A Model Comparison between Neural Architectures of Human Bilingual Sentence Processing

Rasmus Roslund

Master of Science
Cognitive Science
School of Informatics
University of Edinburgh
2021

Abstract

This work investigates phenomena related to human bilingual sentence processing in neural language models. We ask ourselves the question if and how the emergence of these phenomena depend on the model architecture. For this purpose, we train SRNs, LSTMs and Transformers with different hidden layer sizes as bilingual- and monolingual language models. We test these models on three phenomena that have been shown to emerge in at least one of the architectures in the literature. We refer to them as reading time prediction (Frank, 2014); an agreement between monolingual vs. bilingual reading with models trained on monolingual vs. bilingual data, the cognate facilitation effect (Winther et al., 2021); a faster processing of form and meaning-similar words, and the grammaticality illusion (Frank et al., 2016); a preference for the ungrammatical version of a certain class of sentences that is reversed in some languages.

Surprisingly, we found reading time prediction to depend not only on architecture and layer size, but also on the specific random initialization. Failure of reproduction of the effect was confirmed by the author, suggesting the original study to be due to coincidence.

As for the cognate facilitation effect, we found it to be present in the SRN and LSTM, providing further evidence for its emergence in humans to be due to the cumulative frequency of cognates. The effect was found to decrease in magnitude for large layer sizes in the LSTM, which can be linked to the LSTM relying less on corpus frequency. However surprisingly, the effect was found to increase for small layer sizes in the SRN. We do not have an adequate explanation for this trend. Furthermore, it was not found in the Transformer, suggesting Transformers to exhibit less cross-linguistic transfer than the other architectures.

The grammaticality illusion was found to be present in the SRN, but not in the LSTM and Transformer. This provides further evidence for the effect to arise as a result of short-distance language statistics rather than universal working memory constraints. The effect was found to stay fairly constant over layer size, syntactic linguistic transfer to be small. Furthermore, the transformer displayed a consistent preference for grammatical sentences, suggesting it to display superhuman syntactic proficiency on this particular task.

Acknowledgements

First and foremost, I would like to thank my supervisor Yevgen Matusevych for his guidance and support throughout this project. This project relied heavily on his expertise within the subject area and would not have been the same without his help. Second, I would like to thank Stefan Frank for correspondence about his work. He provided both data and useful information for the analysis of our evidence. Third, I would like my two friend Johan Dettmar and Nailia Mirzakhmedova for proof-reading some chapters of this work. Other friends have generously contributed with accomodation, include Ylva Roslund and Mats Göransson, Michael Foertsch, Ian Memgard and Beate Björkengren, Pascal Delabouglise, Michaela Muller-Trutwin and Daniel Boubet. Last but not least, I would like to thank friends and family, especially my dear Michael Foertsch, for their moral support throughout this challenging time.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Rasmus Roslund*)

Table of Contents

Chapter 1

Introduction

Humans have a unique capacity to learn more than one language, and we commonly do so. Although no clear data exist worldwide, it is believed that over half of the world's population has some level of proficiency of another language (Grosjean and Li, 2013, p. 5). This number is only predicted to increase in the future, with the combined forces of globalisation and digitisation.

The research on bilingualism has grown tremendously in the past years. These have presented numerous findings – for example, the speaking of two or more languages seems to be beneficial, rather than harmful, for the development of the brain (Mechelli et al., 2004). Other interesting findings include the fact that bilinguals seem to exhibit reduced decision biases when operating in a non-native language (Keysar et al., 2012), and that the bilinguals seem to be less emotionally attached to their second language, and are more likely to use expletives in it (Gawinkowska et al., 2013).

Research on bilingualism has long been fuelled by the use of computational models. Some well-known examples are given by the BIA+ model of bilingual word recognition (Dijkstra and Van Heuven, 2002), the Multilink model of bilingual word recognition and word translation (Dijkstra et al., 2019), the SOMBIP self-organizing model of bilingual processing (Li and Farkas, 2002), and the DevLex-II unsupervised neural model of bilingual lexical interactions (Zhao and Li, 2010).

By contrast to these word-level models of bilingual language acquisition and processing, there has been fewer models of sentence-level bilingual language processing. A promising model in this field is given by neural language models. Such models are trained to predict the next word in a sentence, and have gained significant improvements in recent years (Bengio et al., 2003; Mikolov et al., 2010; Radford et al., 2018), arguably reaching human-level proficiency (Radford et al., 2019), modelling complex

lexical properties (Mikolov et al., 2013) and syntactic agreement (Gulordava et al., 2018). Popular architectures include recurrent neural networks, as well as the recently proposed Transformer.

In monolingual sentence processing, this has led researchers to investigate the psycholinguistic properties of neural language models. They have been found to be able to significantly predict human reading times (Frank et al., 2015), and model specific effects of human sentence processing such as garden-path effects (Van Schijndel and Linzen, 2018; Futrell et al., 2019). However, there have been few models that explore language models in a bilingual setting. This is surprising, as bilingual neural language models require no architectural changes compared with monolingual ones, and allow for a range of applications related to linguistic transfer. It has been explicitly shown that language models exhibit both word-level lexical transfer (Dhar and Bisazza, 2020) and sentence-level syntactic transfer (Pires et al., 2019; Mueller et al., 2020) by virtue of shared representations. This provides a rich test bed for the probing of models of bilingual sentence processing.

To our knowledge, bilingual language models has only been harnessed in three studies about phenomena specific to human bilingual sentence processing. These studies will be referred to as reading time prediction, the cognate facilitation effect and the grammaticality illusion.

Reading Time Prediction (Frank, 2014) This study looked at how well predictions from monolingual and bilingual language models predicted human L1 and L2 reading times. The study found the monolingual model to predict L1 behaviour better, and the bilingual model to predict L2 behaviour better. This points toward L2 reading being influenced by L1 behaviour beyond word-level lexical phenomena.

Cognate Facilitation Effect (Winther et al., 2021) A reason for the transfer between L1 and L2 is the presence cognates, i.e. translation equivalents that share orthographic and semantic properties. Such words have been found to be processed faster than non-cognates in an effect commonly referred to as the cognate facilitation effect (Costa et al., 2000; Dijkstra et al., 1999). Winther et al. investigated the cognate facilitation effect, where cognates were embedded into English sentences. They found a neural model to reproduce the effect under certain conditions of the presentation of training data.

Grammaticality Illusion (Frank et al., 2016) English native speakers have been consistently found to judge ungrammatical derivations from a certain class of grammatically correct sentences as more acceptable (Frazier, 1985, pp. 129–189; Christiansen

and MacDonald, 2009). The reverse happens in Dutch, where native speakers prefer the grammatical versions (Frank et al., 2016). When Dutch speakers are tested in their L2 English however, they behave as English speakers, showing no evidence of language transfer. Frank et al. tested if a neural language model could reproduce the grammaticality illusion described above (Frank et al., 2016). They found a neural model to correctly reproduce this phenomenon (Frank et al., 2016).

The above mentioned bilingual studies made use of specific neural architectures with specific hidden layer sizes to reproduce their results. Frank (2014), and Frank et al. (2016) used a simple recurrent network (SRN, Elman, 1990), and Winther et al. (2021) used a long short-term memory network (LSTM, Hochreiter and Schmidhuber, 1997). However, the literature suggests that the behaviour of language models depends crucially on type of architecture and layer size. This leads to the following research question: How does the emergence of the described bilingual phenomena vary with architecture? More specifically, what effect does the gated nature of the LSTM have compared with the SRN? How does the emergence of the phenomena change when using attention-based networks instead of recurrent networks? What effect does the hidden layer size have on the emergence of the phenomena?

In order to answer such questions questions, we will consider an SRN, an LSTM and a Transformer (Vaswani et al., 2017) as potential architectures. We will train instances of models from these architectures with various hidden layer sizes in order to analyze how well they account for the bilingual phenomena described above.

The paper is structured in the following order. In chapter 2, we introduce neural language models, and describe how they are used in the psycholinguistics of bilingualism. In chapter 3, we describe our research proposal more in detail, and gives further details on our implementation. In chapter 4, we present our results on language model performance, as well as how well the models account for the three phenomena described above. Finally in chapter 5, we draw conclusions related to the results, and discuss further directions of research.

Chapter 2

Background

This section presents some background knowledge on which the present work is based. We will first start out by describing language models, introducing some of the main neural architectures that are used. Subsequently, we will look at how these language models are used in human sentence processing. Finally, we will focus on bilingual sentence processing, and more specifically, the three studies on which the present work is based.

2.1 Neural Language Models

Consciously or not, humans predict the next word in a sentence. Consider, for example, the following sequence of words:

(1) $S_1 =$ The carpenter who the craftsman ...

A human speaker is likely judge it more probable for the next word to be a verb such as *saw* or *met* than a determiner or noun such as *the* or *net*. Language models capture this intuition by assigning a probability value $P(w_i|w_{<i})$ to a word w_i given its context $w_{<i} = w_1 \dots w_{i-1}$. In order to effectively predict the next word, the model needs to integrate the information from the context by gaining some statistical knowledge of the language. One could of course do this in a rule-based way, but we will discuss learning such information directly from a corpus of text.

How does it gain this statistical knowledge? The simplest way is probably given by counting occurrences in the corpus. For this purpose, the intuition that *met* is more likely to follow than *net* in the above example, one could compute the probabilities of the two sequences by simply counting occurrences of the two sequences

(2) S_2 =The carpenter who the craftsman met

(3) S_3 =The carpenter who the craftsman net

and divide by the context, i.e. S_1.

$$P(\text{“met”}|S_1) = \frac{\#S_2}{\#S_1} \qquad P(\text{“net”}|S_3) = \frac{\#S_2}{\#S_3}$$

However, such an approach is limited, since language is creative – often we encounter sentences that are not present in our corpus. In the corpus used in this work, for example, neither of these sentences appear.

A simplification to make this work is to reduce the context size to a smaller, fixed number. Such approaches are called n-gram language models. In the above example, a 3-gram would correspond to ignoring the first part of the sentence, just using the last two words.

$$P(\text{“met”}|\text{“the craftsman”}) = \frac{\#\text{“the craftsman met”}}{\#\text{“the craftsman”}}$$
$$P(\text{“net”}|\text{“the craftsman”}) = \frac{\#\text{“the craftsman net”}}{\#\text{“the craftsman”}}$$

This works better, but does not generalize to unseen n-grams. Say, for example, that the above trigram is not in our corpus, but instead sequences such as “the father met,” “the secretary met” and “the peasant met” in our corpus. A count-based approach cannot integrate this information in its predictions.

2.1.1 Neural Architectures

A solution is given by language models based on neural networks. Such models are able to harness the lexical similarity between words by representing them as layers of neurons, or vectors in a multidimensional space. Neurons that are holders for numbers called activations. Subsequent layers of neurons are linked by connections of different strengths, referred to as weights. By representing a word as a specific number x between 1 and V, the vocabulary size, the word can be represented to the network as a one-hot encoding, i.e. as a vector of length V with all entries equal to zero except for the xth entry, which is equal to one. This vector works as the first layer of neurons, and the activations of subsequent layers are calculated iteratively as a function of the activations of the previous layer and the weights between the two layers. The final layer can be described as a layer of size V, and represents a probability distribution over the vocabulary.

The neural language models are trained by gradient-based methods that minimize some objective function based on how accurate a prediction is. The error is then back-propagated in the network, adjusting the weights that resulted in the mismatch between desired and predicted result. In order for the language model to learn the language well, it often has to see a lot of data, resulting in multiple iterations over the training corpus, or epochs. Furthermore, in order for the network to be able to learn a useful function, the weights are initialized randomly. In contrast to count-based language models, this leads neural language models to learn different things for each random initialization.

An early approach to neural language models is given by feed-forward neural networks, as first presented by Bengio et al. (2003). Similar to n-gram models, such networks take a context of a fixed size as input and propagates this input through subsequent layers to finally predict the probability of the next word. However, this makes the rough assumption that the to-be predicted word is independent on words beyond the context size. For example, consider the sentence:

(4) Paris is one of Europe's major centres of finance, diplomacy, commerce, fashion, gastronomy, science, and arts, and serves as the capital and most populous city of France.

Obviously, in order to correctly predict *France* as the last word in this sentence, the model would benefit from the first word, *Paris*. Furthermore, language is fundamentally a sequential phenomenon, and this is not reflected in a feed-forward neural network, as it treats each of the input words equally.

In order to solve such issues, a different neural architecture called recursive neural networks (Elman, 1990) have proven more successful. Such networks integrate information in the context by incrementally updating a hidden layer of neurons from word to word. This approach is commonly used for sequential data. Thus it is able to, at least in theory, capture long-term dependencies, and has proven superior to feed-forward based language models (Mikolov et al., 2010).

In their simplest form, they are referred to as simple recurrent networks (SRNs). These networks work better as language models (Mikolov et al., 2010) but are difficult to optimize. The difficulty is commonly described as the vanishing (or exploding) gradient problem, and describes the rapid decrease (or increase) of gradients as they are back-propagated through the network. A solution to this is given by the addition of supplementary gates that control the flow of information, telling the network what to remember and what to forget. A popular such network is given by long short-term

memory network (LSTM), first introduced by Hochreiter and Schmidhuber (1997). These have proven more effective as language models (Sundermeyer et al., 2012).

An alternative to recurrent architectures is given by the recently introduced Transformer (Vaswani et al., 2017), which directly attends to previous words by the use of an attention mechanism (Bahdanau et al., 2014). Specifically, Transformers use a specific form of attention called self-attention, which can be described as computing a score between each word embedding based on how related they are. These scores are then used to compute a weighted average over the word embeddings for each word. Transformers are currently the best-performing language models (Radford et al., 2018).

How is the performance of a language model evaluated? A basic evaluation measure of language model accuracy in language models is a quantity called perplexity as calculated over a sequency of words w_i given their context $w_{<i}$:

$$PPL(w_i, ..., w_n) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\log P(w_i|w_{1...i})\right) \tag{2.1}$$

where a lower perplexity means a higher language model accuracy. It is calculated over a set of sentences, and its accuracy grows with the number of sentences. In order to make sure that the language model is not just learning the particularities of the data on which it is trained, is is common practice to split up the data into a training and validation set, and report values on the validation set.

2.1.2 Bilingual Language Models

In a monolingual setting, sentences from one language are fed into the language model as training, and the model learns statistical knowledge about that language. Basic such knowledge simply includes word frequency. More advanced knowledge includes syntactic properties, such that a noun usually follows a determiner and lexical properties, such that the word *prison* often co-occurs in a sentence with *criminal*.

A bilingual language model requires no structural changes – sentences in two different languages are simply fed to the model as in the monolingual case. Again as in the monolingual case, the language model learns statistical knowledge from both languages. However, bilingual language models additionally need to learn to distinguish between the languages, as words in one language are usually followed by a word in the same language. French (1998) showed that such knowledge about language identity is quickly learned by neural models.

Furthermore, a bilingual setting allows for interaction between languages. Since both languages are captured by the same hidden layers, the languages may interact with each other and share related representations. This can lead to a positive linguistic transfer, where L2 proficiency profits from knowledge in L1, or to negative interference, where knowledge in L1 has detrimental effects on L2 learning. It is commonly believed that highly related languages affect each other more positively than less related languages, but the question of when positive language transfer occurs remains a question open to debate (see for example Mueller et al. (2020) or Dhar and Bisazza (2020)). Furthermore, it is not necessary for the languages to interact in a language model, as is the case if they are represented by different neurons in the hidden layers.

Third, bilingual language models differ from monolingual ones in that they use a shared vocabulary. That is, some of the words have the same form in both languages and are therefore represented by the same one-hot encoding, such as the word *partner* in English and Dutch. In this example, such words are cognates, meaning that they also share meaning. However, the words might mean different things, such as the word *brand*, which means fire in Dutch. If the bilingual language model is given no supplementary information about the language ID, it has to use the same embedding representation for such words. One can, however, enforce two separate embedding representations by including some external language identical information, often called language tags.

2.2 Human Sentence Processing

Language models are useful in a range of applied tasks such as speech recognition, text correction and machine translation. In this work however, we will apply language models as models of psycholinguistics. Such research is interested in the modelling of specific human language processing effects. This field often makes use of a word's surprisal value s_i, given by its negative logarithmic probability:

$$s(w_i) = -\log P(w_i|w_{<i}) \tag{2.2}$$

An example in which surprisal is used is given by garden path effects; a phenomenon which describes syntactially ambiguous sentences to be more difficult to process than others. Consider, for example, the below two sentences:

(5) Even though the girl phoned the instructor was very upset with her for missing a lesson.

(6) Even though the girl phoned, the instructor was very upset with her for missing a lesson.

In example 5, *the instructor* can be interpreted either as the object of *phoned* or as the subject of the upcoming clause. However in example 6, the comma prevents this ambiguity. As a result, the words *was very upset* are often processed with more difficulty by human subjects – they are said to be led down a "garden path". We will use this effect as an example to illustrate how it can be studied using neural language models.

2.2.1 Psychometric Prediction

The difficulty with which such a sentence is processed in humans can be measured in different ways. One approach is to ask participants to rate sentences in terms of their acceptability, or ask subjects facts that depend on a specific parse of a sentence (did the girl phone the intructor?). Another approach is to directly measure some psychometric quantity related to the difficulty of processing. Such quantities include self-paced reading times, eye-tracking, or brain activity.

Psychometric data has been shown to be significantly predicted by surprisal values extracted from neural language models (Smith and Levy, 2013; Frank et al., 2015). Such studies usually measure the *predictive power* of surprisal values by the use of linear mixed model regression. Linear mixed model regression is similar to linear regression, but is able to take both fixed effects (the predicting values, e.g. surprisal) and random effects into account. The inclusion of random effects in a regression model is suitable for when there is non-independence in the data, as is the case in studies with different human subjects. Random effects account for the variability over groups in the data (for example, some human subjects may read slower than others). It is common practice to include not only surprisal values as a predictor, but also more superficial cues such as word length and frequency, as these have been found to significantly predict human psychometric data (Rayner, 1998, pp. 372–422). This allows to isolate the predictive power of surprisal values by measuring the difference in quality of fit, as quantified by the deviance, between a regression excluding surprisal values, and one including surprisal values.

The literature has found that language models with better language model accuracy result in better predictions of psychometric data (Merkx and Frank, 2020). Given the same language model accuracy, predictions from SRNs and LSTMs are comparable in their predictive power (Aurnhammer and Frank, 2019), but predictions from

Transformers are superior (Merkx and Frank, 2020; Wilcox et al., 2020).

2.2.2 Targeted Evaluation Approaches

The relation between surprisal values and psychometric data have inspired some authors to use surprisal values in more targeted settings. For example, Van Schijndel and Linzen (2018) extracted surprisal values using sentences such as examples 5 and 6 above to study the emergence of garden-path effects in neural language models. The authors did find some evidence of garden path effects in the language models, but not to the extent of human subjects.

Such studies shine light on the underlying mechanisms of the effects. In the case of garden path effects, there are two main hypotheses. One theory states that it is due to special reanalysis mechanisms. Another theory states that such effects arise merely from the statistical unpredictability of the words *was very upset* following the immediate context *Even though the girl phoned the instructor*. Since language models capture statistical patterns of language and do not make use of any special reanalyses mechanisms, the absence of a strong effect implies less evidence for the unpredictability hypothesis.

2.3 Bilingual Human Sentence Processing

The example described above elucidates the existence of two different kinds of studies in human bilingual sentence processing – studies that assess the quality of neural language models by relating their predictions to psychometric quantities, and studies that use the models in targeted settings to measure some effect, often to make claims about the underlying statistical mechanisms. This distinction extends naturally to the bilingual realm. Among the three studies that the present work is based on, the first one belongs to the former category, assessing whether L2 psychometric data is better accounted for by a bilingual or monolingual language model, and the two other studies use bilingual models in a targeted context. We will present these studies more in detail below.

2.3.1 Reading Time Prediction

Frank (2014) looked at how well surprisals from monolingual and bilingual language models predicted human L1 and L2 reading times. The study found the monolingual

model to predict L1 behaviour better, and the bilingual model to predict L2 behaviour better.

Furthermore, the findings of this study has implications on human bilingual reading. It has long been found that L1 proficiency affects L2 reading (Duyck et al., 2007), and conversely, that L2 knowledge affects L1 reading (Van Assche et al., 2009). However, such studies have mostly relied on word-level experiments. As such, it remains open to debate whether these effects are modulated by sentence context rather than arising from purely lexical phenomena. Frank's study implies that L2 reading is influenced by L1 proficiency beyond the word level.

2.3.2 Cognate Facilitation Effect

One reason for why the word-level influence of L1 on L2 reading is the presence of cognates. Such form- and meaning-similar words have been used in a range of studies to elucidate the dynamics of the human bilingual lexicon (Costa et al., 2000). A consistent finding is that human bilinguals have been found to process them more easily in their L2 than non-cognates, as confirmed by various sentence-level experiments (Costa et al., 2000; Dijkstra et al., 1999; Libben and Titone, 2009).

In the literature, there are two major theories to account for this effect. One account holds that cognates have a special status in the brain and are therefore more easily processed (Van Hell and Dijkstra, 2002). Another theory is given by the cumulative frequency hypothesis, which explains the effect merely in terms of the higher cumulative frequency of cognates (Voga and Grainger, 2007).

In order to test the cumulative frequency hypothesis, Winther et al. (2021) recently trained a bilingual LSTM on a Dutch-English and Norwegian-English corpora, treating English as L2 and Dutch and Norwegian as L1. They computed surprisal values on cognates and non-cognates (referred to as controls) on sentences such as the one in examples 7 (cognate) and 8 (control).

(7) The attorney consults an *expert* for a detailed opinion on the matter.

(8) The attorney consults a *lawyer* for a detailed opinion on the matter.

By experimenting with different presentations of the training data to the model resulting in different ratios of language model accuracy between L1 and L2, they found evidence of the cognate facilitation effect in some settings. A prerequisite for the effect to occur was for the model to learn L1 well enough, as seen on its perplexity measured

on a validation set. More specifically, the training data consisted of a majority of L1 sentences, and either 1) the model was pretrained on L1 only, and after pretraining, languages were presented in a mixed order, or 2) the model was not pretrained and in each epoch, the model was first trained only on L1, and then only on L2. The presence of the effect reinforces the cumulative frequency theory, as cognates were processed in the same way as other words, with no specific cues related to their status as cognates.

2.3.3 Grammaticality Illusion

An astonishing phenomenon is that some grammatical sentences in English are judged as more grammatical by native speakers when rendered ungrammatical by the omission of a verb. This effect was first described by Frazier (1985, pp. 129–189), and refers to sentences containing double-embedded relative clauses. Such a sentence is depicted in example 11, whose derivation is shown in examples 9 and 10. Example 9 describes a sentence consisting solely of a main clause, example 10 describes the same sentence with the addition of a relative clause (italic). Example 11 adds a second relative clause (bold) within the first relative clause.

(9) The carpenter supervised the apprentice in the garden.

(10) The carpenter *who the craftsman hurt* supervised the apprentice in the garden.

(11) The carpenter *who the craftsman* ***who the peasant carried*** *hurt* supervised the apprentice in the garden.

Double-nested structures such as the one in example 11 have found to be very difficult to process (Hamilton and Deese), However if rendered ungrammatical by the omission of the verb pertaining to the first relative clause, *hurt*, they are easier to process. This results in the sentence depicted in example 12, whose ungrammaticality is indicated by an asterisk.

(12) *The carpenter *who the craftsman* ***who the peasant carried*** supervised the apprentice in the garden.

This somewhat peculiar effect seems to be language-specific rather than universal – it does not appear in L1 German (Vasishth et al., 2010) or Dutch (Frank et al., 2016). Vasishth et al. (2010) argue that the language specific property of the grammaticality illusion arises from the difference in word order in the languages – English subordinate clauses have SVO (subject-verb-object) order, whereas German and Dutch subordniate

clauses have SOV order. The SOV order forces native speakers to retain verb phrases in memory for a longer time, resulting in their working memory being more robust toward structural forgetting.

A different account is outlined by Frank et al. (2016), who argue that if Dutch subjects indeed have a more robust working memory for SOV structures, this should transfer to the behaviour in English. In order to test this hypothesis, they tested Dutch L1 speakers in their L2 English, and found the subjects to display the grammaticality illusion similar to English L1 behaviour. These findings speak against working memory constaints as a modulator of the grammaticality effect, and instead hypothesize that the difference emerges as a result of language statistics.

In order to further test this hypothesis, Frank et al. (2016) trained a bilingual SRN language model on a Dutch-English corpus and tested it on the grammaticality illusion. Similar to human subjects, they found the SRN to display no linguistic transfer, as it exhibited the grammaticality illusion in English but not in Dutch. This reinforces the belief that it is due to language statistics – if the SRN had developed more robust properties in Dutch, this could be expected to have transferred into English.

Chapter 3

Method

In this chapter, we provide details related to the work carried out in this project. The chapter is divided into two sections. In the first part, we will describe the models that we used. In the second part, we will describe how these models were used in the three test settings.

3.1 Design Choices

As previously mentioned, we implement three architectures – a SRN, a LSTM and a Transformer. We chose the SRN and LSTM because they have been succesfully used in previous studies. The reason for the use of Transformers is because they have been successfully in different psycholinguistics. For example, they depict superior behaviour to recurrent architectures in terms of syntactic agreement (Mueller et al., 2020), and have been shown to more accurately predict reading times (Merkx and Frank, 2020).

Another fundamental design choice is to look at the hidden layer size in the model. The reason for this choice is that many bilingual phenomena are related to cross-linguistic transfer. These exists little evidence in the literature on if, and how, cross-linguistic transfer depends on the layer size in a neural network. Our intuition is that a smaller layer size should result in a larger amount of transfer, since it forces the two languages to be squeezes into a small number of neurons. Conversely, in larger networks the layers can more easily develop two separate representations.

Another factor related to the behaviour of a language model is the amount of training one uses. There are several possible settings for this. One can either train a model until convergence, that is, until the model shows no more improvement on a validation

set. Such a setting is often resource-heavy, and since we implement many different models of each architecture, this would be unfeasible within the given time frame. Furthermore, it is not clear if this improves performance on specific tasks. Another possibility is to control the language model quality on a validation set, and stop training a model when it reaches this accuracy. In one way, this would allow for a fair comparison. However in another sense, such an approach is unfair since it would force larger, more expressive networks to stop training long before they have reached their full potential. We chose a simpler approach that can be thought of as a compromise between the two – to train our models for a fixed number of epochs.

The bilingual models are trained on a mixed corpus of the same size as the Dutch and English data, where each new sentence has 50% chance of coming from each language resulting in a "balanced" bilingual model. Furthermore, we follow Frank (2014) and Winther et al. (2021) and train additional monolingual models. Corpus statistics are given in tab. 3.2.

Winther et al. (2021) could not find evidence for the cognate effect in the training regimes described above. They found it necessary to use unbalanced proportions of L1 and L2 data in their training corpus, as well as either 1) split up the training data into separate L1 and L2 parts, or 2) include a period of pretraining. As we are interested in reproducing the cognate effect, we chose to implement one of these models. Because of time and resource constraints, we opted for the former version, such that the training corpus consists of 75 % Dutch data followed by 25 % English data. The Dutch data is presented before the English data in each epoch, thus treating English as L1 and Dutch as L2. This model will only be used for the cognate facilitation effect.

As for testing, we focus on surprisal values extracted from the language models as in the original studies. An alternative approach to this is given by looking at the activations of the hidden layers more directly, as it allows to quantify the concept of linguistic transfer. There are different ways to do this, but one way would be to train diagnostic classifiers to predict the language ID for a set of English and Dutch sentences. The accuracy of such a diagnostic classifier on a validation set gives an idea of the linguistic transfer in the model, since a high accuracy would means that the languages are separable. This has been performed in the literature by, for example, Libovick et al. (2020). However, preliminary testing found this approach to be flawed in a fundamental way. As we are looking at different layer sizes, we are fundamentally comparing vectors of different sizes. An approach using diagnostic classifiers on such vectors is not fair, as larger layer sizes lead to more overfitting. This results in a better

accuracy on a training set, and a worse accuracy on a test set, and it becomes difficult to isolate the amount of linguistic transfer.

3.2 Models

3.2.1 Implementation Details

Our implementations are based on the implementation in PyTorch from Van Schijndel and Linzen (2018), as it includes options for recurrent networks as well as useful functionality for training, testing and layer visualization. For the SRN, we use the hyperparameters from Frank (2014) and Frank et al. (2016). As for the LSTM, we adapt the hyperparameters from Winther et al. (2021), which are identical to those used by Gulordava et al. (2018).

As for the Transformer, there is no such straightforward choice. Most Transformer models used in the literature are pretrained, which is undesirable as we want control over the training process in order to yield a meaningful comparison between architectures. As a result, we implemented a Transformer from scratch using the Transformer blocks described in Vaswani et al. (2017). In order for the model to be comparable to the other architectures, we disposed with many modern tricks for best performance. Mainly, stochastic gradient descent is used instead of a more sophisticated optimizer, no cyclic learning rate scheduling is used, and the only form of regularization is dropout. The full hyperparameters of the models are summarized in tab. 3.1.

BPTT Steps The SRN is inspired by Frank (2014), who used only 3 backpropagation steps. It is thus suitable for modelling simple, short-term dependencies. For the LSTM

Table 3.1: Hyperparameters of the implemented models.

Architecture	**SRN**	**LSTM**	**Transformer**
BPTT Steps	3	35	35
Hidden layers	1	2	8
Learning rate	2	20	2
Batch size	64	64	64
Dropout rate	0.2	0.2	0.2
Training epochs	10	10	10
Number of heads	N/A	N/A	8

and Transformer however, we followed Gulordava et al. (2018) and used 35 BPTT steps.

Number of hidden layers The models also differ in depth – the SRN uses only one hidden layer as in Frank (2014), the LSTM uses two hidden layers as in Gulordava et al. (2018), and the Transformer uses 8 hidden layers, since deep Transformers have proven successful in the language modelling literature (Devlin et al., 2018; Radford et al., 2018).

Embedding size We use the following embedding sizes for all architectures: 32, 64, 128, 256, 512 and 1024. For the SRN and LSTM models, the hidden layer is set equal to the embedding size. For the Transformer, the feed-forward layer is set to twice the embedding size.

Learning Rate The learning rate of each model is chosen from a grid search over 20, 2, 0.2 and 0.02 by monitoring the perplexity on a validation set.

Training time All models are trained for 10 epochs. This is because preliminary testing found the results to stabilize after approximately this time. This is a compromise between successful implementations that correlate model predictions with psychometric data, and implementations that reproduce specific psycholinguistic effects. The former usually use short training times – Aurnhammer and Frank (2019), for example, train their models for 1 epoch or less. The latter use long training times – Winther et al. (2021), for example, train their models for 30 epochs or more. Since we are interested in an implementation that can both correlate with psychometric data, and model specific psycholinguistic effects, we chose a training period that lies between the two. ALSO

3.2.2 Training data

We use the English Wikipedia corpus from Gulordava et al. (2018) and the Dutch Wikipedia corpus from Winther et al. (2021). We follow the preprocessing steps from Winther et al. (2021) a vocabulary of the 50k most frequent tokens is created and out-of-vocabulary words are replaced with the unknown token.[1] Sentences that consist of only one word or contain more than 5 % unknown words are excluded. The English and Dutch corpora are matched in size, resulting in a corpus size of 2.0 million sentences for each language. This is split up in 80-10-10 training-validation-test sets.

[1] An exception to this is made in the Dutch corpus, which is forced to include the tokens in the test sentences from Frank et al. (2016), as it was found that otherwise many crucial words were excluded.

For the bilingual models, we mix the data from the Dutch and English corpora. However, in doing so we only include half the monolingual sentences in each language, such that the total bilingual corpus size equals that of the monolingual corpus sizes. Furthermore, the languages are mixed in a random fashion, such that the language ID of each new sentence is drawn from an unbiased Bernoulli distribution over the two languages.

3.3 Evaluation

3.3.1 Implementation Details

As the models are trained, their perplexity is computed on the validation sets at the end of each epoch during training in order to track their language model accuracy. This indicates the language model accuracy, which is useful for further analysis. For the rest of evaluation, we follow the implementations from the specific studies. That is, we compute surprisal values on the sentences from the three different studies. These are then further evaluated as described in detail for each task below.

Reading time prediction Frank (2014) fit a rather basic linear mixed model regression, using only surprisal values as a fixed effect. We hope to find more exact results including more effects, as this allows to isolate the informativeness of surprisal values better. For this purpose, we follow the approach in Aurnhammer and Frank (2019) and use linear mixed model regression. We fit a "baseline" model excluding surprisal values, and a "full" model including surprisal values. The baseline model includes word length and logarithmic frequency as fixed effects, as well as their interactions. As random effects, we include subject ID and word token as random intercepts, as

Table 3.2: Corpus statistics given in millions.

	Tokens (M)				Sentences (M)
Corpus	**English**	**Dutch**	**Bi-50**	**Bi-75**	**All Corpora**
Train	42.5	29.0	33.4	32.2	1.6
Validation	5.3	3.6	4.2	3.9	0.2
Test	5.3	3.6	4.2	3.9	0.2
Total	53.1	36.3	41.8	40.0	2.0

well as by-subject random slopes for all fixed-effect predictors. The full model additionally includes surprisal values as both fixed effect and by-subject random slope. We compute the goodness-of-fit as the difference between the deviances between the two models:

$$\text{goodness-of-fit} = \text{deviance}_{\text{base}} - \text{deviance}_{\text{full}} \tag{3.1}$$

which describes the ratio of the likelihoods according to the two fits. We use the lme4 package for linear mixed model regression in R (Bates et al., 2014).

Cognate Facilitation Effect As opposed to reading time prediction, we only make use of surprisal values on the cognate and control words. For this purpose, we define the cognate effect size as the difference in surprisal values on the two tokens:

$$\text{cognate effect size} = s(w_i^{control}) - s(w_i^{cognate}). \tag{3.2}$$

These values are further processed by calculating the mean value and standard deviation over sentences. This is performed for three random initializations of the language models, from which we again compute the average value, as well as the standard error.

In order to see how the cognate effect size relates to frequency in the training data, we compute Pearson's correlation between the cognate effect size and the difference in logarithmic frequency between cognate and control token. This is performed for the English, balanced bilingual, and unbalanced bilingual training sets described in tab. 3.2.

For statistical analysis, we make use of a paired t-test between the surprisals on the grammatical sentences and those on the ungrammatical sentences, and report statistical significance with a confidence interval of 0.05. As we are reporting results over multiple models, it is important to keep in mind that a 5% of these are false positives (Benjamini and Hochberg, 1995). This could be accounted for by using some error correction technique. However, since we are more interested in comparing models than in a robust assessment of statistical significance, we do not perform any correction.

Grammaticality Illusion We follow a similar approach as in the cognate facilitation effect – we define the grammatical preference as the difference in surprisal on tokens from the grammatical and ungrammatical versions of a sentence:

$$\text{grammatical preference} = s(w_i^{ungrammatical}) - s(w_i^{grammatical}). \tag{3.3}$$

These values are then further processed as in the cognate effect. We use the the first determinant, *the*, immediately following *supervised* in example 11. In contrast to Frank et al. (2016), we do not present results on the third verb, *supervised* in example 11.

This is because we did not find any consistent trend on these words, which is consistent to their findings that the effect was much smaller on the third verb than the first determinant.

3.3.2 Test Sentences

For reading time prediction, we follow Frank (2014) and use the self-paced reading times on the 56 Dutch and 56 English filler sentences from Frank et al. (2016). Unfortunately, these sentences were found to contain many out-of-vocabulary words, although we used a large voabulary size of 50,000.[2], amounting to about 11% and 14% of the total amount of tokens in the English and Dutch data set, respectively. Since the majority of sentences contained at least one unknown token, it would have been unfruitful to discard each such sentence. Instead, we kept surprisal values for all tokens that were in our vocabulary.

For the cognate effect, we used the same 21 sentences with cognates and controls (totalling 42) as Winther et al. (2021), and found all tokens to be in our vocabulary. The sentences are included in Appendix A.

For the grammaticality illusion, we used the 16 English and Dutch sentences from Frank et al. (2016). We found the majority of the Dutch sentences to include words that were not in our vocabulary. Therefore, we re-processed the entire Dutch corpus to include these crucial words in our vocabulary. The vocabulary size was however kept to the size 50,000 by excluding other words. The sentences are included in Appendix B.

Table 3.3: Statistics of test sentences

	English			Dutch		
Effect	**RTP**	**CF**	**GI**	**RTP**	**CF**	**GI**
Original sentences	56	42	16	56	N/A	16
Unknown tokens (%)	11.1	0.0	2.1	13.8	N/A	0.8
Kept sentences	56	42	14	56	N/A	14

[2]We were only able to procure the sentences and reading times at a late stage in the project, after the networks had been trained. Therefore, we could not include them in our vocabulary.

Chapter 4

Results

In this chapter, we present the results. First, we will discuss the language model accuracy. This serves as an elementary check of the correctness of the implementation, and provides useful information on how well the models learn the languages. Subsequently, we will present the findings on the three probing tasks: reading time prediction, the cognate facilitation effect and the grammaticality illusion.

4.1 Language Model Accuracy

Here we analyse the models' language modelling quality as measured by the perplexity on the English and Dutch validation sets. The language model accuracy is inversely proportional to the perplexity – i.e. a lower perplexity means a higher language model accuracy. The perplexity results are shown for different training epochs and different layer sizes in fig. 4.1. Overall, the depicted trend is comparable on the English and Dutch validation sets. However in all settings the perplexity on the English validation set is roughly 2 times that on the Dutch validation set. This is a common result that arises from differences in the validation sets rather than difference in proficiency across languages.

A comparison between the bilingual and monolingual models reveals the monolingual models to have a better language model accuracy. This is to be expected, as the bilingual models are only trained on a subset of the data the monolingual models are trained on (see section 3.2.2), and additionally has to squeeze another language in the same representations. Furthermore, the balanced bilingual model outperforms the unbalanced model on the English validation set, but falls short on the Dutch validation set. This can be explained by the differences in language proportions – the balanced

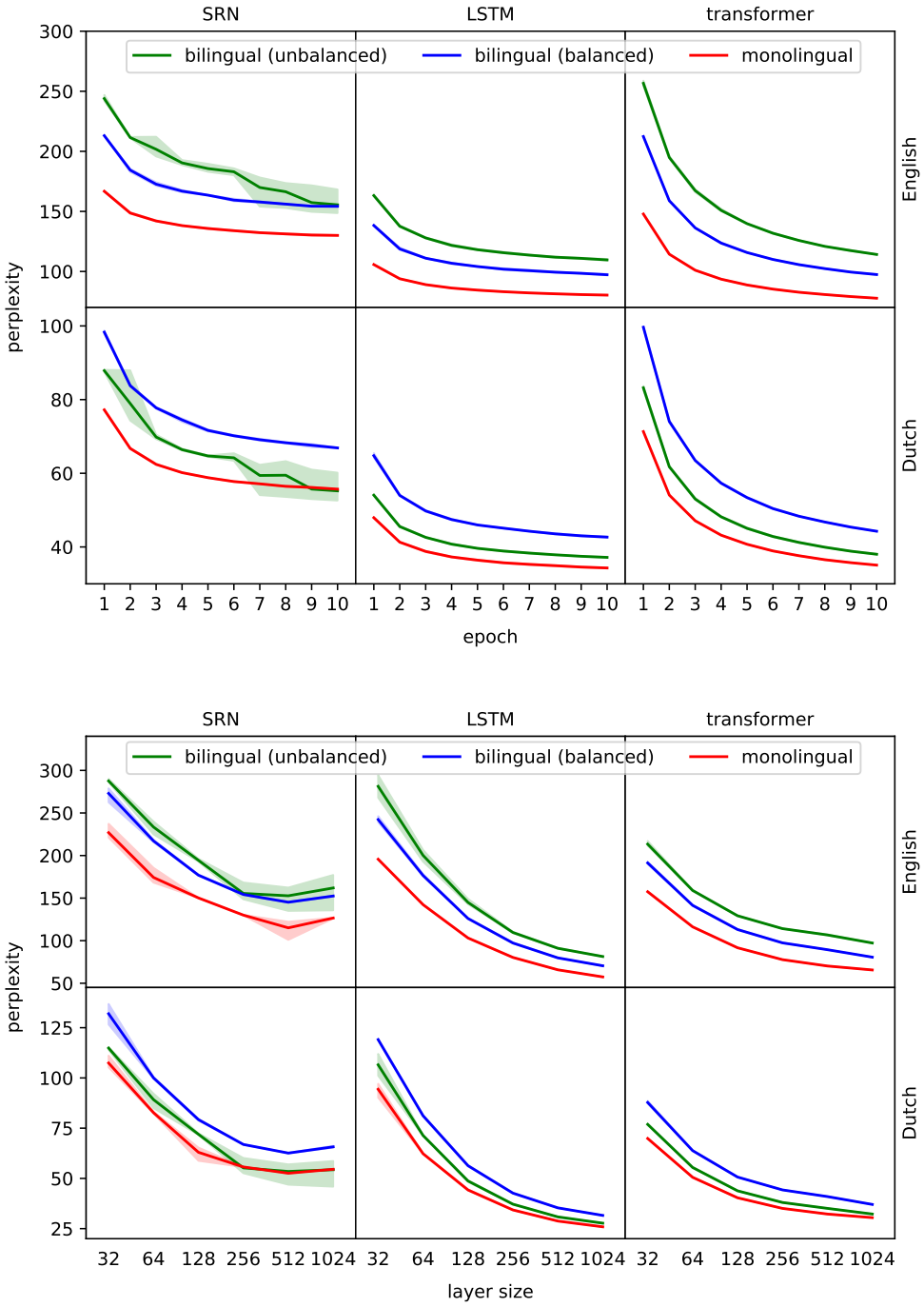

Figure 4.1: The perplexity on the English and Dutch validation sets. The mean perplexity over three seeds is depicted as lines, and the standard standard deviation over the seeds is depicted as shades. Top: perplexity as a function of training time for the models with an embedding size of 256. Bottom: perplexity as a function of embedding size after 10 epochs of training.

model is trained on more English data, but less Dutch data, and it is therefore natural for it to be more proficient in English and less proficient in Dutch.

Overall, the spread of the data, as indicated by the shaded regions, is relatively small. However, there are some exceptions to this, which is most evident in the unbalanced SRN with layer sizes from 256 to 1024, and from epochs 7 to 10. Further scrutiny revealed that this spread is caused by the learning rate scheduling. For these models, the validation set perplexity did not improve over some epoch, and the learning rate was therefore decreased as described in section 3.2.1. This decrease in learning rate resulted in a sudden drop of perplexity. Since this happened only for some random initializations, this is most evident in the graphs as a larger variation, although one can also see a drop in mean perplexity. Moreover, the LSTM displays less variation over random initialization, and the Transformer model exhibits the least variable behaviour.

After 10 epochs of training, the SRN performs worse than the other architectures (fig. 4.1). This is a common result from the literature (Hochreiter and Schmidhuber, 1997), as it can only effectively model short-term language dependencies. Furthermore, a comparison between the LSTM and Transformer reveals that the LSTM achieves better performance for large layer sizes, but worse performance for small layer sizes. The former trend may come across as surprising to some, as large Transformer models generally outperform LSTMs in the literature (Radford et al., 2018). This is likely to be due to two reasons. First, we do not make use of several implementation tricks that are common in the literature, as mentioned in section 3.2.1. Second, the limited training time means that the models do not reach convergence. This is especially true of the Transformer, whose learning curve over epoch is steeper than the other models. Therefore, it would be reasonable to expect the Transformer to perform better than the LSTM if it had been trained for longer.

Furthermore, the models generally profit from a larger embedding size. This is a common result in the literature (e.g. Radford et al., 2019), and can be explained by the fact that high-dimensional representations are more expressive than low-dimensional ones. An exception to this trend is provided by the SRN model with an embedding size of 1024. The loss in performance compared with the 512-dimensional SRN is likely to be due to the limited training time of 10 epochs, since larger networks require more training to reach best performance, but it could also result from the model overfitting to the training data.

Altogether, these results are in line with trends observed in the literature. This validates the correctness of our implementation, and gives us some useful information

for further analysis. In the next sections, we will look at how these models behave in a more targeted setting.

4.2 Reading Time Correlation

In this section, we are looking for evidence of agreement of our models with L1 and L2 reading times. This differs from the previous sections in that we are computing the correlation between surprisal values and reading times. Specifically, we are looking for settings that yield a higher goodness-of-fit for the bilingual model in English, and conversely, a higher goodness-of-fit for the monolingual model in Dutch.

The results are depicted in fig. 4.2, and averaged behaviour is found in tab. 4.1. Note that all of the goodness-of-fit values models take on positive values. This confirms that the surprisal values are informative for the fit. Furthermore, the Dutch goodness-of-fit values are generally lower than the English ones. This is likely to

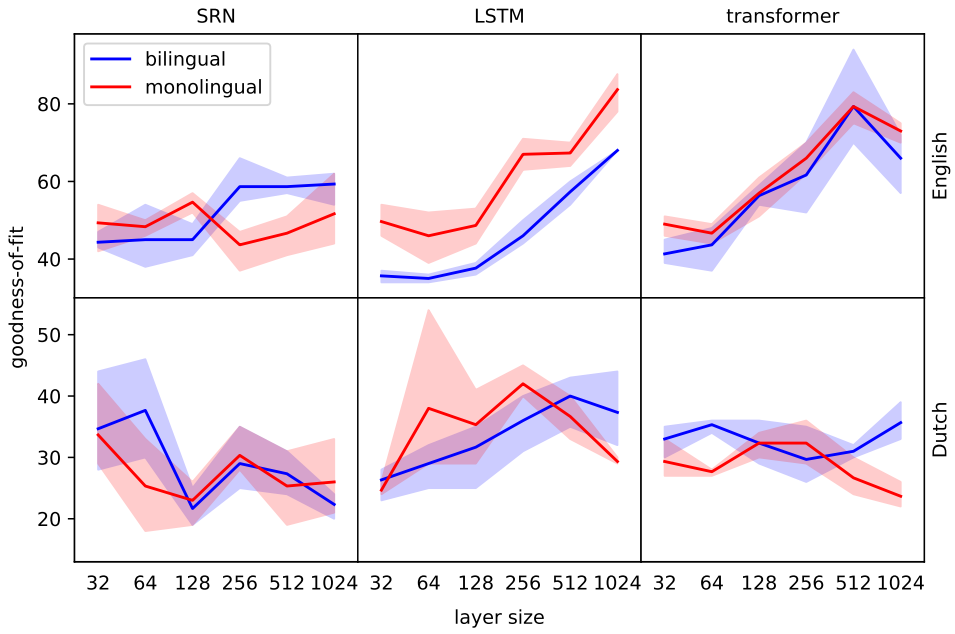

Figure 4.2: Goodness-of-fit as a function of layer size. The mean over three different seeds is depicted, with the shaded areas indicating the standard deviation over seeds. The top row expresses the values related to English reading times, and the bottom row those related to Dutch reading times. The goodness-of-fit is calculated by the difference in deviance of the baseline mixed model and the mixed model that includes surprisal values.

Table 4.1: Summary statistics of the average goodness-of-fit values. Each cell expressed the average value over layers size and random initialization. The bottom row additionally averages over languages.

SRN			LSTM			Transformer		
	Mono	**Bi**		**Mono**	**Bi**		**Mono**	**Bi**
English	1.05	3.83	English	12.38	−1.38	English	13.83	10.05
Dutch	−20.72	−19.22	Dutch	−13.66	−14.61	Dutch	−19.33	−15.16
Total	-9.83	-7.69	**Total**	−0.64	−8.00	**Total**	−2.75	−2.56

be due to pecularities in the respective languages – generally, Dutch is more morphologically rich than English, which poses a challenge for language modelling objectives as this results in less exposure to each word type.

For the specific settings used by Frank (2014), the effect seems to occur for specific random initializations. They used a SRN with layer size 200 and, to the best of out knowledge, only one random initialization. This agrees with the trend depicted in fig. 4.2.

However, this finding does not extend to multiple seeds, different layer sizes, or even different architectures. The LSTM consistently shows a higher goodness-of-fit for the monolingual model on L2 English reading times, and the Transformer shows a higher goodness-of-fit for the bilingual model on L1 Dutch reading times, for most layer sizes. This finding is surprising, especially since reading time correlation has been found to increase with language model quality (Aurnhammer and Frank, 2019; Merkx and Frank, 2020), and the LSTM and Transformer generally display better language modelling quality than the SRN.

Additionally, after correspondence with Stefan Frank, we were informed that they had unsuccessfully tried to reproduce the L1-L1 reading time correlation agreement in an LSTM with more and better data. A more plausible reason for the absence of L1/L2 agreement is thus perhaps simply that there is no noticeable difference in how well bilingual and monolingual language models correspond with reading times.

It is unlikely that the absence of the desired effect is due to errors in our implementation, as it shows several encouraging findings. First, the English goodness-of-fit scores increase with layer size, and by extension, with language model accuracy. This is again to be expected, and agrees with the findings presented in the literature (Aurn-

hammer and Frank, 2019). However, this trend is not present in the SRN, neither is it present in the goodness-of-fit values related to the Dutch reading times. Second, the LSTM and Transformer display a higher average goodness-of-fit score than the SRN (tab. 4.1). This is again encouraging, as these models have a higher language model accuracy than the SRN.

4.3 Cognate Facilitation Effect

In this section, we look for evidence in the cognate facilitation effect. Recall that for this purpose, cognates and non-cognates are substituted into English sentences, and their surprisal values are extracted. We look for evidence that cognates are significantly more easily processed than controls, which corresponds in the cognate effect size (eq. 3.2) being positive and significantly different to zero. As opposed to the grammaticality illusion, where the surprisal was computed on the same word type, each word type is now different.

The results are depicted in fig. 4.3. The fact that we are now dealing with different

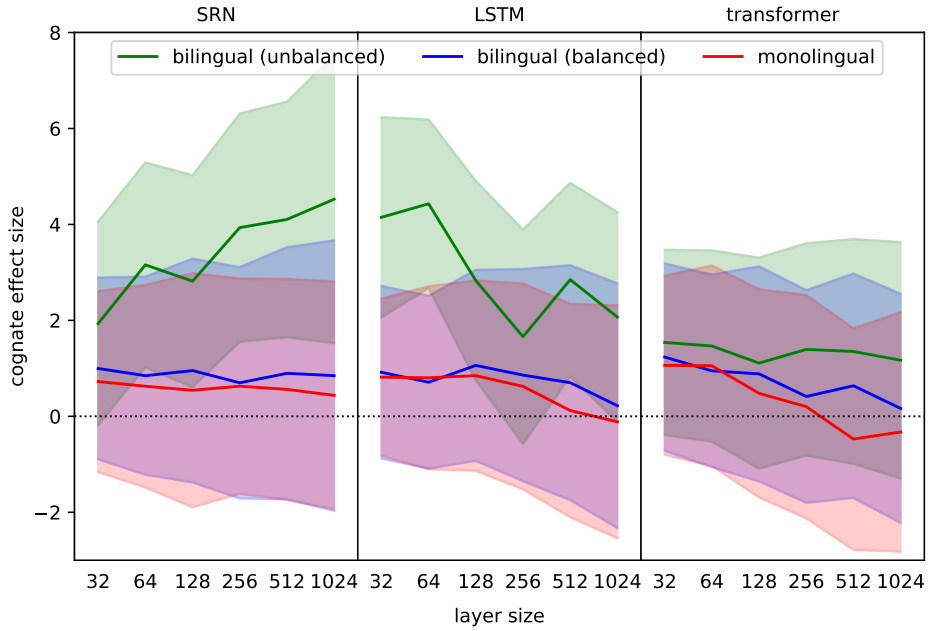

Figure 4.3: Cognate effect size as a function of layer size. Lines indicate the cognate effect averaged over items and seeds, and shaded areas indicate the standard error. Recall that the cognate effect is measured on English data and the monolingual model thus refers to the English monolingual model.

word types is reflected in the variation of the data – the standard error amounts to about 2-3 surprisal units as opposed to the standard error of about 1 surprisal units in the grammaticality illusion. Futhermore, note that the proportion in the training data is reflected in the cognate effect size – the effect is the smallest in the monolingual English model (0 % Dutch data), slightly higher in the balanced bilingual model (50 % Dutch data), and highest in the unbalanced bilingual model (75 % Dutch data).

Neither the monolingual model nor the balanced bilingual model suggest a statistically significant cognate effect. This was confirmed by an independent t-test as presented in tab. 4.2. This is an encouraging finding, asit agrees with the findings of Winther et al. (2021). In the unbalanced bilingual model however, the effect is present – at least for the LSTM and SRN architectures. This again agrees with the findings of Winther et al. (2021), who used an LSTM with a layer size of 650. This was confirmed by statistical analysis for the LSTM (15/18 models) and SRN (17/18 models).

Surprisingly, the cognate effect of the unbalanced bilingual model is less significantly present in the Transformer (2/18 models). The Transformer, in this sense, exhibits less linguistic transfer. This is likely to be a result of a higher separation of the language representations in the Transformer. This is likely to result from the way in which Transformers – by attending to each word in the context separately, it has more information about which language is being processed, as opposed to recurrent architectures, in which this information has to be encoded in a single hidden layer. However, this hypothesis needs further research to be confirmed.

An unexpected observation is that the cognate effect size seems to take a non-zero value for the monolingual models. As these models are only exposed to English data, they should treat cognates and non-cognates equally. A possible explanation for this is in terms of language model accuracy. Models with smaller hidden layers have a smaller language model accuracy, and are more reliant on simple statistical patterns such as word frequency. They are therefore more likely to reflect differences in word frequency in the training data. In order to investigate this, we computed the frequency of the cognates vs. the control words in the English training set, and found the frequency of the cognate words to be slightly higher. The sum of frequencies yielded 0.0021 vs. 0.0016, and a pair-wise comparison showed that 15 of 21 cognates had a higher frequency than the controls. The positive cognate effect in the monolingual models is likely to be due to this difference.

Another unexpected observation is that the cognate effect is primarily present for small layers in the SRN models, and for large layers LSTM models. The latter trend

Table 4.2: Number of models that exhibited a statistically significant cognate effect out of a total of 18 models. The test is an paired t-test with a confidence interval of 0.05 as described in section 3.3.1. “Bi-50” refers to the balanced bilingual model and “Bi-75” refers to the unbalanced bilingual model.

SRN			LSTM			Transformer		
Mono	**Bi-50**	**Bi-75**	**Mono**	**Bi-50**	**Bi-75**	**Mono**	**Bi-50**	**Bi-75**
0	0	17	0	0	15	0	0	2

may be expected, as a small hidden layer is more likely to use share linguistic representations in its hidden layer, and the increased frequency in cognate words in Dutch thus spills over to English. However, the increase in cognate effect with layer size in the SRN is unexpected.

In order these observations further, we decided to look at how well the cognate effect size correlates with difference in logarithmic frequency in the respective training data. The results are depicted in fig. 4.4. For the monolingual models, we observe a

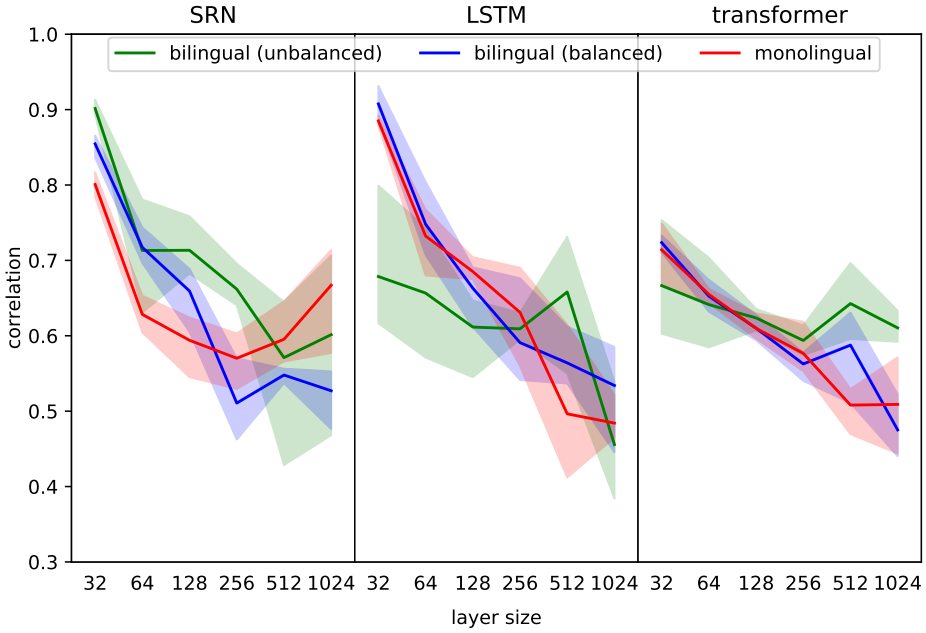

Figure 4.4: Pearson’s correlation between cognate effect size and logarithmic frequency in the respective training data as described in section 3.3.1. Lines indicate mean cognate effects, and shaded areas indicate the standard deviation over seeds.

decrease in correlation with larger layer size. This holds particularly in the case of the LSTM and Transformer, where the decrease in cognate effect over layer size is also the steepest (fig. 4.3). This reinforces the belief that the non-zero cognate effect size is due to a reliance on the higher frequencies of cognates in the training corpus.

The balanced bilingual models exhibit a behaviour similar to that of the monolingual models. Furthermore, whereas the correlation of the unbalanced bilingual models display an overall similar correlation to the monolingual models, it decreases less over layer size. A possible explanation for this is in terms of the order of presentation of the languages. Since testing is performed at the end of the 10th epoch, the unbalanced bilingual layer has seen exclusively English data when surprisal values are extracted. Therefore, they may adjust their predictions disproportionately to the English data. The amount of adjustment is likely to be larger in small layer sizes, as these are able to store less information and are thus more likely to forget in favour of new information.

The behaviour in fig. 4.4 serves as a possible explanation for the decrease of the cognate effect size for large layer sizes in the unbalanced bilingual LSTM. As the layer size increases, the model becomes more proficient in the languages and therefore less reliant on word frequency. However, this leaves the behaviour of the SRN unexplained.

4.4 Grammaticality Illusion

In this section, we present the results on the grammaticality illusion. Recall that the grammaticality illusion describes English sentences to be more easily processed when they are ungrammatical, and Dutch sentences to be more easily processed when they are grammatical. In our results, this corresponds to the *grammatical preference* being negative in English, and positive in Dutch.

The grammatical preference, depicted in fig. 4.5, is calculated as the difference between surprisal on the same word (“the” in English, and its translation equivalent “de” in Dutch) in the context of a grammatical and ungrammatical version of a sentence. The difference in surprisal value between grammatical and ungrammatical sentences thus does not rely on lexical differences in the words, but contextual differences in which the same word appears. This is in line with the observation that the grammatical preference tends toward zero for small layer sizes. Models with such small layer sizes are heavily reliant on superficial cues such as word frequency, and less able to integrate contextual information. Therefore, they do not differ significantly in their predictions on the same word in a different context.

We start out analysis with the SRN. Figure 4.5 suggests that the SRN qualitatively reproduces the grammaticality illusion, preferring English ungrammatical sentences and Dutch grammatical sentences. However, further statistical analysis shows that this effect is not consistent over random initializations. As depicted in tab. 4.3, only 15 of 36 monolingual, and 12 of 36 bilingual SRNs produce a statistically significant grammaticality effect in English or Dutch. Note that this is consistent with the findings of Frank et al. (2016), who only found a statistically significant grammaticality illusion on the Dutch test sentences for a specific random initialization.

Furthermore, the Transformer captures the negative grammaticality illusion in Dutch, as it consistently prefers grammatical sentences. Teh effect is statistically significant in 16 of 18 monolingual models and 15 of 18 bilingual models (tab. 4.3). However, it fails to capture the positive grammaticality illusion in English, as it prefers grammatical sentences also in this case. Statistical analysis reflects this – only 1 of 36 Transformer models showed a statistically significant positive grammaticality illusion (4.3). The strong preference for grammatical sentences in the Transformer is related to language model accuracy, as the preference for grammatical sentences grows with layer size, as does the language model accuracy. However, this is not the full explanation, as the

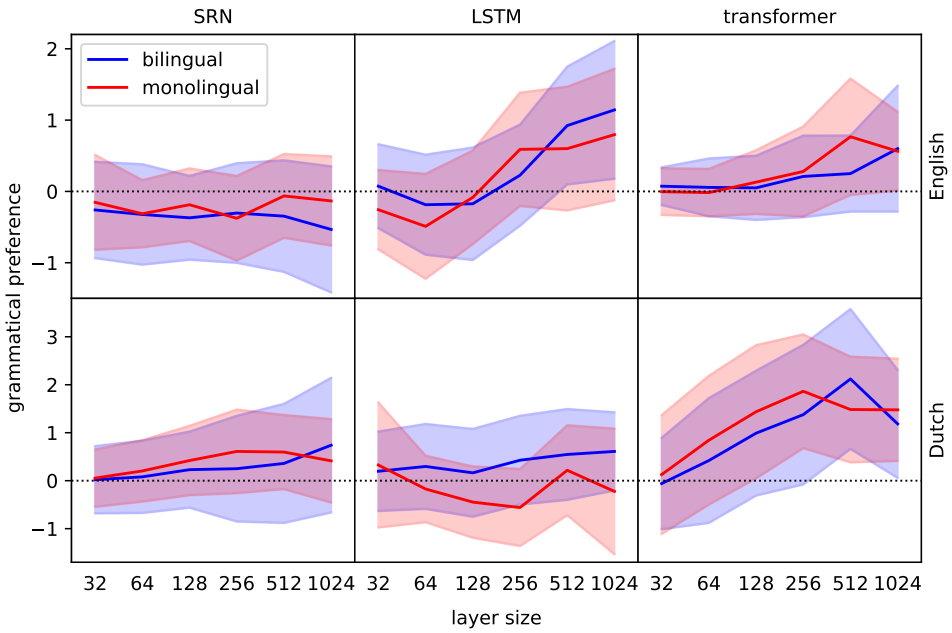

Figure 4.5: Grammatical preference (eq. 3.3) as a function of layer size. The mean values are depicted as lines, and the standard error. A positive grammatical preference corresponds to grammatical sentences being preferred, and vice versa.

Table 4.3: Summary statistics of the number of models that exhibited a significant grammaticality illusion. A paired t-test was used with confidence interval of 0.05. Each cell describes how many times the effect was observed in a specific language. Since we used 6 different layer sizes and 3 different seeds, the maximum number of models is 18 per cell, and the total is taken from 36 models.

SRN		
	Mono	**Bi**
English	4	8
Dutch	11	4
Total	15	12

LSTM		
	Mono	**Bi**
English	4	1
Dutch	2	4
Total	6	5

Transformer		
	Mono	**Bi**
English	1	0
Dutch	16	15
Total	17	15

language model accuracy of the Transformer is similar to that of the SRN and LSTM, who do not consistently prefer grammatical sentences. Rather, this is likely to be due to the way information is integrated in the Transformer. Since it attends to each word individually, it is more capable of modelling syntactic agreement, as has bee shown in the literature by, for example, Mueller et al. (2020).

Finally, the grammatical preference in the LSTM is differs between languages. On the English sentences, it prefers the ungrammatical version for smaller layer sizes, successfully modelling the English grammaticality illusion in some cases – statistical analysis reveals a significant effect in 4 of 18 monolingual models, and in 1 of 18 bilingual models. However, for larger embedding sizes, it fails to capture the grammaticality illusion, preferring the grammatical versions of the sentences (fig. 4.5). A possible explanation for this is again related to language model accuracy. For small layer sizes, the language model accuracy is low, and the LSTM relies heavily on short-term language statistics, as these are easier to model. The behaviour is thus similar to the SRN. For large layer sizes however, the language model accuracy increases, allowing the model to capture long-term dependencies, which results in it preferring the grammatical version of the sentence. Furthermore, on the Dutch sentences, the behaviour is inconsistent – the monolingual model exhibits a slight preference for ungrammatical sentences, and the bilingual model seems to prefer grammatical sentences. This is unexpected, and possibly results from to particularities in the training and test sets.

A comparison between the monolingual and bilingual implementations suggests that the monolingual implementations captures the grammaticality illusion better (tab.

4.3). However, this difference is small, as the monolingual and bilingual implementations exhibit similar values of the grammatical preference (fig. 4.5). The similarity tells us two things about the implementations. First, the smaller exposure to each individual language in the bilingual implementation does not have a large effect on its behaviour on this task. Second, this suggests that the amount of syntactic transfer is small in the bilingual models.

In summary, the SRN is able to capture the grammaticality illusion, as opposed to the LSTM and Transformer models. This behaviour confirms our hypothesis and hints toward the hypothesis that the grammaticality illusion is primarily a result of short-term language statistics. Although statistical analysis showed the effect to be heavily dependent on the specific random initialization, we showed that the effect does appear to be consistent when averaging over different initializations.

Chapter 5

Conclusions

In this work, we reproduced baseline results of three studies that use neural language models trained on two languages to model psycholinguistic effects of bilingualism: reading time prediction (Frank, 2014), the cognate facilitation effect (Winther et al., 2021) and the grammaticality illusion (Frank et al., 2016). We extended these baselines in two directions – by considering different architectures, and different layer sizes. We saw that the emergence of these effects depended crucially on architecture and hidden layer size. In the rest of this section, we will analyze how this dependency played out in more detail by first looking at each effect individually, and subsequently perform a general analysis.

5.1 Reading Time Prediction

We tried to replicate the results from Frank (2014), but with two modeifications – a model extension over layer sizes and architectures, and a more sophisticated evaluation procedure, inspired by recent work on reading time prediction. However, we found inconsistent evidence of their results, even when using the same hyperparameters as the original study. After correspondence with the author, they informed us that they too had been unsuccessful at reproducing the results, using a LSTM network and more reading data of higher quality (Stefan Frank, personal communication, June 22, 2021). This suggests the results from the original study to have been a coincidence.

One can imagine the monolingual and bilingual language models as embodying two points on a spectrum. The monolingual model has no exposure to the other language, and represents L1 processing. The bilingual model, however, is exposed to another language and can model the other language. The evidence presented in this

work is inconsistent in which one of these models better represents bilingual L2 reading.

It is possible that the ambiguity of our findings results from L2 reading lying somewhere between the monolingual and (balanced) bilingual models. This could be tested for by training an unbalanced language model with less exposure to L1 than L2. As our unbalanced model is unbalanced in the other direction (is has more exposure to L1 Dutch data than L2 English data), it is unlikely that it would be superior to the balanced bilingual model.

5.2 Cognate Effect

We managed to reproduce the findings of Winther et al. (2021) – in the recurrent networks, the cognate effect was absent in the monolingual and balanced bilingual models, but present in the unbalanced bilingual models. We extended their findings by showing that the cognate effect depends on architecture – it was not present in the Transformer – and on layer size.

In order to gain insight into our results, it may be helpful to ask why we would expect a language model to exhibit the cognate effect in the first place. An ideal bilingual language model, i.e. one that approximates the true probability distribution over words $P(w_i|w_{<i})$ perfectly, would not exhibit the cognate effect as the English probability distribution over words should be independent of the frequency of (cognate) words in Dutch. This requires language-specific information to be extracted from the context and effectively integrated when predicting the next word. Furthermore, since cognates are forced to have the same embedding in both English and Dutch, the only language-identical knowledge that the language model possesses comes from the context. The size of the cognate effect is thus related to how contextual language-identical information is integrated with a cognate's embedding. The Transformer is thus doing a better job than the recurrent networks at separating between the languages.

Similarly, the dependency of the cognate effect with layer size is surprising – for the SRN, it increases with layer size, but the trend is reversed in the LSTM. Which of these behaviours is more plausible? One hypothesis is that the behaviour seen in the LSTM is related to the amount of language-specific information that can be stored in each layer. Since this amount increases with layer size, the network is equipped to separate between the languages, resulting in a smaller cognate effect.

In order to test such a hypothesis, one could measure the amount of language iden-

tical information in the model by looking at its activations. As mentioned in section 3.1, we tried this using diagnostic classifiers in preliminary experiments, but found the method to be inherently flawed.

5.3 Grammaticality Illusion

Finally, we reproduced the grammaticality illusion from (Frank et al.) in the SRN, but not in the LSTM and Transformer. This provides further evidence for the hypothesis This supports the hypothesis presented in Frank et al. (2016), i.e. that the grammaticality illusion arises as an effect of language statistics rather than working memory contraints. Because English subordinate clauses have a Subject-Verb-Object structure, whereas Dutch subordinate clauses have an Subject-Object-Verb structure, it is much more common for sentences to contain three subsequent verbs in Dutch than in English. The Transformer and LSTM models failed to model this effect, as they are more effective at capturing long-range dependencies.

In order to further test the short-term language statistics hypothesis, our work could be extended by training a feed-forward language model with different context lengths. This would allow to isolate short-term effects by testing if shorter context lengths indeed result in a stronger grammaticality illusion. A cheaper alternative would be given by an n-gram language model, but as n-grams do not generalize well for unseen words n-grams, this is unlikely to work.

Furthermore, the Transformer displayed a consistent preference for grammatical sentences, regardless of language. A possible reason for this is that it captures syntactic agreement in the sentences. The superior syntactic properties of Transformers has found evidence in the literature (Mueller et al., 2020), however to our knowledge they have not been tested explicitly on subordinate clauses, let alone double-embedded subordinate clauses. This could be tested for in future work by explicitly looking at how strongly it attends to each of the words in the context.

5.4 Overall Analysis

Altogether, the Transformer model seems to behave too well as a language model to model human behaviour, exhibiting little lexical transfer and super-human syntactic abilities. This stands in contrast to recent evidence that claim Transformers to relate more strongly than recurrent models to psychometric data (Merkx and Frank, 2020).

Attention-based networks and recurrent networks represent two extreme views on sentence processing, and it is possible for the best model to be a mixture between the two.

The SRN and LSTM to behaved more similarly in our tasks. This is particularly true for the LSTMs with small layer sizes and SRNs with large layer sizes (cf. figs. 4.3 and 4.5). This suggests there to be little difference between the two as models of psycholinguistics beyond their language model accuracies. This is supported by the literature, where they show no difference in how well they predict psychometric data (Merkx and Frank, 2020).

A comparison between the monolingual and balanced bilingual models show little difference. This is perhaps most true in reading time prediction, but holds as a general trend in the other tests as well. This suggests there to be little overall transfer in balanced bilingual models – they learn to distinguish between the languages well and behave like two monolingual models in one model. This stands in contrast to the unbalanced bilingual model, suggesting it to be a better model of human bilingual behaviour.

Bibliography

Aurnhammer, Christoph and Stefan L Frank (2019). "Evaluating information-theoretic measures of word prediction in naturalistic sentence reading". In: *Neuropsychologia* 134, p. 107198.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473*.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2014). "Fitting linear mixed-effects models using lme4". In: *arXiv preprint arXiv:1406.5823*.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin (2003). "A neural probabilistic language model". In: *The journal of machine learning research* 3, pp. 1137–1155.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300.

Christiansen, Morten H and Maryellen C MacDonald (2009). "A usage-based approach to recursion in sentence processing". In: *Language Learning* 59, pp. 126–161.

Costa, Albert, Alfonso Caramazza, and Nuria Sebastian-Galles (2000). "The cognate facilitation effect: implications for models of lexical access." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26.5, p. 1283.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Dhar, Prajit and Arianna Bisazza (2020). "Understanding Cross-Lingual Syntactic Transfer in Multilingual Recurrent Neural Networks". In: *arXiv preprint arXiv:2003.14056*.

Dijkstra, Ton, Jonathan Grainger, and Walter JB Van Heuven (1999). "Recognition of cognates and interlingual homographs: The neglected role of phonology". In: *Journal of Memory and language* 41.4, pp. 496–518.

Dijkstra, Ton and Walter JB Van Heuven (2002). "The architecture of the bilingual word recognition system: From identification to decision". In: *Bilingualism: Language and cognition* 5.3, pp. 175–197.

Dijkstra, Ton, Alexander Wahl, Franka Buytenhuijs, Nino Van Halem, Zina Al-Jibouri, Marcel De Korte, and Steven Rekké (2019). "Multilink: a computational model for bilingual word recognition and word translation". In: *Bilingualism: Language and Cognition* 22.4, pp. 657–679.

Duyck, Wouter, Eva Van Assche, Denis Drieghe, and Robert J Hartsuiker (2007). "Visual word recognition by bilinguals in a sentence context: evidence for nonselective lexical access." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33.4, p. 663.

Elman, Jeffrey L (1990). "Finding structure in time". In: *Cognitive science* 14.2, pp. 179–211.

Frank, Stefan L (2014). "Modelling reading times in bilingual sentence comprehension". In:

Frank, Stefan L, Leun J Otten, Giulia Galli, and Gabriella Vigliocco (2015). "The ERP response to the amount of information conveyed by words in sentences". In: *Brain and language* 140, pp. 1–11.

Frank, Stefan L, Thijs Trompenaars, and Shravan Vasishth (2016). "Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics?" In: *Cognitive Science* 40.3, pp. 554–578.

Frazier, Lyn (1985). "Syntactic complexity". In: *Natural language parsing: Psychological, computational, and theoretical perspectives*, pp. 129–189.

French, Robert M (1998). "A simple recurrent network model of bilingual memory". In: *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, pp. 368–373.

Futrell, Richard, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy (2019). "Neural language models as psycholinguistic subjects: Representations of syntactic state". In: *arXiv preprint arXiv:1903.03260*.

Gawinkowska, Marta, Michał B Paradowski, and Michał Bilewicz (2013). "Second language as an exemptor from sociocultural norms. Emotion-related language choice revisited". In: *PloS one* 8.12, e81225.

Grosjean, François and Ping Li (2013). *The psycholinguistics of bilingualism*. John Wiley & Sons.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni (2018). "Colorless green recurrent networks dream hierarchically". In: *arXiv preprint arXiv:1803.11138*.

Hamilton, Helen W and James Deese (1971). "Comprehensibility and subject-verb relations in complex sentences". In: *Journal of Verbal Learning and Verbal Behavior* 10.2, pp. 163–170.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Keysar, Boaz, Sayuri L Hayakawa, and Sun Gyu An (2012). "The foreign-language effect: Thinking in a foreign tongue reduces decision biases". In: *Psychological science* 23.6, pp. 661–668.

Li, Ping and Igor Farkas (2002). "3 A self-organizing connectionist model of bilingual processing". In: *Advances in psychology*. Vol. 134. Elsevier, pp. 59–85.

Libben, Maya R and Debra A Titone (2009). "Bilingual lexical access in context: evidence from eye movements during reading." In: *Journal of Experimental Psychology: Learning, memory, and cognition* 35.2, p. 381.

Libovick, Jindřich, Rudolf Rosa, and Alexander Fraser (2020). "On the language neutrality of pre-trained multilingual representations". In: *arXiv preprint arXiv:2004.05160*.

Mechelli, Andrea, Jenny T Crinion, Uta Noppeney, John O'Doherty, John Ashburner, Richard S Frackowiak, and Cathy J Price (2004). "Structural plasticity in the bilingual brain". In: *Nature* 431.7010, pp. 757–757.

Merkx, Danny and Stefan L Frank (2020). "Comparing Transformers and RNNs on predicting human sentence processing data". In: *arXiv e-prints*, arXiv–2005.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Mikolov, Tomáš, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur (2010). "Recurrent neural network based language model". In: *Eleventh annual conference of the international speech communication association*.

Mueller, Aaron, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen (2020). "Cross-linguistic syntactic evaluation of word prediction models". In: *arXiv preprint arXiv:2005.00187*.

Pires, Telmo, Eva Schlinger, and Dan Garrette (2019). "How multilingual is multilingual BERT?" In: *arXiv preprint arXiv:1906.01502*.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). "Improving language understanding by generative pre-training". In:

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8, p. 9.

Rayner, Keith (1998). "Eye movements in reading and information processing: 20 years of research." In: *Psychological bulletin* 124.3, p. 372.

Smith, Nathaniel J and Roger Levy (2013). "The effect of word predictability on reading time is logarithmic". In: *Cognition* 128.3, pp. 302–319.

Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney (2012). "LSTM neural networks for language modeling". In: *Thirteenth annual conference of the international speech communication association*.

Van Assche, Eva, Wouter Duyck, Robert J Hartsuiker, and Kevin Diependaele (2009). "Does bilingualism change native-language reading? Cognate effects in a sentence context". In: *Psychological science* 20.8, pp. 923–927.

Van Hell, Janet G and Ton Dijkstra (2002). "Foreign language knowledge can influence native language performance in exclusively native contexts". In: *Psychonomic bulletin & review* 9.4, pp. 780–789.

Van Schijndel, Marten and Tal Linzen (2018). "Modeling garden path effects without explicit hierarchical syntax." In: *CogSci*.

Vasishth, Shravan, Katja Suckow, Richard L Lewis, and Sabine Kern (2010). "Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures". In: *Language and Cognitive Processes* 25.4, pp. 533–567.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in neural information processing systems*, pp. 5998–6008.

Voga, Madeleine and Jonathan Grainger (2007). "Cognate status and cross-script translation priming". In: *Memory & cognition* 35.5, pp. 938–952.

Wilcox, Ethan Gotlieb, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy (2020). "On the predictive power of neural language models for human real-time comprehension behavior". In: *arXiv preprint arXiv:2006.01912*.

Winther, Irene Elisabeth, Yevgen Matusevych, and Martin John Pickering (2021). "Cumulative Frequency Can Explain Cognate Facilitation in Language Models". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 43.

Zhao, Xiaowei and Ping Li (2010). “Bilingual lexical interactions in an unsupervised neural network model”. In: *International Journal of Bilingual Education and Bilingualism* 13.5, pp. 505–524.

Appendix A

Cognate Effect Items

1. He convinces her to buy the art at the department store in town .
2. He convinces her to buy the bed at the department store in town .
3. He does not like to talk about the error out of a sense of guilt .
4. He does not like to talk about the drama out of a sense of guilt .
5. I barely recognized the witch on the black and white cover of the magazine .
6. I barely recognized the title on the black and white cover of the magazine .
7. My uncle calls the editor to get the announcement on the exhibition .
8. My uncle calls the museum to get the announcement on the exhibition .
9. The attorney consults a lawyer for a detailed opinion on the matter .
10. The attorney consults an expert for a detailed opinion on the matter .
11. The boys ask permission to use the spoon and then suddenly turn around .
12. The boys ask permission to use the motor and then suddenly turn around .
13. The campaigners halt before the window and do not dare to move further .
14. The campaigners halt before the student and do not dare to move further .
15. The children visit the farm on their annual outing to Germany .
16. The children visit the race on their annual outing to Germany .
17. The flyers contact their airport before taking off on a six-hour flight .
18. The flyers contact their partner before taking off on a six-hour flight .
19. The fortune-tellers know the destiny of the wealthy gentlemans fiancé .

20. The fortune-tellers know the dilemma of the wealthy gentlemans fiancé .
21. The governor worries about the safety of the big aircraft after the crash .
22. The governor worries about the status of the big aircraft after the crash .
23. The hostess discovers the liar in the kitchen behind the wall .
24. The hostess discovers the menu in the kitchen behind the wall .
25. The inspectors review the case thoroughly to pinpoint their mistakes .
26. The inspectors review the week thoroughly to pinpoint their mistakes .
27. The ladies watch the bottle in the cupboard with great interest .
28. The ladies watch the detail in the cupboard with great interest .
29. The officers catch the fear of the burglar as he reaches for his knife .
30. The officers catch the hand of the burglar as he reaches for his knife .
31. The painting depicts the city from above in a beautiful manner .
32. The painting depicts the baby from above in a beautiful manner .
33. The parents are surprised by the fairy in their childrens bedroom .
34. The parents are surprised by the chaos in their childrens bedroom .
35. The participants submit the vote for the contest at the festival .
36. The participants submit the film for the contest at the festival .
37. The residents dislike the prison for the trouble experienced in the past .
38. The residents dislike the winter for the trouble experienced in the past .
39. The superiors invite the unit for a short briefing in the office .
40. The superiors invite the team for a short briefing in the office .
41. The wives send a reply to their sick husbands in the nursery home .
42. The wives send a plant to their sick husbands in the nursery home .

Appendix B

Grammaticality Illusion Items

1. The carpenter who the craftsman who the peasant carried [hurt] supervised the apprentice in the garden .
2. The mother who the daughter who the sister found [frightened] greeted the grandmother on the tricycle .
3. The worker who the tenant who the foreman looked for [injured] questioned the shepherd in the office .
4. The trader who the businessman who the professor hired [confused] annoyed the investor in the morning .
5. The painter who the musician who the father missed [sheltered] cooked for the artist in the kitchen .
6. The saxophonist who the trumpeter who the conductor brought along [distracted] thanked the violinist in his speech .
7. The pharmacist who the optician who the stranger saw [troubled] questioned the customer at the counter .
8. The cleaner who the janitor who the doctor recognized [hurt] surprised the patient in the hallway .
9. The dancer who the singer who the bystander admired [hurt] tipped the doorman at the door .
10. The artist who the sportsman who the guard shouted at [annoyed] instructed the newscaster in the studio .
11. The clerk who the bureaucrat who the visitor forgot about [helped] annoyed the neighbor at the town hall .

12. The son who the father who the teacher saw [disturbed] visited the grandfather in the nursing home .
13. The conductor who the choirmaster who the worker ignored [hit] berated the musician at the festival .
14. The defence who the prosecutor who the spy looked at [surprised] convinced the judge in the courtroom .
15. The cousin who the brother who the peasant described [pleased] hated the uncle from the farm .
16. The painter who the musician who the friend liked [disturbed] admired the poet in the pyjamas .
17. De timmerman die eergisteren de vakman die zaterdag de boer droeg [bezeerde] begeleidde de leerling in de tuin .
18. De moeder die vrijdag de dochter die toen de zus vond [beangstigde] begroette de oma op de driewieler .
19. De arbeider die onlangs de huurder die toen de voorman zocht [verwondde] ondervroeg de herder in het kantoor .
20. De handelaar die net de zakenman die destijds de professor inhuurde [verwarde] irriteerde de investeerder in de ochtend .
21. De schilder die laatst de muzikant die al zo lang de vader miste [beschutte] kookte voor de kunstenaar in de keuken .
22. De saxofonist die zondag de trompettist die altijd de dirigent meebracht [afleidde] bedankte de violist in een toespraak .
23. De apotheker die woensdag de opticien die gisteren de vreemdeling zag [verontrustte] ondervroeg de klant aan de balie .
24. De schoonmaker die vanochtend de conciërge die snel de dokter herkende [verwondde] verraste de patiënt in de gang .
25. De danser die gisteren de zanger die laatst de toeschouwer bewonderde [bezeerde] beloonde de portier met een tientje .
26. De artiest die vanmorgen de sportman die soms de bewaker riep [irriteerde] instrueerde de nieuwslezer in de studio .
27. De klerk die gisteren de ambtenaar die soms de bezoeker vergat [hielp] irriteerde de buurman op het stadhuis .

28. De zoon die laatst de vader die vaak de leraar zag [stoorde] bezocht de grootvader in het bejaardenhuis .
29. De conducteur die vanmorgen de dirigent die altijd de arbeider negeerde [sloeg] bekritiseerde de musicus op het festival .
30. De verdediging die vanmiddag de aanklager die eventjes de spion aankeek [verraste] overtuigde de rechter in de rechtbank .
31. De neef die eergisteren de zus die ooit de boer in detail beschreef [plezierde] haatte de oom na de familieruzie .
32. De schilder die toentertijd de muzikant die ooit een vriend sloeg [stoorde] bewonderde de dichter op het poëziefestival .

Appendix C

Statistical Significance of Grammaticality Illusion

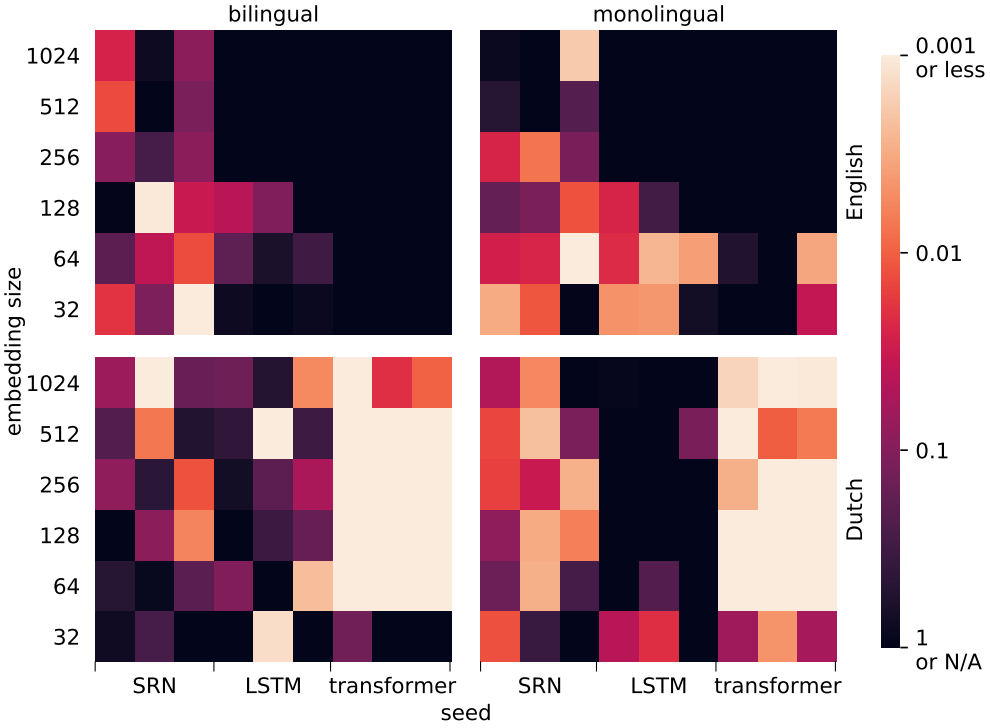

Figure C.1: A heat map of the p-values related to the grammaticality illusion. The p-values are obtained by a paired t-test. Each row represents a specific model and seed, for a total of 3 seeds. The darkest color represents p-values near one, or settings where the opposite behaviour was observed (i.e. where the model preferred English grammatical sentences or Dutch ungrammatical sentences).