No. of phonemes	4.60 (0.63)	3.30 (0.52)	1.30* (0.61)
No. of syllables	1.70 (0.46)	1.00 (0.00)	0.70* (0.46)
PLD 20	1.86 (0.28)	1.11 (0.14)	0.75* (0.33)

Note. *$p < .05$

Along with the 40 HD and 40 LD items, two sets of non-word primes were created. The first set of non-word primes consisted of 80 form related primes. For each target word, a related non-word form prime was created by changing one letter of the target word (e.g., slibe - SLIDE). The related primes were always created in a way that they were still pronounceable and obeyed the phonotactics of English. To avoid onset effects (see Schiller, 2008; Forster & Davis, 1991), the first letter was never changed. The result was that each related non-word form-prime matched one of the 80 words with regards to first letter, number of letters, number of syllables and number of phonemes. The second set of non-word primes consisted of 80 unrelated primes. Each unrelated prime matched one of the 80 target words only with regards to the first letter (e.g., spack - SLIDE). Unrelated non-word primes were also still pronounceable, obeyed the phonotactics of English and matched the targets with regards to the number of letters, number of syllables and number of phonemes.

Design

A 2x2 Latin square design was used to distribute items across four Prime (unrelated, related) by Task (word naming, picture naming) conditions resulting in four subject groups. This distribution was applied to each density group (HD and LD) separately. As a result, within each of the two density groups every item appeared in every condition, acting as its own control, and while each participant was exposed to all conditions, they only experienced each item once, as illustrated in Table 3.2.

Table 3.2: Illustration of the Latin Square Distribution of Items across Conditions in Experiment 2

Task	Word naming		Picture naming	
Group	Unrelated Prime	Related Prime	Unrelated Prime	Related Prime
1	curp - COMB (1)	nuil - NAIL (4)	lult - LAMP (3)	poar - PEAR (2)
2	pizt - PEAR (2)	cemb - COMB (1)	nuef - NAIL (4)	lemp - LAMP (3)
3	lult - LAMP (3)	poar - PEAR (2)	curp - COMB (1)	nuil - NAIL (4)
4	nuef - NAIL (4)	lemp - LAMP (3)	pizt - PEAR (2)	cemb - COMB (1)

Note. Illustrative item numbers are given in parentheses

Apparatus

Participants were tested individually and seated approximately 50cm in front of a PC running Windows XP with a 14-inch CRT monitor running at 85Hz (refresh rate of 11.76ms). The DMDX display software developed by Forster and Forster (2003) was used for stimulus presentation and data collection. Participants wore a headset with a microphone to record their spoken answers and to keep outside noises to a minimum. CheckVocal (Protopapas, 2007) was used to process the voice recordings, checking for accuracy and timing of responses.

Procedure

The form priming task consisted of two phases: A word naming phase and a picture naming phase. It was felt that any generalised practice effects were more likely to impact on picture naming than on the more automatic word naming. Practice effects would increase the sensitivity of the task, and so the picture naming task was always presented after word naming task. With each phase taking no more than five minutes to complete, there was no likelihood of fatigue effects. In the word naming phase of the form priming task participants were instructed to name an uppercase target word, in the picture naming phase they were instructed to name a picture. Apart from an instruction to start naming pictures, there was no break between the two phases. In both phases, after 10 practice trials, each participant completed a total of 40 test trials. Each participant thereby completed a total of 80 test trials. At the beginning of each trial, a forward mask comprised of six hashes was presented for 505ms, immediately followed by a non-word prime in lowercase letters for 47ms. Immediately after the prime, a word target in uppercase letters during the word

naming phase or a picture target during the picture naming phase appeared (ISI = 0) until the participant made a response by naming the target, activating the voice switch. If no naming response was given after 2000ms, a time out message appeared on screen, reminding the participant to respond to upcoming items faster. No other feedback was given to the participants throughout the task. The participants completed this task and the tasks in the experiments outlined in Chapter 4 in one session with the opportunity to have a quick break in between tasks if needed.

3.2.2 Results

An overview of the mean reaction times (RTs; in ms), error rates (in %) and respective standard deviations for Prime by Density conditions in both the word naming and the picture naming task can be found in Table 3.3. RTs will be analysed first, followed by the analysis of the error rates, which (throughout this and the following experiments) reflect incorrect responses (generating a wrong name). Premature responses (RTs faster than 300ms) and failures to provide a response within 2000 ms were excluded from analysis.

Reaction Times

RTs were examined to ascertain whether there were parallel form priming effects across tasks. In both tasks, a significant positive form priming effect was expected for LD items, with a smaller or absent form priming effect for HD items. A four-way mixed-design ANOVA was used to analyse the factors: Prime (unrelated vs. related), Density (HD vs. LD), Task (word naming vs. picture naming) and Group. A non-repeated factor, Group, was always added to extract variance due to counterbalancing (Latin square design). All analyses were carried out for both subjects (F_1) and items (F_2).

The analysis revealed a significant Prime by Density by Task interaction, $F_1(1, 55) = 17.4$, $MSE = 1037$, $p < .001$, $\eta^2 = .002$; $F_2(1, 78) = 4.04$, $MSE = 2094$, $p < .05$, $\eta^2 = .002$. Further analysis of the subject data using Tukey's $HSD = 18.6$ with a familywise α of .05 showed significant form priming effects for LD and HD items in the picture naming task, but not in the word naming task. Given the mean and variance differences between the tasks and the conservative nature of Tukey's HSD, RTs for each task were analysed

separately using two three-way mixed design ANOVAs.

Table 3.3: Mean RTs (ms), Error Rates (%), and Standard Deviations (in parentheses) for HD and LD Density Words Preceded by Unrelated or Related Primes in the Word or Picture Naming Task

Task	Condition	Example	RT	Error rate
Word naming	HD			
	Unrelated	curp - COMB	455 (64)	1.52 (4.00)
	Related	cemb - COMB	443 (64)	0.34 (1.82)
	Priming Effect		12 (26)	1.18
	LD			
	Unrelated	fiet - FROG	460 (64)	1.52 (3.62)
	Related	freg - FROG	443 (65)	1.18 (3.75)
	Priming Effect		17 (27)	0.34
	Difference		3	0.84
Picture naming	HD			
	Unrelated	curp - COMB	710 (90)	12.1 (10.8)
	Related	cemb - COMB	656 (92)	6.16 (6.74)
	Priming Effect		54 (73)	5.94
	LD			
	Unrelated	fiet - FROG	737 (74)	13.1 (10.3)
	Related	freg - FROG	627 (98)	6.10 (8.31)
	Priming Effect		110 (84)	7
	Difference		56	-1.06

Word naming task An analysis of word naming RTs did not reveal a significant Prime by Density interaction, $F_{1\&2} < 1$. However, a significant main effect was found for Prime, $F_1(1, 55) = 38.4$, $MSE = 341$, $p < .001$, showing a large effect $\eta^2 = .150$; $F_2(1, 78) = 13.6$, $MSE = 724$, $p < .001$, showing a large effect $\eta^2 = .147$; while the main effect for Density was not significant, $F_{1\&2} < 1$. As illustrated in Figure 3.1, there was a reliable form priming effect in the word naming task, but contrary to what was expected, this effect did not vary with density.

Picture naming task For the picture naming task, a significant Prime by Density interaction was found, $F_1(1, 55) = 21.3$, $MSE = 2061$, $p < .001$, showing a small effect $\eta^2 = .048$; $F_2(1, 78) = 6.2$, $MSE = 6911$, $p < .05$, showing a small effect $\eta^2 = .035$. Further analysis using Tukey's HSD of 22.2 with a familywise α of .05 showed significant

form priming effects for LD and HD items. As expected and illustrated in Figure 3.2, the priming effect was bigger for LD pictures than for HD pictures.

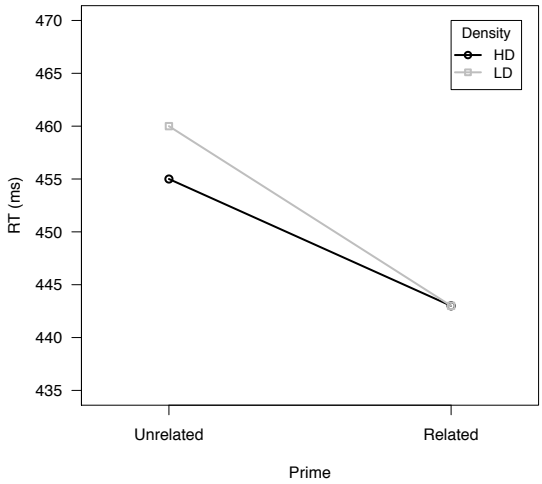

Figure 3.1: Interaction between type of prime and type of density in the word naming task.

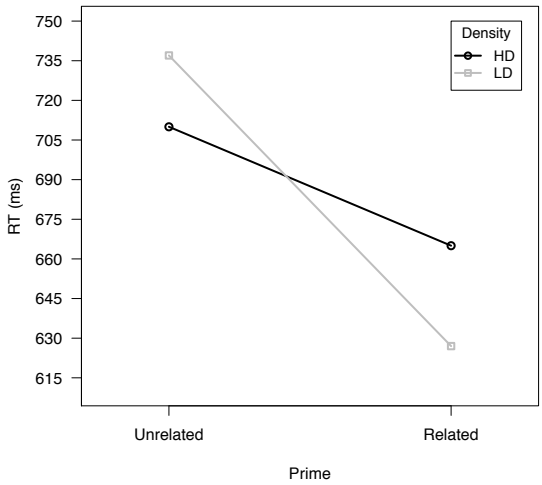

Figure 3.2: Interaction between type of prime and type of density in the picture naming task.

Error Rates

As with the RTs, error rates were analysed using a four-way mixed-design ANOVA for the factors: Prime (unrelated vs. related), Density (HD vs. LD), Task (word naming vs.

picture naming) and Group. The analysis revealed a non-significant Prime by Density by Task interaction, $F_{1\&2} < 1$. Only a significant Prime by Task interaction was revealed, $F_1(1, 55) = 23.5$, $MSE = 44.3$, $p < .001$, showing a small effect $\eta^2 = .036$; $F_2(1, 78) = 17.2$, $MSE = 170$, $p < .001$, showing a small effect $\eta^2 = .023$. Further analysis using Tukey's $HSD = 3.25$ with a familywise α of .05 showed a significant difference between error rates on unrelated vs. related Prime trials in the picture naming, but not in the word naming task. As illustrated in Figure 3.3, in contrast to the word naming task where overall fewer errors were made, in the picture naming task, the presence of a related form prime reduced the error rates significantly.

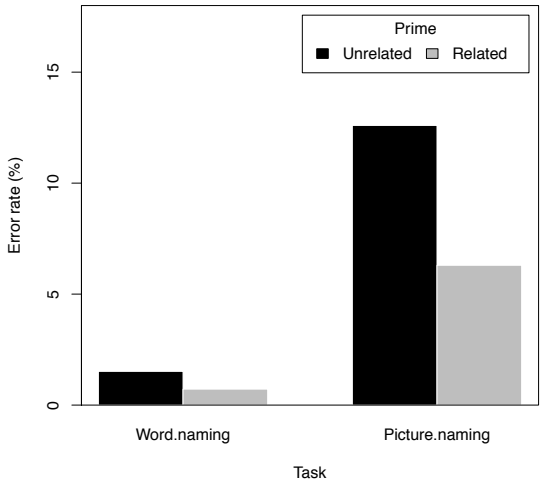

Figure 3.3: Mean error rates (%) across primes and tasks.

3.2.3 Discussion

In Experiment 2, the form priming patterns of items from LD and HD phonological neighbourhoods in a word naming and a picture naming task were investigated. The main aim was to ascertain whether previously reported differences in form priming between LD and HD items found in word naming (Forster & Davis, 1991) and lexical decision tasks (Forster & Davis, 1991; Forster & Taft, 1994) could be replicated in a picture naming task. Parallel form priming effects across both word and picture naming tasks were predicted,

with significant priming effects for items in LD neighbourhoods and reduced form priming, or even an absence of priming, for items in HD neighbourhoods. The results revealed a difference in form priming patterns across tasks. In the word naming task, both HD and LD items showed an equivalent amount of RT form priming. In the picture naming task, both HD and LD items showed form priming, with LD items showing a greater amount of form priming. While, contrary to our expectations, there were no parallel form priming effects across both tasks, the difference in priming between HD and LD items does point in the expected direction. The findings will be discussed for each task separately below.

Word Naming Task

The finding that form priming effects were present and equivalent for both HD and LD words in the word naming task is not in line with the expectations and the results in Forster and Davis (1991) study. Forster and Davis did find a difference in form priming for words from HD and LD neighbourhoods in a naming task, whereby there was no form priming effect for HD words once onset sounds were controlled for. However, in contrast to Forster and Davis (1991) and other previous studies reporting no form priming effects for HD items in the lexical decision task (Castles et al., 2007; Forster & Taft, 1994), this experiment used a different approach to classify neighbourhoods. First, neighbourhoods were calculated according to phonological neighbours in contrast to the orthographic neighbours used by Forster and Davis (1991).While orthographic and phonological neighbourhoods do overlap in most cases, not all phonological neighbours are also orthographic neighbours (e.g., bait – fate), especially in a language with a deep orthography like English. In our item set, there was a strong correlation ($r(78)$ = -.78, $p < .001$) between the PLD20 used to classify neighbourhoods in this study and the items Coltheart's orthographic N used in previous studies. However, upon inspection, it was found that a number of HD items as classified by the PLD20 would have been LD items according to Coltheart's orthographic N. Given that word naming, in contrast to picture naming, involves visual word recognition, the orthographic neighbourhood might have played a bigger role than the phonological neighbourhood, in which case the number of HD items in our set which should have been classified as LD is likely to have invalidated the density contrast and

prevented the emergence of an interaction. Furthermore, a number of LD items would have been HD items according to Coltheart's orthographic N, causing the priming effect in both groups as classified by the PLD20 to be equivalent.

Second, Forster and Taft (1994) have shown that it is not only the neighbourhood of the whole word that determines form priming, but also frequency and neighbourhood density of segments such as the body or the antibody. For example if HD words had a low frequency antibody (e.g., pru in prune), and the prime shared this antibody (e.g., prute – PRUNE), the word showed a priming effect regardless of the whole word density neighbourhood. Given the restrictive nature of using targets from an existing picture set, it was not possible to manipulate targets and primes in terms of their segmental overlap in this fashion. Since the onset of the primes and targets had to match to avoid onset effects, a few target and prime pairs shared an antibody. Therefore, it is possible that some of the HD words that were used showed priming due to shared low frequency and/or low neighbourhood density segments with the prime.

Picture Naming Task

The difference in the magnitude of priming effects between HD and LD items, with LD items showing greater form priming effects, is in line with our expectations and previous findings (Castles et al., 2007; Forster & Davis, 1991; Forster & Taft, 1994). This also confirms that a form priming effect as predicted by the LTH and the LRM can be found in a pure production task such as picture naming and validates this method for further use.

As expected, there was a reduced priming effect for HD pictures, in contrast to the absence of priming effects for HD items in previous studies. As described above, shared low frequency segments between the target and the prime could have resulted in a priming effect for these HD items. Additionally, as mentioned before, a new, less conservative neighbourhood measure was used in this experiment. In contrast to the traditional neighbourhood measure, Coltheart's N, the PLD20 takes into account neighbours with different degrees of similarity to the target. A PLD of 1.45, which was the high end (closer to the LD range) of the range for the items in the HD neighbourhood condition, translates

into the target having 11 direct neighbours by substitution, deletion or addition, and nine neighbours through two operations of substitution deletion or addition. In terms of word recognition studies, a target with 11 direct neighbours would be considered to be a HD item, however, in one word production study using phonological neighbourhoods (Vitevitch, 2002) the LD neighbourhood included items with a mean direct neighbourhood size of 14. Also, none of the previous studies looked at the influence of neighbours that differ by two or more operations (e.g., mouse – HORSE), which are also considered in the PLD20. As a result, it is not evident whether, for instance, an item with 11 direct (one operation) and nine (two or possibly more operations) indirect neighbours can be considered more or less of an HD item than an item with 14 direct and one indirect neighbour. In this sense, if not all of our HD items are considered as 'higher' HD items, a small but possibly significant priming effect would be expected for this condition as well. This could also be the case if the lexical tuning, as suggested in the LTH and the LRM, is not necessarily dichotomous, but a graded process: The degree of tuning might vary with the number of neighbours in a given neighbourhood. The more neighbours a word has, the finer the tuning in that neighbourhood, possibly determined by a threshold. The degree of tuning might then also be influenced by further characteristics of a word, such as segment frequency (Forster & Taft, 1994). As a result, a certain degree of priming could be expected in some HD items as found here. Considering these aspects and given the difference found in priming effects between HD and LD items, the results for the picture naming task do show that the same evidence for lexical tuning or restructuring as proposed for orthographic (Castles et al., 2007; Forster & Davis, 1991; Forster & Taft, 1994) and phonological neighbourhoods in recognition (Metsala & Walley, 1998), is also evident in phonological neighbourhoods in production.

Conclusion

In sum, a difference in form priming effects for HD and LD items in a picture naming task was demonstrated in accordance with the LTH showing that it is possible to determine whether picture names are associated with HD or LD neighbourhoods using the magnitude of their priming effects in the form priming task. Therefore, the picture naming task

proves to be an appropriate task to use in the second experiment where the object is to investigate the organisation of the bilingual lexicon. However, given the issues surrounding item selection mentioned above, the item set was adjusted as described in the method below.

3.3 Experiment 3

The aim of this experiment was to look at how words from a L2 get integrated in the mental lexicon. If L2 words are functionally incorporated (at least initially) in the L1 lexicon, an L1 density dependent form priming effect would be expected for L2 words. However, if this is not the case, no such effect would be expected and L2 words would show density priming effects only with respect to the L2 neighbourhood density. It is also possible that these effects are modulated by proficiency such that low proficient late bilinguals would be expected to show density dependent priming effects with respect to L1 neighbourhoods while in highly proficient speakers, L2 density should determine the magnitude of form priming effects either as well as or instead of the neighbourhood density as defined by the L1 lexicon. The degree to which priming in highly proficient bilinguals depends on the neighbourhood density of the L1 lexicon should also be an indication of how functionally independent the two lexicons of a late bilingual are. Following the findings in Experiment 2, significant form priming effects were expected for all items regardless of neighbourhood with this effect being significantly bigger for LD items than for HD items. It was also expected that this density difference would surface when density was defined by the L1 because of the integration of L2 word forms into the L1 word form stores assuming a single lexicon. Lastly, a difference in priming patterns was expected between low and high proficient late bilinguals, with a greater influence of L2 in the latter group indicating the development of L2 items within the L1 lexicon towards more independence over time.

3.3.1 Method

Participants

The same forty-eight (24 female) German (n = 33) and Icelandic (n = 15) English late bilinguals as in Experiment 1 participated in this study.

Materials

Line drawings from the coloured Snodgrass and Vanderwart set by Rossion and Poutois (2004) were used in this task. A different image set to the one used in Experiment 2 was chosen because this set contained HD and LD items that were more homogenous within their density group (see Appendix B, Table B.2). For example, in this second set, HD items did not vary as much in terms of their density and were less close to the LD items. Once again, only four to six letter words with two to six phonemes, a maximum of two syllables and one morpheme were considered. Also, as pointed out earlier, only low frequency words tend to reliably show form priming. Therefore, as in Experiment 2, only words with a frequency less than 40000 based on Lund and Burgess (1996) Hyperspace Analogue to Language (HAL) frequency norms were considered.

Initially, the intention was to analyse the effect of density in the L2 (English) only. So first, a median split was carried out to divide the sample of words in two neighbourhood density groups according to the items' English density. The median split was performed for a PLD20 of 1.50. Accordingly, 80 words were selected with 40 targets belonging to the English LD (LD pictures) and 40 targets belonging to the HD group (HD pictures). An overview of the mean lexical characteristics for both LD and HD items can be found in Table 3.4. See note in Experiment 2, p. 42, on significant differences for number of letters, phonemes and syllables. Eighty non-words were matched to the target words as in Experiment 2.

After the 80 words had been selected, it was decided to also investigate the effect of cross-language neighbourhood density. German and Icelandic PLD20s were also calculated for these 80 items. This was achieved by taking International Phonetic Alphabet (IPA) transcriptions of the 80 selected English words and calculating the Levenshtein

Table 3.4: Means and Standard Deviations (in Parentheses) of the Number of Letters, HAL Frequency, Number of Phonemes and Syllables as well as the English PLD20 for LD and HD items and their Difference along with Results of Significance Testing of this Difference

	LD	HD	Difference
HAL frequency	5113 (3447)	5454 (3656)	341 (794)
No. of letters	5.45 (0.64)	4.45 (0.59)	1.00* (0.13)
No. of phonemes	4.37 (0.54)	3.48 (0.51)	0.90* (0.11)
No. of syllables	1.68 (0.47)	1.05 (0.22)	0.62* (0.08)
PLD 20	1.78 (0.22)	1.18 (0.19)	0.59* (0.05)

Note. *$p < .05$

distance (see Kruskal, 1983) between these 80 IPA transcriptions, and 150,000 German word-IPA transcriptions found in the Deutsches Aussprachewörterbuch / German Pronunciation Dictionary (Krech, Stock, Hirschfeld, & Anders, 2009), and 60,000 Icelandic word-IPA transcriptions kindly provided by the The Icelandic Centre for Language Technology (http://www.iclt.is/). Subsequently, for each of the 80 English words, the 20 closest neighbours in German and Icelandic respectively were chosen, and their mean Levenshtein distance was calculated, resulting in German and Icelandic PLD20s. Two median splits were performed to create HD and LD groups according to the items' densities in these two languages. Classifying the items according to their German neighbourhood density, a median split was performed at a PLD20 of 2.80. For the Icelandic neighbourhood, the median split was performed at a PLD20 of 2.90. The items' PLD20s according to the German and the Icelandic neighbourhoods are much higher than the items' PLD20 according to the English neighbourhood. This is not surprising given that the distance between words in two languages (German - English; Icelandic - English) would be expected to be greater than the distance between words within the same language.Unfortunately, it was impossible to find items that had lower German and Icelandic PLD20s given the constraints of using pictures from a standardised pictures set ($n = 260$) that had a low frequency (leaving $n = 100$). Given the constraints of the item set, classifying the items using a cut-off as low as 1.50 would have left only a few HD items for the German and Icelandic neighbourhood classification. Therefore, the higher cut-offs were used.

The initial intention was to analyse the effect of density in the L2 (English) and the L1 (German or Icelandic) separately. However, it became evident that most items corresponded with regards to their density across languages (e.g., goat has many neighbours

in English and in the L1 of the participants). And indeed, there was a significant correlation between German and English PLD20s, $r(78) = .62$, $p < .001$, and between Icelandic and English PLD20s, $r(78) = .59$, $p < .001$. Analysing the data according to L1 and L2 neighbourhoods separately yielded very similar results. This way, it was not clear whether the priming effects were driven by the participants' L1 or their L2. Consequently, it was later decided to define and analyse the density groups according to their density in both languages (L1 Density (HD vs. LD) x L2 Density (HD vs. LD)), allowing for contrasting density groups.

In the end, the German HD item group consisted of 39 items with a mean PLD20 of 2.10 ($SD = .37$) ranging from 1.10 to 2.75. The German LD item group consisted of 36 items with a mean PLD20 of 3.34 ($SD = .41$) ranging from 2.85 to 4.00. The Icelandic HD item group contained 39 items with that had a mean PLD20 of 2.10 ($SD = .35$) ranging from 1.65 to 2.85. The Icelandic LD item group contained 36 items with that had a mean PLD20 of 3.30 ($SD = .45$) ranging from 2.95 to 4.35. Not all of these L1 items corresponded across German and Icelandic neighbourhoods (e.g., cloud is HD according to its German neighbourhood, but LD according to its Icelandic neighbourhood). As a result, it was not possible to run an item analysis.

Design

A 2x2 (Group x Prime per Density) Latin square design was used similar to the one in Experiment 2 except that there was no Task factor and there were only two instead of four subject groups. As illustrated in Table 3.5, within each of the density groups every item appeared in every condition, acting as its own control, and while each participant was exposed to all conditions, they only experienced each item once.

Table 3.5: Illustration of the Latin Square Distribution of Items across Conditions in Experiment 2

	Density	Either HD or LD	
Group	Prime	Unrelated	Related
1		atlam-ARROW (1)	geat-GOAT (2)
2		gelm-GOAT (2)	arfow-ARROW (1)

Note. Illustrative item numbers are given in parentheses

Apparatus

See Experiment 2.

Procedure

See Experiment 2, picture naming section. In Experiment 3 only the picture naming task was used, allowing for twice as many items per condition compared to Experiment 2. Because there was one less repeated factor (no task factor, compared to Experiment 2), after 10 practice trials, a total of 80 test trials were completed by each participant. The participants completed this task after completing the tasks outlined in Chapter 2 and before completing the tasks in the following experiments in one session with the opportunity to have a quick break in between tasks if needed.

3.3.2 Results

To investigate overall priming patterns when density was defined by the L2, four-way mixed-design ANOVAs were used to analyse the factors: Prime (unrelated vs. related), L1 Density (HD vs. LD), L2 Density (HD vs. LD), and Group. Only a subject analysis was conducted. Analysing the items was not possible, as their membership in the four Density groups did not always correspond across German and Icelandic.

Reaction Times

Mean RTs (ms) and standard deviations across conditions as well as for the priming effects can be found in Table 3.6.

Analysis revealed a significant L1 Density by L2 Density by Prime interaction, $F_1(1,46) = 5.40$, $MSE = 7007$, $p < .05$, $\eta^2 = .008$. Further analysis using Tukey's $HSD = 54.2$ with a familywise α of .05 revealed a significant form priming effect for items from all four Density combinations, as illustrated in Table 3.6. The form priming effect was equivalent across all conditions except for when the word was LD in both L1 and L2, where it showed the largest priming effect. A median split was performed on the participants' LexTALE (Lemhöfer & Broersma, 2011) scores to create two proficiency groups. The median split

Table 3.6: Mean RTs (ms) and Standard Deviations (in parentheses) for Pictures across L1 Density (HD vs. LD) by L2 Density (HD vs. LD) Preceded by Unrelated or Related Primes and Their Priming Effect

L2 Density	L1 Density		
Prime	HD	LD	Difference
HD			
Unrelated	838 (152)	832 (157)	
Related	711 (153)	716 (149)	
Priming Effect	127 (151)	116 (158)	11
LD			
Unrelated	842 (164)	859 (121)	
Related	727 (159)	674 (117)	
Priming Effect	115 (144)	185 (76)	-70
Difference	12	- 69	81

was performed at a LexTALE score of 75.15% (see Chapter 2 for more details). Mean RTs and standard deviations across the conditions for both proficiency groups can be found in Table 3.7.

Table 3.7: Mean RTs (ms) and Standard Deviations (in parentheses) for Pictures across the Four Density Combinations Preceded by Unrelated or Related Primes and Their Priming Effect in the Low and High Proficient Bilingual Groups

	Low Proficient Bilinguals			High Proficient Bilinguals		
L2 Density	L1 Density			L1 Density		
Prime	HD	LD	Difference	HD	LD	Difference
HD						
Unrelated	875 (168)	871 (170)		800 (127)	794 (136)	
Related	723 (126)	733 (124)		700 (177)	698 (172)	
Priming Effect	152 (170)	138 (180)	14	100 (127)	96 (134)	4
LD						
Unrelated	876 (171)	862 (115)		808 (152)	855 (130)	
Related	731 (123)	682 (108)		723 (191)	666 (129)	
Priming Effect	145 (154)	180 (81)	-35	85 (128)	189 (71)	-104
Difference	7	-42	49	15	-93	108

Low proficient bilinguals Analysis revealed no significant three-way L1 Density by L2 Density by Prime interaction, $F_1 < 1$. Only a significant main effect for prime was found, $F_1(1, 22) = 43.6$, $MSE = 25897$, $p < 001$, showing a large effect $\eta^2 = .424$. As shown in Table 3.7, there was a substantial priming effect for all Density combinations. Contrary to our expectations, priming effects did not vary significantly across L1 and/or L2 Densities.

High proficient bilinguals In contrast to the low proficient group, a significant three-way L1 Density by L2 Density by Prime interaction was found, $F_1(1,22) = 5.67$, $MSE = 6312$, $p < .05$, showing a small effect $\eta^2 = .021$. Further analysis using Tukey's $HSD = 63.7$ with a familywise α of .05 revealed a significant form priming effect for items from all four Density combinations, as illustrated in Table 3.7. The form priming effect was bigger when not only the L1, but also the L2 defined an item as LD, which is the source of the three-way interaction effect.

Error Rates

Following the findings in Experiment 2, smaller error rates on related prime trials compared to unrelated prime trials regardless of L1 or L2 Density were expected. Mean Error rates and their standard deviations can be found in Table 3.8.

Table 3.8: Mean Error Rates (%) and Standard Deviations (in parentheses) for Pictures across L1 Density (HD vs. LD) by L2 Density (HD vs. LD) Preceded by Unrelated or Related Primes and Their Priming Effect

L2 Density	L1 Density		
Prime	HD	LD	Difference
HD			
Unrelated	10.7 (9.28)	15.2 (15.2)	
Related	6.45 (6.57)	6.21 (10.3)	
Priming Effect	4.25 (9.80)	9.03 (15.9)	-4.78
LD			
Unrelated	9.55 (11.1)	12.9 (11.1)	
Related	6.51 (10.0)	5.48 (9.23)	
Priming Effect	3.04 (14.0)	7.42 (12.2)	-4.38
Difference	1.21	1.61	-0.40

Analysis revealed no significant three-way L1 Density by L2 density by Prime interaction, $F_1 < 1$. However, results showed a significant L1 Density by Prime interaction, $F_1(1, 46) = 8.75$, $MSE = 61.5$, $p < .01$, showing a small effect $\eta^2 = .016$, as illustrated in Figure 3.4. The priming effect for the LD items was greater than that for the HD items.

Further analysis using Tukey's $HSD = 4.27\%$ at a familywise α of .05 showed that, as expected, there were 8.4% fewer errors made on related prime trials compared to unrelated prime trials for LD items. The difference of 3.66% for HD items was not significant. Again, separate analyses were then undertaken on each proficiency group. Mean error rates and

standard deviations across the conditions for both proficiency groups can be found in Table 3.9.

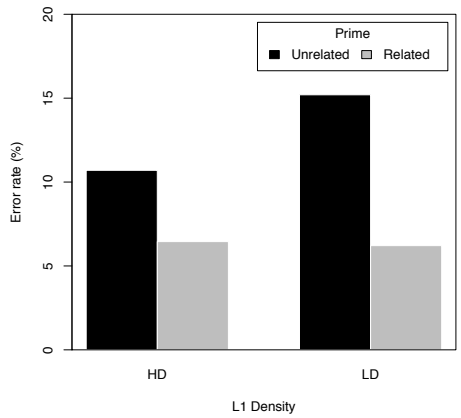

Figure 3.4: Mean error rates (%) across L1 densities and primes.

Table 3.9: Mean Error Rates (%) and Standard Deviations (in parentheses) for Pictures across the Four Density Combinations Preceded by Unrelated or Related Primes and Their Priming Effect in the Low and in the High Proficient Bilingual Groups

	Low Proficient Bilinguals			High Proficient Bilinguals		
L2 Density	L1 Density			L1 Density		
Prime	HD	LD	Difference	HD	LD	Difference
HD						
Unrelated	12.1 (10.4)	16.2 (14.0)		9.39 (8.02)	14.2 (16.6)	
Related	7.53 (6.83)	5.24 (9.68)		5.36 (6.26)	7.20 (10.9)	
Priming Effect	4.57 (10.4)	10.9 (16.2)	-6.33	4.03 (9.38)	7.00 (15.7)	-2.97
LD						
Unrelated	6.87 (10.1)	13.4 (12.5)		12.2 (11.5)	12.4 (9.75)	
Related	6.64 (12.2)	6.10 (7.21)		6.39 (7.51)	4.86 (11.0)	
Priming Effect	0.23 (13.5)	7.30 (12.8)	-7.07	5.81 (14.2)	7.54 (11.8)	-1.73
Difference	4.34	3.60	0.74	-1.78	-0.54	-1.24

Low proficient bilinguals Analysis revealed no significant three-way L1 Density by L2 Density by Prime interaction, $F_1 < 1$. However, results showed a significant L1 Density by Prime interaction, $F_1(1, 22) = 7.48$, $MSE = 73.5$ $p < .05$, showing a small effect $\eta^2 = .032$, driving the combined effect reported on the previous page. Further analysis using Tukey's HSD of 6.87% at a familywise α of .05 revealed that there was a significant error

difference between unrelated and related prime trials for LD items ($Diff = 9.17\%$), but not for HD items ($Diff = 2.41\%$). As expected, fewer errors were made on related Prime trials, however, contrary to expectations this was only true for LD items, similar to the effect shown in Figure 3.4.

High proficient bilinguals In the high proficient group, no significant three-way L1 Density by L2 Density by Prime interaction was found either, $F_1 < 1$. Only a significant main effect for prime was found, $F_1(1,22) = 112$, $MSE = 55.3$, $p < 01$, showing a large effect $\eta^2 = .882$. As expected, 6.24% fewer errors were made on related Prime trials compared to unrelated Prime trials.

3.3.3 Discussion

Experiment 3 confirmed our expectations by showing a significant difference in the magnitude of form priming effects for LD and HD items, with LD items showing greater effects. This is consistent with previous studies (Castles et al., 2007; Forster & Davis, 1991; Forster & Taft, 1994) and with the findings in Experiment 2. The initial expectation was that in the whole sample only the L1 neighbourhood would determine priming, however, both the L1 and the L2 densities influenced the magnitude of form priming. Items that belong to an LD neighbourhood according to both the L1 and the L2 showed the greatest amount of form priming compared to items with other density combinations. If previous interpretations of this effect are applied to these results, they suggest that, in bilinguals, phonological word forms in an L2 are integrated into the mental lexicon according to both their L1 and L2 neighbourhood density (at least the L2 density of the words already acquired), suggesting a single shared lexicon.

Looking at low and high proficient speakers separately, the results then show that, as expected, this integration is dependent on the proficiency of the bilingual. For low proficient late bilinguals no difference in the amount of priming was found between different density combinations. In contrast, high proficient bilinguals showed greater form priming for items that belonged to an LD neighbourhood in both the L1 and the L2, compared to all other density combinations. The finding that bilinguals differ in terms of L2 words'

integration into the mental lexicon is consistent with the previous suggestion that the organisation of L2 entries differs with varying L2 proficiency (Jiang, 2000). The finding is also consistent with the suggestion that recoding and restructuring of representations is a developmental process that is driven by the discriminative demands of a growing vocabulary (Castles et al., 2007; Storkel, 2002). The fact that neither the L1 nor the L2 density appeared to influence the amount of priming in low proficient bilinguals will be discussed in more detail below.

The results also show that, overall, fewer errors were made on related prime trials compared to unrelated prime trials. The amount of error reduction was bigger for LD items than for HD items as defined by the participant's L1. Assuming the LTH is correct, HD items should be more fine-tuned and less error prone than LD items. One could, however, assume that the error reduction in LD trials was caused by the advantage these items had in priming. In previous form priming experiments, no priming effects have been found for error rates (Castles et al., 2007; Forster & Taft, 1994). However in those studies recognition tasks were used, eliciting very few errors overall as also seen in the word naming task used in Experiment 2. In contrast to word naming, picture naming requires the participant to retrieve the name of the object. So if a picture is not recognised, its name cannot be simply read from the surface form (i.e., its orthography, using something like grapheme-to-phoneme correspondence rules). In that case, the presence of a related prime might then have served as a 'hint,' reducing error rates on related prime trials. And since HD items are less error prone in nature, their difference between related and unrelated trials would not be as big as for the more error prone LD items.

Curiously, when error rates were analysed for low and high proficient bilinguals separately, an interaction effect for Prime and L1 Density was found for low proficient bilinguals, most likely driving the effect in the overall sample. On the other hand, for high proficient bilinguals, only a difference in error rates between unrelated and related prime trials regardless of density was found. The latter finding is in line with the pattern of error rates of monolingual speakers found in the picture naming task in Experiment 2. It seems as if low proficient bilinguals benefit from the expected density difference in priming according to the L1 density when it comes to error rates, while the priming effect regardless

of which language defined density reduced error rates in high proficient bilinguals. In the high proficient group, it seems as if analyses could not detect a three-way interaction. On the other hand, the error pattern in the latter group is in line with what was found for monolingual speakers and might indicate a more native-like processing of the items.

3.4 General Discussion

The aim of this study was to investigate the development and storage of L2 entries in late bilinguals for production using a masked form priming picture naming task. As discussed above, the results indicate that L2 items are integrated into the mental lexicon according to both the L1 and the L2 density. At least this seems to be the case in high proficient bilinguals, since the results also suggest that this process differs between low and high proficient bilinguals. Assuming the LTH is correct, the results would be compatible with one of two possible scenarios for the development of the L2 mental lexicon.

3.4.1 Possible Scenarios for the Development of the L2 Lexicon

For example, as illustrated in Figure 3.5, as a late bilingual starts to learn an L2, L2 entries are integrated into the L1 mental lexicon. Some of those L2 entries belong in HD neighbourhoods in the L1 (e.g., barn and blouse), others belong to LD neighbourhoods in the L1 (e.g., wrench, snail). Those L2 entries that have many neighbours in the L1, would not only be integrated into that HD neighbourhood, but would also be recoded according to it.

However, as the individual acquires a greater vocabulary, more and more neighbours of L2 HD entries will be acquired too. As a result, and as illustrated in Figure 3.6, there would be four different neighbourhood configurations: HD_{L1}/HD_{L2}, LD_{L1}/LD_{L2}, HD_{L1}/LD_{L2} and LD_{L1}/HD_{L2}. The situation in corresponding neighbourhoods (HD/HD and LD/LD) would not change much as more L2 words are acquired; however, development is a bit more challenging for the HD_{L1}/LD_{L2} and LD_{L1}/HD_{L2} neighbourhoods.

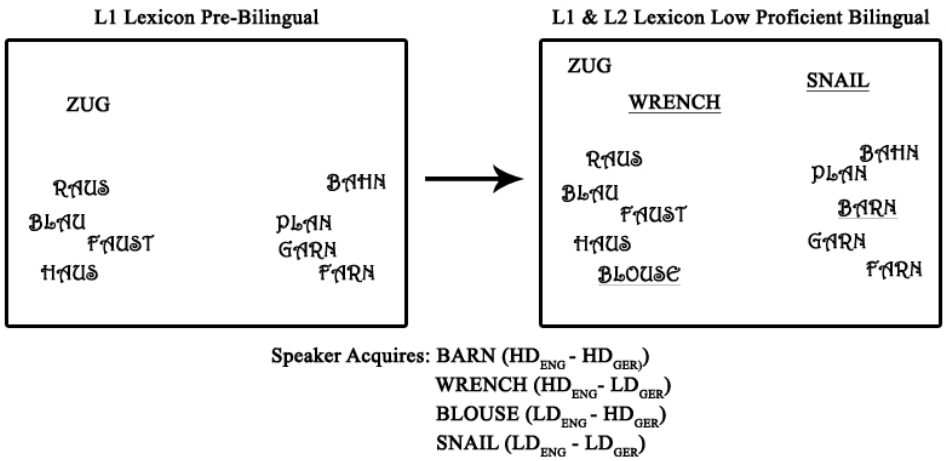

Figure 3.5: Lexical development in a low proficient bilingual.

In the former, L2 LD words would need to be recoded to meet the demands of the L1 HD neighbourhood (e.g., blouse), while in a monolingual L2 lexicon they would not be expected to undergo recoding. In the $\text{LD}_{L1}/\text{HD}_{L2}$ neighbourhood, the development of a greater L2 vocabulary is the most challenging (e.g., wrench). As shown in Figure 3.6, the arrival of the L2 words creates a HD neighbourhood in an original LD neighbourhood. As a result, at least the L2 entries would need to be recoded, if not both the L1 (e.g., Mensch) and the L2 items. In total, there would be less truly LD neighbourhoods in a bilingual's common lexicon than in the monolingual case. So bilinguals would have a greater percentage of retuned words.

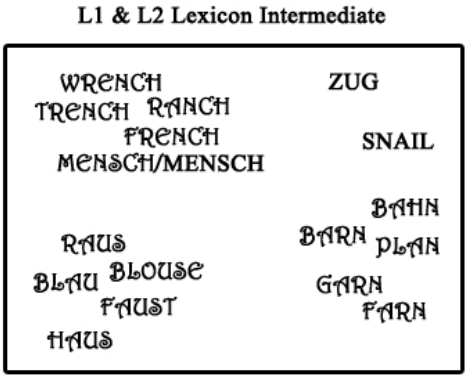

Figure 3.6: Lexical development in an intermediate bilingual.

Alternatively, as illustrated in Figure 3.7, it can be also envisioned that the lexicon is separated once the L2 portion of the integrated lexicon has grown in size. Maybe it

is 'uneconomical' to house word forms from two languages in one lexicon or functional separation is fostered as the two languages are used more independently (Kroll et al., 2010). As the lexicon splits, however, previously recoded L2 LD items from HD_{L1}/LD_{L2} neighbourhoods would probably keep their recoding, even in the absence of the L1 HD items. As a result, only the items that were never recoded, which stem from the former LD_{L1}/LD_{L2}, neighbourhood should show form priming as they do in Experiment 3 for high proficient bilinguals.

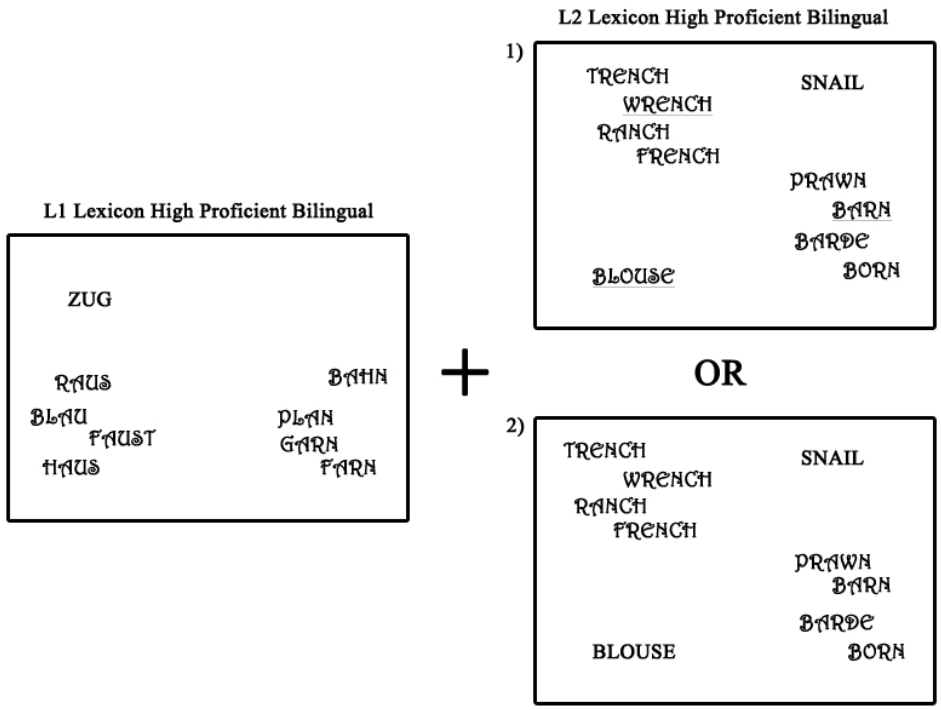

Figure 3.7: Lexical development in a high proficient bilingual.

3.4.2 L1 Influence on the L2 Production Process

As bilingual word production has gained more attention over the last decade, the accumulated evidence points to L1 being able to influence the production process in the L2 and vice versa (Colomé & Miozzo, 2010; Costa et al., 2003; Hermans et al., 1998; Kroll et al., 2006). Most of the studies argue that non-target language influence on the production of the target language is a result of interactivity between the L1 and the L2 lexicons, either through direct connections between word forms, or feedback activity at previous

levels. Assuming that the LTH is correct, the results of this study show that at least in late bilinguals, L1 and L2 must have been in contact at least at some point and that cross-language influence could also be a result of similar coding in a (previously) shared lexical space instead of being the result of cross-language activity.

After finding cross-language effects of phonologically related distracters in a picture-picture interference paradigm, Colomé and Miozzo (2010) argued that a serial, modular model of word production as proposed by Levelt et al. (1999) could not explain these effects because it restricts lexeme selection to the selected target lemma. However, assuming that L2 phonological word forms are coded at least in part according to L1 neighbourhood density, a phonologically similar L1 distracter could be able to either prime lexeme selection due to a previously similar coding or be able to be rejected faster at an output stage. As Jiang (2000) argues, more research is needed into how the two languages of a bilingual are not only activated but, more importantly, represented in order to shed more light on how efficient bilingual production is possible.

3.4.3 Limitations

There are some shortcomings to this study. First of all, the method to determine storage in the mental lexicon is based on methods used in monolingual production studies. While there is the assumption that both production and recognition use the same lexical entries and since Experiment 2 shows similar effects in a production task as had been found in a comprehension one, this does not mean that the effects have the same causes. While the overall results fit nicely with the LTH and the LRM, the conclusion over L1 and L2 representations and their coding should be treated with caution.

It would have also been preferable to look at the Icelandic and German participants separately to see whether the effects were the same in both groups. However, there were not enough participants (especially in the Icelandic group) to do so without critically reducing the statistical power. In fact, when the analyses were initially conducted per group, the three-way interaction was not significant in either group, displaying below recommended power, even though both groups were comparable in terms of both RT and error patterns. Given the fact that the groups were remarkably similar in their L2

history and given that their L1 belonged to the same language family, it was considered acceptable to combine them.

Also, while the results support our predictions, it has to be noted that in Experiment 3 no item analyses could be conducted as the item categorisation into HD and LD items did not correspond across German and Icelandic. Given the difficulties with item selection in this experiment, some of the items might not have been ideal. As mentioned in the method section, item selection was constrained by the fact that using a standardised picture set restricted the number of possible items to a small number. More participants would have been required to analyse the data in each language group separately. This would then allow item analyses to check whether the items were ideal and whether the effects found for subjects could be found for the items as well.

3.5 Summary

In this chapter, the development and storage of L2 entries in the mental lexicon of late bilinguals was explored. Adapting a form priming paradigm used in previous studies to study the development and storage of L1 entries (Castles et al., 2007; Forster & Taft, 1994), it was found that L2 entries seem to be integrated into the L1 lexicon. Additionally, the findings suggest that this integration in the L1 lexicon is most likely to occur for low proficient late bilinguals. Subsequently, as the late bilingual becomes more proficient, it then seems as if not only the L1 but also the L2 influences the integration or the coding into the lexicon. This influence might then lead to a separation into two lexicons.