Cumulative Frequency Can Explain Cognate Facilitation in Language Models

Irene E. Winther (irene.winther@ed.ac.uk)
Centre for Doctoral Training in Natural Language Processing
School of Informatics, 10 Crichton Street
Edinburgh, EH8 9AB, UK

Yevgen Matusevych (yevgen.matusevych@ed.ac.uk)
School of Informatics, 10 Crichton Street
Edinburgh, EH8 9AB, UK

Martin J. Pickering (martin.pickering@ed.ac.uk)
School of Philosophy, Psychology & Language Sciences, 7 George Square
Edinburgh, EH8 9JZ, UK

Abstract

Cognates – words which share form and meaning across two languages – have been extensively studied to understand the bilingual mental lexicon. One consistent finding is that bilingual speakers process cognates faster than non-cognates, an effect known as cognate facilitation. Yet, there is no agreement on the underlying factors driving this effect. In this paper, we use computational modeling to test whether the effect can be explained by the cumulative frequency hypothesis. We train a computational language model on two language pairs (Dutch–English, Norwegian–English) under different conditions of input presentation and test it on sentence stimuli from two existing studies with bilingual speakers of those languages. We find that our model can exhibit a cognate effect, lending support to the cumulative frequency hypothesis. Further analyses reveal that the size of the effect in the model depends on its linguistic accuracy. We interpret our results within the literature on cognate processing.

Keywords: cognate facilitation effect; cumulative frequency; bilingualism; language model; sentence processing

Introduction

Cognates are words that share form and meaning across languages, like the noun *winter*, which has identical orthography and meaning in English and Dutch.[1] Cognates are widely studied in bilingual language comprehension and production as they may be central to our understanding of bilingual lexical access (Costa, Caramazza, & Sebastian-Galles, 2000). One consistent finding about cognates is that bilingual speakers process them faster than non-cognates (and than monolingual speakers do), known as the *cognate facilitation effect*. This effect has been reported across various experimental tasks, including sentence processing (Costa et al., 2000; Dijkstra, Grainger, & van Heuven, 1999; Libben & Titone, 2009; Schwartz & Kroll, 2006, etc.). For instance, Dutch–English bilinguals read the English example (1) faster with the cognate word *winter* than with the control word *prison* (‘gevangenis’ in Dutch), even though the two words are matched on their English corpus frequencies (Bultena, Dijkstra, & van Hell, 2014).

(1) The residents dislike the [*winter* / *prison*] ...
cognate control

The exact nature of the facilitation effect, however, is not known. One open question is whether it is driven by the higher cumulative frequency of cognates compared to non-cognates (Midgley, Holcomb, & Grainger, 2011; Voga & Grainger, 2007): due to the identical form, Dutch–English bilinguals encounter a cognate such as *winter* more frequently than a non-cognate such as *prison*. This effectively makes cognates similar to words with very high frequency within one language. Although the cumulative frequency hypothesis has found some empirical support (Voga & Grainger, 2007; Strijkers, Costa, & Thierry, 2010), it is not universally accepted. For example, van Hell and Dijkstra (2002) argue that cognate facilitation is not just due to cumulative frequency and that cognates have a special type of representation in the mental lexicon.

The current paper investigates whether the cognate facilitation effect in sentence processing can be explained by cumulative word frequency. We use computational language modeling, a well-established paradigm in research on online language processing (e.g., Smith & Levy, 2013). Language models are well-suited for studying online processing, because they predict the next word in a sentence, and their predictions correlate with human reading times (e.g., Goodkind & Bicknell, 2018). The exact relationship between a word’s predictability in context and its (unigram) frequency is not known (e.g., Staub, 2015; Shain, 2019), but more frequent words tend to also be more predictable in neural language models (Xie et al., 2015). Importantly, such models provide full control over the input data, so that the exact frequencies of cognates and non-cognates can be counted, and they treat cognates and non-cognates in exactly the same way, without assigning cognates any special status. This makes computational language modeling a suitable paradigm to test the cumulative frequency hypothesis.

We train the model on two (Dutch–English, Norwegian–English) languages and test it on stimuli from two experiments on the reading of identical cognates in English sentence contexts, carried out with corresponding bilingual populations (Bultena et al., 2014; Winther, 2017). The sentence contexts contain a target word which is either a cognate or a non-cognate (control) word, as in (1) above. Because the

[1] Cognates can be spelled with small differences across languages, but here we focus on identical cognates.

Dutch–English cognate *winter* occurs in both languages, a model trained on Dutch and English will encounter this word overall more frequently than the control word, *prison*. If our model predicts cognates to be more likely than non-cognates in the same sentence context, this would support the cumulative frequency hypothesis. We consider multiple model variants, trained under different conditions of input data presentation, through manipulating three variables: first language (L1) pre-training, inter-sentential language mixing in the input, and first-to-second language (L1:L2) input ratio. These manipulations change the cumulative frequencies of cognate and control words and allow us to investigate under which conditions, if any, the cognate facilitation effect is observed in the model.

Our main contribution in this study is a test of the cumulative frequency hypothesis using a computational language model trained on two languages, which we evaluate against human experimental data with Dutch–English and Norwegian–English bilinguals (Bultena et al., 2014; Winther, 2017). We show that the model can predict the cognate facilitation effect observed in human data. Our cumulative frequency analysis confirms that frequency explains the model's predictions better than the words' cognate status alone. We also find that the size of the cognate effect in the model depends on its linguistic accuracy.

Background

Cognate Facilitation Effect

A cognate facilitation effect in sentence processing has been observed across many language pairs (e.g., Bultena et al., 2014; Schwartz & Kroll, 2006; Libben & Titone, 2009). Here, we focus on Dutch–English and Norwegian–English, due to the availability of human experimental data with identical cognates. Specifically, we evaluate our model on the stimuli from two studies on cognate processing in second language (L2) sentence contexts by Dutch–English (Bultena et al., 2014, henceforth B14) and Norwegian–English bilinguals (Winther, 2017, henceforth W17). In both studies, bilingual participants read L2 English sentences containing either cognates or non-cognates (matched on their English frequency) whilst their eye movements were measured. Shorter reading times were observed for cognates than for control words.

Language Modeling and Bilingualism

In research on monolingual online sentence processing, language models are commonly used to study a variety of phenomena (e.g., Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018; Arehalli & Linzen, 2020). In the computational study of bilingualism, to our knowledge, only Frank (2014) and Frank, Trompenaars, and Vasishth (2016) trained such models on natural language data from two languages at the same time (see also Frank, 2021). These two studies found a significant correlation between the bilingual model's predictions and reading times of L1 Dutch and L2 English sentences by Dutch–English bilinguals.

Table 1: Number of sentences and tokens in the training corpora. The corpora are matched on number of sentences.

	Sentences (mln.)	Tokens (mln.)		
	All languages	Norw.	Dutch	English
Train	1.6	28.2	29.0	42.5
Validation	0.2	3.5	3.6	5.3
Test	0.2	3.5	3.6	5.3
Total	2.0	35.2	36.3	53.1

While it is common to compare various language model architectures on how well they predict human data (e.g., Futrell et al., 2019), we are interested to test whether the cognate facilitation effect can be predicted by *any* language model trained on two languages. We choose to use a long short-term memory (LSTM; Hochreiter & Schmidhuber, 1997) model, because it has been commonly used in online processing studies mentioned above and is arguably a cognitively plausible model (Merkx & Frank, 2020).

Methods

Our general approach is to train the model on data from one or two languages and evaluate its linguistic accuracy and its ability to predict the data from the two experimental studies with cognates, B14 and W17.

Training Data

To simulate L1 input, we use two corpora created from Dutch and Norwegian Wikipedia. For L2 input, we use an existing English Wikipedia corpus (Gulordava et al., 2018). We follow standard practices for data preprocessing: limit the vocabulary to the 50k most frequent words for each corpus, replace the remaining tokens with the unknown symbol, and exclude sentences with more than 5% unknown tokens. Each preprocessed corpus is divided into the standard $80/10/10$ split for training, validation and test sets. The corpus statistics are set out in Table 1.

Model

We use an LSTM language model, namely the baseline implementation of van Schijndel and Linzen (2018)[2], adopting their architecture and hyperparameters: two hidden layers (650 units each), with 650-dimensional word embeddings, a learning rate of 20, a dropout rate of 0.2 and a batch size of 128. The models are trained for 30 epochs (without early stopping) and are tested every 10 epochs, unless specified otherwise.

The model's task is to predict the next word w_i given its preceding context, which yields a probability distribution $P(w_i|w_1...w_{i-1})$. The model's performance is measured by

[2] `https://github.com/vansky/neural-complexity`. Note that van Schijndel and Linzen (2018) trained the model for 40 epochs, but we stop after 30 epochs due to very small improvements after that.

Table 2: Log-frequencies of each word type per L1:L2 training condition. Mean (*M*) and standard deviation (*SD*) are shown.

	English		50:50		75:25	
	M	*SD*	*M*	*SD*	*M*	*SD*
B14 (Dutch–English data)						
Cognate	3.32	0.56	3.18	0.54	3.08	0.55
Control	3.05	0.69	2.75	0.71	2.43	0.77
W17 (Norwegian–English data)						
Cognate	2.85	0.65	2.64	0.64	2.46	0.64
Control	2.79	0.51	2.50	0.52	2.20	0.53

perplexity, the inverse probability of the (unseen) test data:

$$PPL = \left(\prod_{i=1}^{N} P(w_i|w_1...w_{i-1}) \right)^{-\frac{1}{N}} \quad (1)$$

where N is the total number of tokens in the test set. Lower perplexity indicates a higher probability assigned to the test data, i.e., a better fit of the language model to the test data.

Model Evaluation

To estimate whether our model successfully learns two languages, we first report the model's perplexity on the Wikipedia test data. This indicates the model's overall linguistic accuracy, i.e., its ability to predict the following word in natural language sentences.

Our main goal is to determine under which conditions, if any, the bilingual model displays a cognate effect. For this, we use surprisal (S), a commonly used measure thought to reflect the cognitive effort of processing a word in a given context (Hale, 2001; Smith & Levy, 2013):

$$S(w_i) = -\log_2 P(w_i|w_1...w_{i-1}) \quad (2)$$

We expect the model to display lower surprisal values for cognates than their controls on two sets of test stimuli: 21 pairs of B14's noun sentence stimuli and 28 pairs of W17's semantically constraining sentences. We also measure the size of the cognate effect for a given pair of words as the difference in surprisal values: a greater difference between the control and the cognate words occurring in the same context indicates a larger effect size:

$$\Delta S(w_i) = S_{\text{control}_i} - S_{\text{cognate}_i} \quad (3)$$

In all the statistical analyses, we use mixed-effects models, as implemented in *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) for *R* (R Core Team, 2020).

Simulations

Using the described architecture and data sets, we train monolingual and bilingual models under different conditions of input presentation, as described below. Each model is trained with 5 different random seeds (i.e., initializations of model weights) to ensure the results are robust. We evaluate the models on linguistic accuracy and their ability to predict the cognate facilitation effect observed in B14 and W17.

Table 3: Characteristics of the six bilingual models.

Model	Variables			Type of bilinguals	
	L1 pretr.	Lang. mix.	L1:L2 ratio	Unbal./ balanc.	Simult./ sequen.
PT-MX-50	+	+	50:50	Un-/bal.	Sequen.
PT-MX-25	+	+	75:25	Unbal.	Sequen.
NPT-NMX-50	−	−	50:50	Balanc.	Sequen.
NPT-NMX-25	−	−	75:25	Unbal.	Sequen.
NPT-MX-50	−	+	50:50	Balanc.	Simult.
NPT-MX-25	−	+	75:25	Unbal.	Simult.

Monolingual Models

We first train monolingual models on each of the three languages independently: English, Dutch, and Norwegian. We report each model's linguistic accuracy (perplexity) on the test set for the respective language. These are used as monolingual points of comparison to evaluate the performance of the bilingual models.

Bilingual Models

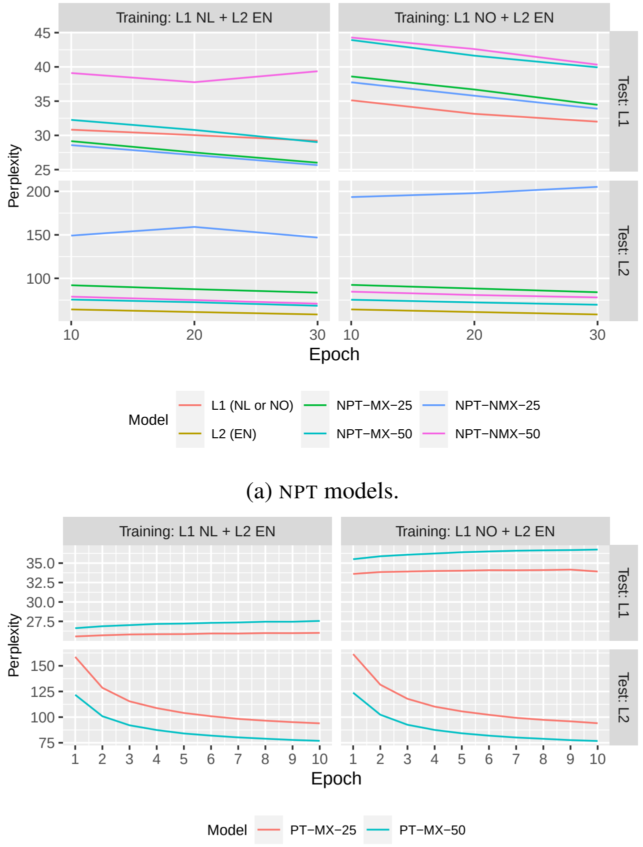
The bilingual models are trained on samples from two languages' data, keeping the total number of sentences constant. The vocabulary for the bilingual models is the unified vocabulary for the L1 and L2 corpora. There are various ways to train the model on two languages, depending on the relative amount of training data in each language and the order of presentation. We explore these differences by manipulating three variables:

1. L1 pretraining. We consider models with pretraining (pretrained, PT) and without pretraining (non-pretrained, NPT). NPT models are trained on data from two languages in parallel. PT models are first trained on the L1 data for 30 epochs and then on combined bilingual data for 10 more epochs. Unlike NPT models, PT models are saved for further testing after every epoch during bilingual training.

2. Language mixing. Within each training epoch, we either present the model with L1 data followed by L2 data (non-mixed models, NMX), or with randomly shuffled combined L1–L2 sentences (mixed models, MX).

3. L1:L2 ratio. We combine either 75% of the L1 and 25% of the L2 corpora, or 50% each. These proportions are arbitrary, as our aim is to compare models trained on different and equal proportions of L1:L2. Manipulating this ratio changes the cumulative frequencies of English cognates and control words in the training data, as shown in Table 2.

The 3 variables with 2 values each result in 8 bilingual models. We do not consider the 2 pretrained non-mixed models, because our preliminary simulations showed that a model cannot learn L2 from non-mixed input if its exposure to L1 is much higher than to L2. This leaves us with 6 bilingual mod-



(a) NPT models.

(b) PT models.

Figure 1: Mean model perplexities for (a) non-pretrained (NPT) and (b) pretrained (PT) models, on the L1 Dutch (NL) or Norwegian (NO), or L2 English (EN) Wikipedia test sets.

els. We use the values of the 3 variables above to refer to each model: e.g., PT-MX-25 is a pretrained mixed model with 25% L2 in the bilingual part of its training data. Table 3 shows the characteristics of all 6 models and how they can be related to various types of bilingualism: *unbalanced* bilingualism (larger exposure to L1 than to L2) vs. *balanced* bilingualism (approximately equal exposure), and *simultaneous* bilingualism (both languages are acquired from birth) vs. *sequential* bilingualism (learning an L2 later in life). Note that the mapping between model variants and types of bilinguals is approximate: e.g., PT-MX-50 sees equal amounts of L1 and L2 in the bilingual data, yet overall it is exposed to more L1 due to pretraining.

Results

We are primarily interested in whether the bilingual models exhibit the cognate effect and under which conditions. First, however, we present the models' overall linguistic accuracy. All results are averaged over 5 random initializations.

Models' Linguistic Accuracy

We first look at the overall linguistic accuracy of each bilingual model in comparison to the monolingual models, to see how well the bilingual models learn two languages. The models' perplexity over time is shown in Figure 1. Note that the values cannot be compared across panels, as the test sets are different (except the L2 English test set), and the absolute perplexity values vary substantially across languages (Gerz, Vulić, Ponti, Reichart, & Korhonen, 2018). We observe similar patterns in each of the two language pairs. First, L1 test perplexity is low (< 45) and stable across training epochs for all models (top panels in each subplot), indicating that L1 learning is successful and that L2 learning does not jeopardize L1 linguistic accuracy. Second, the L2 English perplexity for most of the bilingual models is only somewhat higher than for the monolingual English model (bottom panels in Figure 1a). One exception is NPT-NMX-25, which has high (> 145) perplexity on the L2 test set compared to all other models (< 100) and does not improve over time. This suggests that when a model is exposed predominantly to one language (L1) *and* sees that language first within each epoch, it cannot achieve high linguistic accuracy in the other language (L2), an issue we return to in the discussion. Lastly, all -25 models have lower L1 perplexity, but higher L2 perplexity than the corresponding -50 models: accuracy in a given language increases with its share in the training data. To summarize, our model displays general learning patterns that are consistent with the conditions of L1 and L2 input data presentation, defending our use of this model for the study of bilingualism.

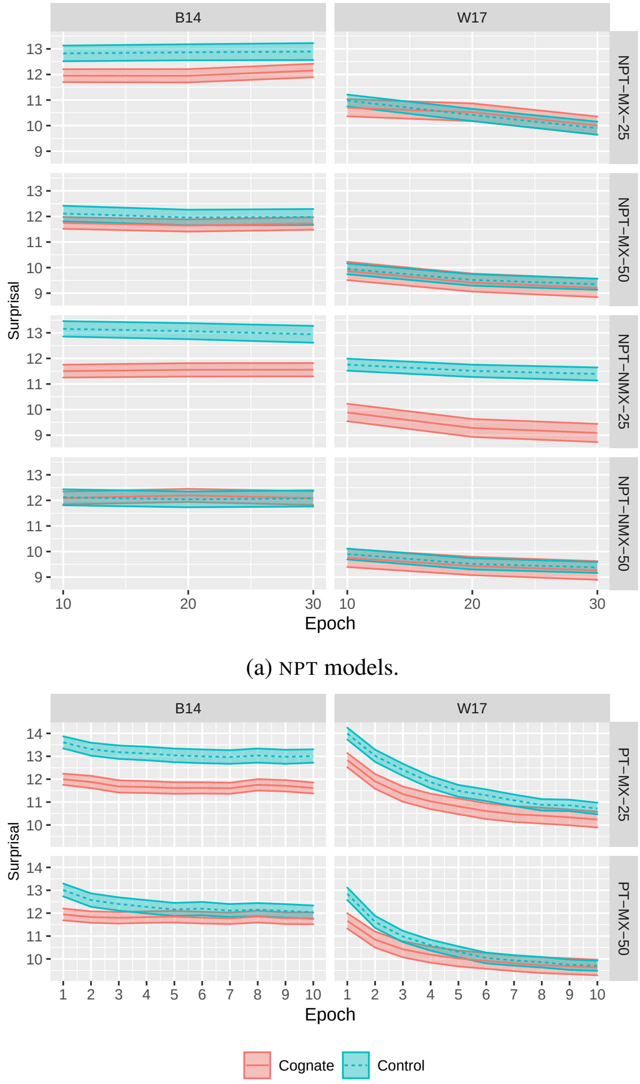
Cognate Facilitation Effect in the Model

Here, we analyse the relationship between words' cognate status and models' surprisal values, to see if the models display the cognate facilitation effect. We expect surprisal for cognates to be lower than for control words. We fit a mixed-effects linear regression to the surprisal values predicted by each bilingual model variant (and the monolingual model, as a sanity check). Fixed effects include cognate status (cognate vs. control), epoch (to account for the model's learning over time), and their interaction. Random effects include intercepts for model's random initialization, item, and sentence context.[3]

As a simple check, we first consider the monolingual English model. Because the frequencies of cognates and control words are matched in English, we expect no cognate effect. Indeed, the model shows no significant effect: the regression coefficients ($b = 0.01$ and 0.08, $|\, t \,| = 0.02$ and 0.11 for B14 and W17, respectively) suggest no meaningful differences between the surprisal values for cognates vs. control words.

The surprisal values for bilingual models are shown in Figure 2. We first look at the four NPT models (Figure 2a). For both language pairs, a significant effect of cognate status is only observed for NPT-NMX-25 (note the gaps between the pairs of lines): $b = -1.79$ and -1.70 ($|\, t \,| = 2.30$ and 2.34) on

[3] Full mixed-effects model structure in *R*: surprisal ∼ cognate_status*epoch + (1|word) + (1|sentence) + (1|initialization). Random slopes could not be included due to convergence issues. Some random intercepts were excluded to ensure model convergence: random initialization for NPT-MX-50, random initialization and sentence for the English model.



(a) NPT models.

(b) PT models.

Figure 2: Mean model suprisal values for cognates and control words from B14 and W17. Ribbons indicate standard error. Note that the absolute surprisal values on B14 vs. W17 cannot be compared, because the test data sets are different.

B14 and W17, respectively, meaning lower surprisal values for cognates than for control words. Additional by-epoch analyses for this model variant confirmed the presence of the cognate facilitation effect at each time point for both data sets, except at epoch 30 on the B14 stimuli.

Next, we look at the two PT models (Figure 2b). We observe a significant effect of cognate status for PT-MX-25 on both data sets: $b = -1.56$ and -1.23 ($|\, t \,| = 2.37$ and 2.03), again suggesting lower surprisal values for cognates than for control words. We also observe significant positive interactions between cognate status and epoch for both PT-MX-25 and PT-MX-50 on both data sets, suggesting that the difference between control and cognate words gets smaller with more training (the gap between the pairs of lines in Figure 2b decreases over time). Again, our by-epoch analyses show that on the B14 stimuli the cognate effect was significant for most epochs of the PT-MX-25 model (but not at all for the for PT-MX-50 model). On the W17 stimuli, the effect was significant only at epoch 1 for both PT-MX-25 and PT-MX-50.

Cumulative Frequency Analysis

We have observed a consistent cognate facilitation effect in two of our bilingual models: NPT-NMX-25 and PT-MX-25. Because our models treat cognates and control words in the same way, this effect must be attributed to the different cumulative frequencies of the two types of words in the training data. If that is the case, a word's frequency must predict its surprisal value better than the cognate status does. To test this, we again consider the surprisal values of NPT-NMX-25 and PT-MX-25 on B14 and W17. Just as before, we fit four mixed-effects regressions to these values, but this time including cumulative word frequency as an additional predictor variable.[4] The regression analyses show that word frequency is a significant predictor of surprisal for both language models on both B14 and W17, while cognate status is no longer significant in these new regressions. For NPT-NMX-25 tested on B14 and W17, the regression coefficients for frequency are $b = -2.43$ and -3.42 ($|\, t \,| = 4.68$ and 5.65), and for cognate status $b = -0.22$ and -0.81 ($|\, t \,| = 0.30$ and 1.15). Analogously, for PT-MX-25, $b = -2.32$ and -3.67 ($|\, t \,| = 4.96$ and 6.77) for frequency, and $b = 0.97$ and 0.96 ($|\, t \,| = 1.28$ and 1.53) for cognate status. This confirms that cumulative frequency explains the cognate facilitation effect in our models.

Linguistic Accuracy and the Cognate Effect

We have found that cumulative word frequency can predict surprisal in our two bilingual language models that show the cognate facilitation effect. Yet, frequency alone cannot determine the presence of the effect in a given model. Consider two models trained on the same data under different conditions, NPT-**NMX**-25 and NPT-**MX**-25. Despite identical word frequencies, only the former displays a cognate effect. We know that these two models differ in their linguistic accuracy, and here we ask whether a model's linguistic accuracy (measured by perplexity) can predict the size of the cognate effect in that model. We consider by-epoch perplexity values for each of our six bilingual models and test whether perplexity can predict the size of the cognate effect (measured as ΔS, see Equation 3) at a given training time for a given model. To do so, we fit a series of mixed-effect regressions with ΔS as the dependent variable, L2 (English) perplexity and epoch (with their interaction term) as the fixed effects, and random intercepts over sentence pair and random initialization. As before, we run separate analyses for NPT and PT models. For both types of

[4] For NPT models, we use a word's total frequency in each bilingual training sample. For PT models, we compute the total number of times each word is seen during training: $F(w)_{\text{L1}} \times 30 + F(w)_{\text{biling.}} \times 10$ (where 30 and 10 are the number of L1 pretraining and bilingual training epochs, respectively).

models, L2 perplexity is a significant predictor of ΔS on both data sets. For NPT models, $b = 0.02$ and 0.02 ($| t | = 5.26$ and 6.22) on B14 and W17, respectively. For PT models, the respective values are: $b = 0.01$ and 0.01 ($| t | = 4.28$ and 3.43). This suggests that the cognate effect is observed in the models with higher L2 perplexity, or lower L2 linguistic accuracy.

Discussion

Our main goal was to test whether the cognate facilitation effect in bilingual speakers can be explained by cumulative word frequency. To do this, we trained computational language models on two languages (Dutch–English or Norwegian–English). We manipulated three variables determining the patterns of input data presentation to the model – L1 pretraining, language mixing, and L1:L2 ratio – to explore a number of model variants. We tested each variant on the stimuli from sentence reading experiments with bilingual speakers (Bultena et al., 2014; Winther, 2017), and analyzed the models' predictions for cognate vs. control words. We observed a significant cognate effect for two (out of six) model variants across both language pairs. Because our model does not process cognates differently from other words, the effect observed in our models cannot be explained by cognates' special status, as suggested by van Hell and Dijkstra (2002). Instead, we found that a word's frequency predicts its surprisal value better than its cognate status does. This finding supports the cumulative frequency hypothesis: the higher frequency of cognates (compared to non-cognates) facilitates their processing in sentences. At the same time, we did not test a similar computational model that would explicitly assign cognates a special status, and we cannot argue against van Hell and Dijkstra's proposal based on our result. While further research is needed in this respect, we have shown that the cumulative frequency hypothesis *can* in principle explain the cognate effect.

The common properties of the two model variants that displayed the cognate effect are their higher exposure to L1 than to L2 and the presentation of L1 before L2. Both of these characteristics are associated with lower linguistic accuracy in L2, and our analysis of accuracy confirms that the models' L2 accuracy can predict the size of the cognate effect. Therefore, language accuracy (or "proficiency") also affects the presence of the cognate effect in a given bilingual model. This mirrors the trend found in bilingual speakers: the cognate effect is stronger in less proficient L2 speakers (Bultena et al., 2014; Pivneva, Mercier, & Titone, 2014; Libben & Titone, 2009). Therefore, the effect could be due to the lack of exposure to non-cognate words. In less proficient speakers, cognate words behave like high-frequency words do in a monolingual setting, while their control words (matched on L2 frequency) function like lower-frequency words. With increased proficiency, the differences in exposure to the two types of words become smaller, leading to smaller cognate effects.

Based on the finding that the cognate effect in our model is observed only when the model's exposure to L2 (relative to L1) is low, one could argue that in our case the effect is just an artefact. However, it is important to note the differences in the amount of data required for successful learning in humans vs. neural language models. As pointed out by, e.g., Linzen (2020), compared to models such as the one used in this study, humans learn language from considerably less exposure. Therefore, further research is required to reveal the exact relationship between language exposure, the learning of two languages, and the presence of a cognate effect in both humans and models. In particular, the dynamics of bilingual learning and forgetting in neural language models are not the same as in human speakers, and one challenge is to better understand under which conditions of L1 and L2 exposure the model can successfully learn both languages, and how these conditions differ from what we know about bilingual learning in human speakers. Here we have made first steps in this direction. For example, we showed that with low L2 exposure the order of L1–L2 presentation within each training epoch has a strong effect on the model's L2 proficiency, to the extent that models pretrained on L1 may be unable to learn L2 if it comes after L1 in each epoch. This result, which may seem anomalous from the cognitive perspective, helps us better understand how the model's variables such as pretraining, language mixing and L1:L2 ratio map onto L2 learning settings in human speakers.

Our approach and findings align with Frank's (2021) desiderata (such as parsimony) for useful models of bilingual sentence processing. We argue that computational language models can be used to address open questions in this domain. Despite the mentioned challenges, such models can be theoretically appropriate for drawing conclusions about whether specific effects in bilingual cognition can be explained by learning mechanisms implemented in such models, and under which conditions.

Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

References

Arehalli, S., & Linzen, T. (2020). Neural language models capture some, but not all, agreement attraction effects. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.*

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using {lme4}. *Journal of Statistical Software*, *67*, 1–48.

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.

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.

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

Frank, S. L. (2014). Modelling reading times in bilingual sentence comprehension. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society.*

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

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.

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 the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Gerz, D., Vulić, I., Ponti, E. M., Reichart, R., & Korhonen, A. (2018). On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.*

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018).*

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).*

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics.*

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*, 1735–1780.

Libben, M. R., & Titone, D. A. (2009). Bilingual lexical access in context: Evidence from eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 381–390.

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*

Merkx, D., & Frank, S. L. (2020). Comparing Transformers and RNNs on predicting human sentence processing data. *arXiv:2005.09471.*

Midgley, K. J., Holcomb, P. J., & Grainger, J. (2011). Effects of cognate status on word comprehension in second language learners: An ERP investigation. *Journal of Cognitive Neuroscience*, *23*, 1634–1647.

Pivneva, I., Mercier, J., & Titone, D. (2014). Executive control modulates cross-language lexical activation during L2 reading: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 787–796.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schwartz, A. I., & Kroll, J. F. (2006). Bilingual lexical activation in sentence context. *Journal of Memory and Language*, *55*, 197–212.

Shain, C. (2019). A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).*

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, *9*, 311–327.

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.

van Hell, J. G., & Dijkstra, T. (2002). Foreign language knowledge can influence native language performance in exclusively native contexts. *Psychonomic Bulletin & Review*, *9*, 780–789.

van Schijndel, M., & Linzen, T. (2018). A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.*

Voga, M., & Grainger, J. (2007). Cognate status and cross-script translation priming. *Memory & Cognition*, *35*, 938–952.

Winther, I. E. (2017). *The effect of semantic constraint on lexical access in bilingual word recognition.* [Master's thesis, University of Edinburgh]. Edinburgh Research Archive.

Xie, Y., Daga, P., Cheng, Y., Zhang, K., Agrawal, A., & Choudhary, A. (2015). Reducing infrequent-token perplexity via variational corpora. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers).*