Modeling Sentence Processing Effects in Bilingual Speakers: A Comparison of Neural Architectures

Rasmus Roslund (roslund.rasmus@gmail.com)
School of Informatics
University of Edinburgh

Yevgen Matusevych (yevgen.matusevych@ed.ac.uk)
School of Informatics; School of Philosophy, Psychology and Language Sciences
University of Edinburgh

Abstract

Neural language models are commonly used to study language processing in human speakers, and several studies trained such models on two languages to simulate bilingual speakers. Surprisingly, no work systematically evaluates different neural architectures on bilingual speakers' data, despite the abundance of such studies in the monolingual domain. In this work, we take the first step in this direction. We train three neural architectures (SRN, LSTM, and Transformer) on Dutch and English data and evaluate them on two data sets from experimental studies. Our goal is to investigate which architectures can reproduce the cognate facilitation effect and grammaticality illusion observed in bilingual speakers. While all three architectures can correctly predict the cognate effect, only the SRN succeeds at the grammaticality illusion. We additionally show how the observed patterns change as a function of the models' hidden layer size, a hyperparameter that we argue may be more important in bilingual models.

Keywords: language modeling, neural networks, bilingualism, cognate facilitation, grammaticality illusion

Introduction

In recent years, neural language models (LMs) have been extensively used to study human sentence processing. Such models have been shown to significantly predict human reading times (e.g., Frank et al., 2015; Goodkind & Bicknell, 2018) and exhibit various processing effects observed in human speakers, such as number agreement (Mueller et al., 2020) and garden-path effects (Van Schijndel & Linzen, 2018; Futrell et al., 2019). These models even make some syntactic errors similar to those of human speakers (Linzen & Leonard, 2018). The use of language models in psycholinguistics has become so pervasive that online platforms for their automatic testing have been designed (Gauthier et al., 2020), including tools that introduce human-in-the-loop evaluation (Kiela et al., 2021).

In psycholinguistics, neural LMs are commonly trained on data from a single language and evaluated on data from human speakers, to test whether they make good models of (monolingual) sentence processing. By contrast, little work exists that considers models trained on two (or more) languages in the context of bilingual speakers' sentence processing. This is surprising, both because the use of multilingual language models is very common in NLP applications (Devlin et al., 2019; Guo et al., 2020), and because it has long been argued that the field of bilingualism and bilingual sentence processing in particular could benefit from having more formal models (Frank, 2021; Li, 2002).

Neural LMs trained on two languages *have* been used to study various effects in the field of bilingualism, such as the grammaticality illusion (Frank et al., 2016), gender pronoun errors (Tsoukala et al., 2017), cognate facilitation (Winther et al., 2021), code-switching (Tsoukala et al., 2021), crosslinguistic structural priming (Khoe et al., 2021) and sentence-level reading times (Frank, 2014). These studies focus on a single effect of interest and look at the ability of a particular architecture to correctly predict that effect, making it difficult to say which architectures make better models of bilingual sentence processing. Moreover, these studies tend to use models such as simple recurrent networks or networks trained on miniature languages, and not the modern architectures common in the monolingual domain, such as LSTMs or Transformers (but see Winther et al., 2021). There is thus a lack of work comparing the architectures of bilingual neural LMs in terms of their ability to predict a variety of effects in bilingual processing, as has been done for monolingual models (e.g., Merkx & Frank, 2021; Wilcox et al., 2020).

In this study, we take the first steps in this direction. We focus on three neural architectures that have been commonly used in studies on monolingual sentence processing: Simple Recurrent Networks (SRNs; Elman, 1990), Long Short-Term Memory networks (LSTMs; Hochreiter & Schmidhuber, 1997), and Transformers (Vaswani et al., 2017). We train these models on two languages, Dutch and English, in parallel and test them against the available human data on two effects from the domain of bilingual sentence processing, the cognate facilitation effect and the grammaticality illusion, which we present in more detail in the next section. Existing studies (Winther et al., 2021; Frank et al., 2016) showed that these effects can be predicted by neural LMs in principle, and here we extend the existing results by testing each of the three architectures on both effects. Our main goals are to test whether the results reported in the two studies mentioned above can be replicated across model architectures, and, by extension, whether one of them makes a better model of bilingual sentence processing.

In addition, we consider whether and how the models' ability to predict human-like patterns depends on their hidden layer size. While it is not uncommon to experiment with various hidden layer sizes (e.g., Gulordava et al., 2018), we believe that this parameter may be more important in the case of models trained on two languages, because a model with a large hidden layer may be able to separate the representations

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

for the two languages, reducing the amount of cross-linguistic influence (i.e., the extent to which the representations from the two languages are shared in the model's hidden layers). We test whether this is the case by systematically running our experiments with various hidden layer sizes for each model.

To preview our findings, all three architectures could correctly predict the cognate facilitation effect, while only the bilingual SRN model was successful at reproducing the grammaticality illusion in both languages. The magnitude of the two effects depended on the layer size to a greater extent in some bilingual models than in the monolingual models, especially so for the cognate facilitation effect.

Background

Processing effects in bilingual speakers

We focus on two effects in bilingual sentence processing – cognate facilitation and grammaticality illusion – because the existing studies (Winther et al., 2021; Frank et al., 2016) showed that neural LMs trained on two languages could correctly predict these effects in principle.

Cognate facilitation effect. Cognates are words that share their form and meaning across two languages, such as the word *drama* in Dutch and English. A common finding in the literature on bilingualism is that they are processed faster than non-cognates (Costa et al., 2000; Dijkstra et al., 1999). In particular, Bultena et al. (2014) showed that Dutch–English bilingual speakers were faster at reading sentences with cognates than with non-cognates, as in (1), where *drama* is a cognate word, but *error* (Dutch 'fout') is not. Note that the words in each pair were matched on various characteristics including corpus frequency, to avoid potential confounds.

(1) He does not like to talk about the *drama*/*error* out of a sense of guilt.

Winther et al. (2021) investigated if neural LMs could reproduce this effect by training bilingual LSTMs on English and Dutch data and testing them on Bultena et al.'s sentences. They found evidence for the cognate effect under certain conditions of input presentation, namely when the models were trained so as to simulate unbalanced/sequential bilingual speakers, through exposing the models first to Dutch and then to English input data, with more Dutch exposure overall.

Grammaticality illusion. Speakers normally consider grammatical sentences to be more acceptable than ungrammatical ones. However, English native speakers have been consistently found to judge ungrammatical derivations of a certain class of sentences as more acceptable than their grammatical equivalents (Frazier, 1985; Christiansen & MacDonald, 2009). This effect is observed in sentences containing double-nested relative clauses, as in English (2) and Dutch (3).

(2) The carpenter *who the craftsman* **who the peasant carried** *hurt* supervised the apprentice in the garden.
V1 V2 V3 D1

(3) De timmerman *die eergisteren de vakman* **die zaterdag de boer droeg** *bezeerde* begeleidde de leerling in de tuin.
V1 V2 V3 D1

The addition of the first relative clause (in italics) and the second relative clause (in bold) nested within the first clause makes this sentence difficult to process, although still grammatical. As a result, English speakers tend to process more easily an ungrammatical version in which the second verb V2, *hurt*, is omitted. This effect, referred to as the grammaticality illusion, has been shown using acceptability ratings (Christiansen & MacDonald, 2009) and reading times on the first determinant after the third verb, D1[1] (Vasishth et al., 2010).

Frank et al. (2016) carried out a similar test in Dutch with native Dutch speakers and found that they, unlike English speakers, process the grammatical version more easily. Interestingly, when the same Dutch speakers were tested in their second language (L2) English, they behaved as native English speakers and read the ungrammatical version more quickly. In their study, Frank et al. also tested if a neural LM could reproduce this grammaticality illusion effect across languages. They trained a bilingual SRN model on Dutch and English data and observed a human-like behavior when the model was tested specifically at D1: it preferred ungrammatical English sentences but grammatical Dutch sentences.

Neural LMs in monolingual sentence processing

In the monolingual context, a number of studies compare how well different neural LMs predict human reading data such as self-paced reading times and event-related potentials in the brain. SRNs and GRUs (Gated Recurrent Units; Cho et al., 2014) have been reported to be equally successful, provided the language model accuracy is accounted for (Aurnhammer & Frank, 2019). The same is likely to hold between SRNs and LSTMs, since LSTMs and GRUs both include gates to control the flow of information in a similar fashion. Furthermore, Transformers generally outperform recurrent architectures (Merkx & Frank, 2021; Wilcox et al., 2020).

Another line of research evaluates neural LMs in more targeted settings, to measure their grasp of various syntactic phenomena (Linzen et al., 2016; Gulordava et al., 2018; Marvin & Linzen, 2018; Wilcox et al., 2018; Warstadt et al., 2020). The results indicate LSTMs to be superior to SRNs (Linzen et al., 2016), and Transformers to be superior to recurrent networks (Mueller et al., 2020).

Methods

Our general approach is to train language models on text corpora from one or two languages (Dutch and/or English) and evaluate their performance on two sets of sentences, in order to test whether they exhibit the cognate facilitation effect and the grammaticality illusion observed in human speakers.

[1] Here, we follow the numeric notation of Frank et al. (2016), even though D1 is not the first determiner in the sentence.

Table 1: Hyperparameters of our models.

Parameter	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
No. heads	N/A	N/A	8

Models

We test three architectures: SRNs, LSTMs, and Transformers. The former two are recurrent networks, and thus predict the next word from a hidden state that is updated incrementally for each new word. The memory gates in the LSTM allow it to learn more long-term dependencies compared to the SRN. The Transformer predicts the next word by attending to each word in the context directly, and thus disposes with the recency bias in recurrent networks. We chose these three architectures because numerous studies have used them to study human sentence processing (e.g., Linzen et al., 2016; Frank et al., 2015; Wilcox et al., 2018; Hollenstein et al., 2021; Wilcox et al., 2020). We adapt our SRN implementation from Frank et al. (2016), and our LSTM implementation from Van Schijndel & Linzen (2018).[2] Our Transformer model is based on a standard PyTorch implementation.

For each architecture, we use the hyperparameters reported in the literature. The most important of these are summarized in Table 1. Note that the number of layers differs across the three models, because we decided to closely follow the computational setup reported in earlier studies: the SRN only uses 1 hidden layer, the LSTM uses 2 stacked layers, and the Transformer uses 8 stacked Transformer blocks. Ideally, we would also experiment with the number of layers and other hyperparameters, but due to the lack of space we focus on manipulating the embedding/hidden layer sizes. We train each model with six different embedding sizes: 32, 64, 128, 256, 512 and 1024. In the SRN and LSTM, the hidden layer is set equal to the embedding size. In the Transformer, the feed-forward layer is set to twice the embedding size. All models are trained for 10 epochs since our preliminary experiments showed that the main qualitative patterns for the two target effects are unlikely to change after more training. The learning rate for each architecture is chosen using a grid search from a set of values $\{20, 2, 0.2, 0.02\}$ to minimize perplexity on the validation set. The models are trained using standard stochastic gradient descent, gradients larger than 0.25 are clipped, and the learning rate is divided by a factor of 4 if there is no decrease in perplexity on the validation set for 3 consecutive epochs. Each model is trained with 3 different random initializations, and the results are averaged.

[2] https://github.com/vansky/neural-complexity

Training regime

We train the models on English and/or Dutch text. For English, we use the Wikipedia corpus from Gulordava et al. (2018), and for Dutch, we use the Wikipedia corpus from Winther et al. (2021). We follow the preprocessing steps in Winther et al. and use their vocabulary size of 50k word types.[3] The English and Dutch corpora are matched in size, resulting in a corpus of 2M sentences for each language, which are split into training, validation, and test sets in proportion 80:10:10. The test data is only used for intrinsic model evaluation, which we do not report here for the sake of space. The resulting corpora are used to train the monolingual models.

For the bilingual models, we create a balanced bilingual corpus that consists of Dutch and English sentences in a 50:50 ratio, mixed in a random fashion such that each new sentence has a 50% chance of coming from each language. Furthermore, since Winther et al. (2021) only found a significant cognate effect in models trained on an unequal number of sentences from each language, we also create an unbalanced bilingual corpus that consists of Dutch and English sentences in a 75:25 ratio. The unbalanced corpus further differs from the balanced corpus in that languages are presented consecutively, such that in each epoch, the model is exposed to the Dutch sentences before the English sentences (following Winther et al.'s method of simulating an unbalanced/sequential bilingual speaker with higher exposure to Dutch).

Evaluation

To test whether our models can correctly predict human readers' data on the two target tasks, we follow the evaluation setup from the corresponding studies (Frank et al., 2016; Winther et al., 2021). Specifically, for each target item we compute the surprisal values, which are then processed as described below. Surprisal s of a word w_i at position i is a standard measure shown to be associated with the processing effort for that word in a given context (e.g., Levy, 2008):

$$s(w_i) = -\log P(w_i | w_{1\ldots i-1}) \quad (1)$$

Cognate facilitation effect. Following Winther et al. (2021), we use the 21 English pairs of stimuli selected from Bultena et al. (2014). Each pair consists of two nearly identical sentences that only differ in one word: cognate or non-cognate control, where cognates are spelled identically in English and Dutch. We compare surprisal values of the cognate and control words embedded in the same sentence and compute the *cognate effect size*, CE, as the difference between the two surprisal values:

$$\text{CE} = s(w_i^{\text{control}}) - s(w_i^{\text{cognate}}) \quad (2)$$

Because higher surprisal values correspond to *less* likely sequences, a positive CE value indicates a model's preference for a cognate (rather than non-cognate) word in a given context.

[3] An exception to this is made in the Dutch corpus, where we extend the vocabulary by 27 words that occur in the test sentences from Frank et al. (2016), to ensure the testing is not carried out on words unknown to the model.

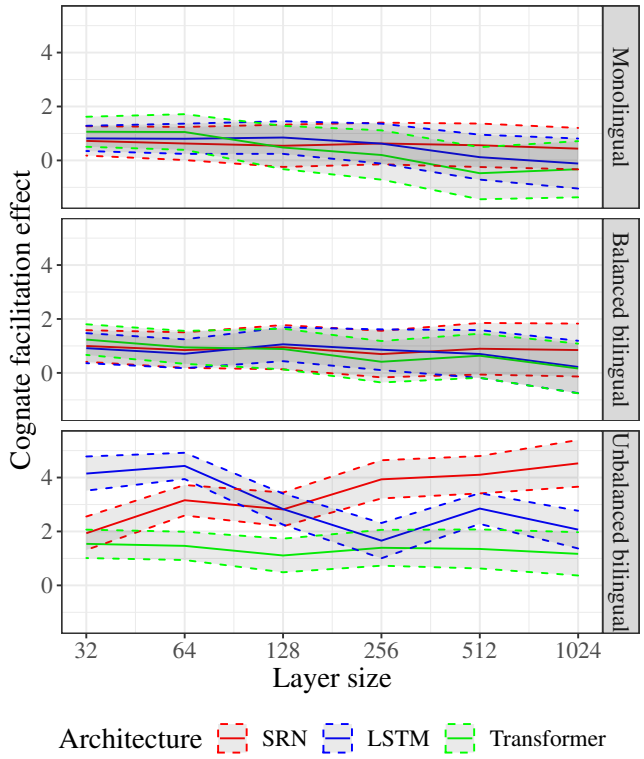

Figure 1: Size of the cognate facilitation effect (CE) as a function of hidden layer size across three model types (English monolingual model and two bilingual models) and three architectures. The results are averaged over test items and random initializations. Error bands show the standard error of the mean across the 21 test items.

Grammaticality illusion. We use the English and Dutch sentences from Frank et al. (2016), but exclude 2 sentences per language that contain out-of-vocabulary words, resulting in 14 test sentences per language. We consider the model's surprisal value of the determiner directly following the verbs (D1), and compute the model's *grammaticality preference*, GP, as the difference in surprisal on D1 embedded in an ungrammatical vs. grammatical version of the sentence:

$$\text{GP} = s(w_i^{\text{ungram.}}) - s(w_i^{\text{gram.}}) \quad (3)$$

A positive GP value in (3) means that the model prefers grammatical sentences, and a negative GP value indicates its preference for ungrammatical sentences.

Results

Here, we present the results on the two evaluation tasks: the cognate facilitation effect and the grammaticality illusion. Due to the lack of space, we do not present models' intrinsic evaluation in terms of their perplexity on validation/test data, but lower perplexity was observed in models with larger hidden layer sizes, and also in LSTMs and Transformers rather than SRNs. More details can be found in Roslund (2021).

Cognate facilitation effect

The results for the cognate facilitation effect across all models are presented in Figure 1.

Monolingual models. As a sanity check, we first look at the monolingual models, which are not expected to show the cognate effect. Indeed, we observe that the effect stays close to zero for all models and all hidden layer sizes (top panel). To test this result statistically, we fit a mixed-effects regression model predicting the CE from the model architecture, log-transformed hidden layer size, and their interaction, with random intercepts and random slopes for individual predictors over items and random initializations. This analysis confirms that there is no statistically significant cognate effect for any model: intercept (i.e., SRN as the reference level) is 0.59, $p = .399$; β (LSTM) $= -0.07$, $p = .805$; β (Transformer) $= -0.25$, $p = .462$. This is in line with the result of Winther et al. (2021) for their monolingual models. Also, on average there are no differences across the three model architectures.

Balanced bilingual models. A visual examination of the results (middle panel) suggests that the patterns are similar to those observed in the monolingual models. Our statistical analysis of the data with an analogous mixed-effects regression confirms this observation: on average, there is no statistically significant cognate effect: intercept (SRN as the reference level) is 0.87, $p = .274$; β (LSTM) $= -0.13$, $p = .621$; β (Transformer) $= -0.16$, $p = .553$. Again, there are no significant differences across the three architectures.

Unbalanced bilingual models. The patterns for the unbalanced models (bottom panel) are different. In particular, there is a positive cognate effect, and its size depends on the architecture and the hidden layer size. A mixed-effects regression suggests that all three architectures consistently show a statistically significant cognate facilitation effect, although the effect size on average is significantly smaller in the Transformer than in the LSTM and SRN: intercept (here, Transformer as the reference level) is 1.34, $p = .046$; β (SRN) $= 2.07$, $p < .001$; β (LSTM) $= 1.66$, $p < .001$.

Hidden layer size. Unlike in balanced bilingual and monolingual models, the cognate effect size in the unbalanced models changes as a function of the hidden layer size, and the direction of this change depends on the architecture. In the Transformer the effect stays stable across the hidden layer sizes (the main effect of layer size in the mixed-effects regression is $\beta = -0.05$, $p = .583$), in the SRN it increases with the hidden layer size (the interaction term 'SRN $\times$ layer size' is $\beta = 0.53$, $p < .001$), and in the LSTM it decreases (the interaction term 'LSTM $\times$ layer size' is $\beta = -0.41$, $p = .005$). This partially supports our intuition about the role of the hidden layer size in bilingual models: recall that we hypothesized that bilingual models can be more susceptible to changes in their hidden layer sizes due to the interplay of the representations from the two languages. While the cognate effect is observed consistently in the unbalanced models, its magnitude changes depending on the hidden layer size in the SRN and LSTM (but not in the Transformer), and the direction of that change

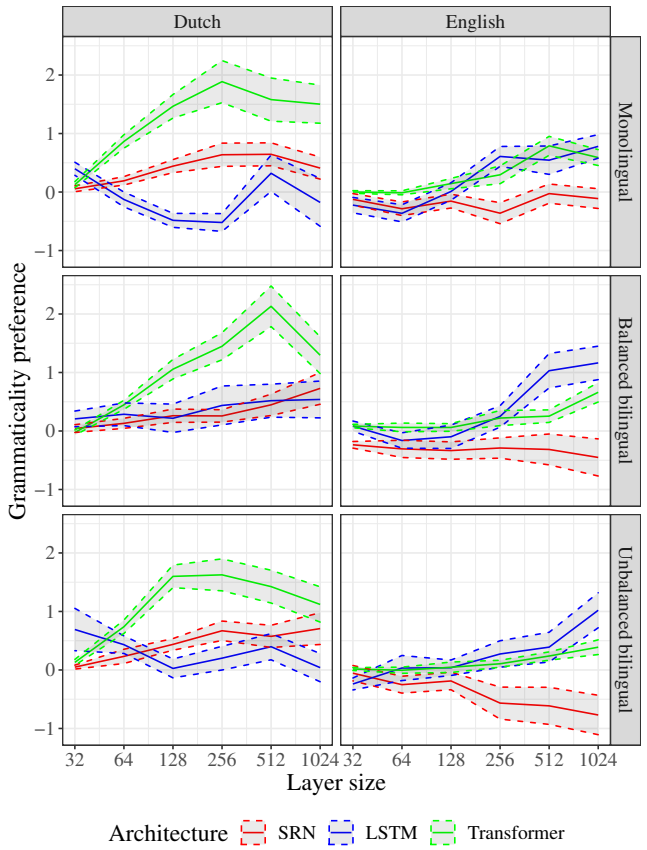

Figure 2: Grammatical preference (GP) as a function of hidden layer size across three model types (one monolingual model and two bilingual models) and three architectures. The results are averaged over test items and random initializations. Error bands show the standard error of the mean across the 14 test items.

differs across the two models.

To summarize our results, all three models show the cognate facilitation effect in the unbalanced, but not in the balanced models, which replicates the existing findings for the LSTM (Winther et al., 2021) and also extends these findings to the two other architectures, SRN and Transformer. Although the effect is consistently present in all the three models, its size varies depending on the architecture and the hidden layer size.

Grammaticality Illusion

The grammaticality preference values of each model are presented in Figure 2.

Monolingual models. We first examine the results for the monolingual models (top panels). Recall that we expect a positive grammaticality preference in Dutch, and a negative preference in English. For the **Dutch** models (top left panel) we see a positive preference for the Transformer and SRN, but not for the LSTM. Again, we fit a linear mixed-effects regression to the Dutch GP values of the monolingual model, with fixed effects of architecture (SRN vs. LSTM vs. Transformer), log-transformed hidden layer size, and their interaction, with random intercepts and slopes for the individual predictors over items and random initializations. This statistical analysis indicates no GP effect for the LSTM, but a positive GP for the SRN and the Transformer, and this effect is larger in the Transformer compared to the SRN: β (SRN) $= 0.40$, $p = .010$; β (LSTM) $= -0.10$, $p = .604$; β (Transformer) $= 1.24$; $p < .001$. While for the LSTM the preference on average is not different from zero, it varies in a non-linear way depending on the hidden layer size, as we can see in the top left panel of Figure 2. For **English** monolingual models, a visual examination suggests that all architectures struggle with reproducing the expected negative GP (top right panel). Indeed, a mixed-effects regression analogous to the one described above suggests that on average, none of the three architectures show a statistically significant negative GP, and the Transformer even predicts a preference in the 'wrong', i.e., positive, direction: β (SRN) is -0.18, $p = .147$; β (LSTM) $= 0.23$, $p = .088$; β (Transformer) $= 0.30$; $p = .004$. Additional analyses of individual hidden layer sizes show that only the SRN and the LSTM with hidden layer size 64 can predict a statistically significant negative GP. However, considering the large number of hidden layer sizes that we tested, it is unclear whether this result is a statistical error.

Balanced bilingual models. Just as for the cognate effect in the previous section, the patterns of the balanced models are overall similar to those of the monolingual models (compare the middle panels in Figure 2 to the top panels). Again, for Dutch, the Transformer and SRN, but not the LSTM, correctly predict a statistically significant positive GP across the hidden layer sizes: β (SRN) is 0.31, $p = .021$; β (LSTM) $= 0.37$, $p = .114$; β (Transformer) $= 1.06$; $p < .001$. For English, surprisingly, a mixed-effects regression suggests that, unlike in monolingual models, the bilingual balanced SRN shows a statistically significant negative GP across the hidden layer sizes, while the Transformer with larger hidden layer sizes and the LSTM show preference in the 'wrong', positive, direction: β (SRN) is -0.32, $p = .026$; β (LSTM) $= 0.38$, $p = .011$; β (Transformer) $= 0.22$, $p = .110$; β (Transformer $\times$ layer size) $= 0.13$, $p = .001$.

Unbalanced bilingual models. The overall qualitative patterns for the unbalanced models are also similar to those in the balanced models (compare the middle vs. the lower panels in Figure 2). Again, the Transformer and the SRN (but not the LSTM) show a statistically significant positive GP in Dutch: β (SRN) is 0.44, $p = .007$; β (LSTM) $= 0.30$, $p = .101$; β (Transformer) $= 1.11$, $p < .001$. Also, only the SRN shows a negative GP in English: β (SRN) is -0.41, $p = .009$; β (LSTM) $= 0.25$, $p = .092$; β (Transformer) $= 0.13$, $p = .369$.

Hidden layer size. We test whether the preferences of the bilingual models depend on the hidden layer size to a larger extent than in the monolingual models. Since the patterns in this case are less obvious than for the cognate effect in the previous section, we fit two mixed-effects regressions (one per language) to *all* models, i.e., monolingual and bilingual.

These models predict GP from main effects of model type and architecture, their two-way interaction, their three-way interaction with the hidden layer size, and random intercepts and slopes over items and random initializations. We find that in English, the unbalanced (but not balanced) bilingual SRN models (but not LSTM and Transformer) are more susceptible to the changes in the hidden layer size than their monolingual counterparts. Specifically, in the monolingual SRN models the GP stays stable across all hidden layer sizes: β (layer size $\times$ SRN) $= 0.02$, $p = .591$. In unbalanced bilingual SRNs it decreases significantly in larger hidden layer sizes (compare the slopes of the red line in the top right vs. bottom right panel): β (layer size $\times$ SRN $\times$ unbalanced) $= -0.16$, $p < .001$. In balanced bilingual SRNs this decrease is relatively small: β (layer size $\times$ SRN $\times$ unbalanced) $= -0.05$, $p = .309$. In Dutch, we find no statistically significant patterns in this regard: bilingual models are susceptible to changes in the hidden layer size approximately to the same degree as monolingual models.

To summarize our results, the SRN is the only architecture that can correctly reproduce the human grammaticality illusion in both languages. Surprisingly, for English data the monolingual models show a less robust positive preference than the bilingual models. The LSTM exhibits a variable behavior across different hidden layer sizes, while the Transformer consistently prefers grammatical sentences in both languages. Finally, the preferences of the unbalanced bilingual (but not monolingual and balanced bilingual) SRNs in English depend on the hidden layer size, although this pattern does not hold for the other two architectures or for the Dutch data.

Discussion

In this study, we evaluated three commonly used neural language model architectures – SRN, LSTM, and Transformer – trained on two languages, Dutch and English, in terms of their ability to predict two processing effects commonly observed in bilingual speakers, namely cognate facilitation and grammaticality illusion. While systematic comparisons of monolingual neural LMs do exist (e.g., Merkx & Frank, 2021; Wilcox et al., 2020), to our knowledge this is the first study of this kind for models trained on two languages.

We found that all three architectures were able to correctly predict the human-like behavior for the processing of cognate vs. non-cognate words in English sentences, known as the cognate facilitation effect. The effect was only observed in the 'unbalanced' models trained on larger amounts of Dutch than English. This result replicates the findings of Winther et al. (2021), who demonstrated the effect in an LSTM model, and extends their findings to two other architectures, SRN and Transformer, providing further support to frequency-based explanations of the cognate facilitation effect (e.g., Strijkers et al., 2010). We also found that the size of the cognate effect was substantially smaller in the Transformer compared to the other architectures.

For the grammaticality illusion, the models exhibit more variable patterns. For Dutch sentences, where human speakers tend to read grammatical sentences more quickly, two out of the three architectures (SRN and Transformer) make the correct predictions. For English sentences, where human speakers tend to read *ungrammatical* sentences more quickly, only the SRN is able to show this effect. Moreover, the effect is highly unstable in the monolingual SRN model, as it is only present for one of the layer sizes. Interestingly, the bilingual SRN models (both 'balanced' and 'unbalanced') consistently predict the effect. This result is in line with the findings of Frank et al. (2016) for their bilingual SRN model. However, the lack of the stable predictions in our monolingual SRN model across the layer sizes and the lack of the effect in the other two architectures requires further investigation. We can speculate that this pattern of results is because faster processing of the ungrammatical sentences (i.e., grammaticality illusion) requires human speakers to track the nested clauses *incorrectly*, while the LSTM and Transformer are good at capturing long-term dependencies (e.g., Gulordava et al., 2018; Mueller et al., 2020), thus failing to reproduce the grammaticality illusion in English. This line of reasoning also provides further support to Frank et al.'s (2016) explanation of the grammaticality illusion: since the SRN relies more on local linguistic information compared to the other two architectures, its success corroborates the language statistics hypothesis, which explains speakers' behavior in a given language by the likelihood of three verbs occurring in a sequence in that language.

Across the two evaluation tasks, the SRN was the only architecture that could predict both effects. This suggests that, even though models with more parameters and more complex architectures, such as LSTM and Transformer, are superior in many syntactic tasks, they may be less successful in replicating some psycholinguistic effects, in particular those related to bilingual speakers' sentence processing. This is in line with Merkx & Frank's (2021) argument that more complex architectures do not necessarily result in better models of human sentence processing.

Finally, our hypothesis that neural language models trained on two languages are more sensitive to changes in hidden layer size than monolingual models has been partially supported. The size of the cognate effect in the bilingual SRN and the LSTM, as well as the grammaticality preference in English for the bilingual SRN, were affected by the layer size to a greater extent than in their monolinguals counterparts. One practical consequence of this finding is that the hyperparameters of bilingual models should be selected empirically rather than being directly adopted from analogous monolingual models.

Acknowledgments: We thank Irene Winther, Sameer Bansal, and five anonymous reviewers for their useful feedback on the earlier version of this paper.

References

Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in natu-

ralistic sentence reading. *Neuropsychologia*, *134*, 107198.

Bultena, S., Dijkstra, T., & Van Hell, J. G. (2014). Cognate effects in sentence context depend on word class, L2 proficiency, and task. *Quarterly Journal of Experimental Psychology*, *67*, 1214–1241.

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST*.

Christiansen, M. H., & MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, *59*, 126–161.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1283–1296.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of ACL-HLT*.

Dijkstra, T., Grainger, J., & Van Heuven, W. J. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language*, *41*, 496–518.

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

Frank, S. L. (2014). Modelling reading times in bilingual sentence comprehension. In *Proceedings of CogSci.*

Frank, S. L. (2021). Toward computational models of multilingual sentence processing. *Language Learning*, *71*, 193–218.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Frank, S. L., Trompenaars, T., & Vasishth, S. (2016). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, *40*, 554–578.

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

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of NAACL-HLT*.

Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. (2020). SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of ACL*.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of CMCL*.

Gulordava, K., Bojanowski, P., Grave, É., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL-HLT*.

Guo, M., Dai, Z., Vrandečić, D., & Al-Rfou, R. (2020). Wiki-40b: Multilingual language model dataset. In *Proceedings of LREC*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hollenstein, N., Pirovano, F., Zhang, C., Jäger, L., & Beinborn, L. (2021). Multilingual language models predict human reading behavior. In *Proceedings of NAACL-HLT*.

Khoe, Y. H., Tsoukala, C., Kootstra, G. J., & Frank, S. L. (2021). Is structural priming between different languages a learning effect? Modelling priming as error-driven implicit learning. *Language, Cognition and Neuroscience*, 1–21.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., ... others (2021). Dynabench: Rethinking benchmarking in NLP. In *Proceedings of NAACL-HLT*.

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

Li, P. (2002). Bilingualism is in dire need of formal models. *Bilingualism: Language and Cognition*, *5*, 213–213.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Linzen, T., & Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of CogSci.*

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of EMNLP*.

Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? In *Proceedings of CMCL*.

Mueller, A., Nicolai, G., Petrou-Zeniou, P., Talmina, N., & Linzen, T. (2020). Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of ACL*.

Roslund, R. (2021). *A model comparison between neural architectures of human bilingual sentence processing.* [Master's thesis, University of Edinburgh]. Edinburgh Research Archive. Retrieved from `http://dx.doi.org/10.7488/era/2212`

Strijkers, K., Costa, A., & Thierry, G. (2010). Tracking lexical access in speech production: Electrophysiological correlates of word frequency and cognate effects. *Cerebral Cortex*, *20*, 912–928.

Tsoukala, C., Broersma, M., van den Bosch, A., & Frank, S. L. (2021). Simulating code-switching using a neural network model of bilingual sentence production. *Computational Brain & Behavior*, *4*, 87–100.

Tsoukala, C., Frank, S. L., & Broersma, M. (2017). "He's pregnant": Simulating the confusing case of gender pronoun errors in L2 English. In *Proceedings of CogSci.*

Van Schijndel, M., & Linzen, T. (2018). Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of CogSci.*

Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, *25*, 533–567.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Proceedings of NeurIPS.*

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392.

Wilcox, E., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of CogSci.*

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? In *Proceedings of BlackboxNLP.*

Winther, I. E., Matusevych, Y., & Pickering, M. J. (2021). Cumulative frequency can explain cognate facilitation in language models. In *Proceedings of CogSci.*