Is Cross-Lingual Transfer in Bilingual Models Human-Like? A Study with Overlapping Word Forms in Dutch and English

Iza Škrjanec[1,2], **Irene Elisabeth Winther**[3], **Vera Demberg**[1], **Stefan L. Frank**[3]
[1]Saarland University, Germany [2]Zuse School ELIZA, Germany [3]Radboud University, the Netherlands
{skrjanec, vera}@lst.uni-saarland.de, {irene.winther, stefan.frank}@ru.nl

Abstract

Bilingual speakers show cross-lingual activation during reading, especially for words with shared surface form. Cognates (friends) typically lead to facilitation, whereas interlingual homographs (false friends) cause interference or no effect. We examine whether cross-lingual activation in bilingual language models mirrors these patterns. We train Dutch-English causal Transformers under four vocabulary-sharing conditions that manipulate whether (false) friends receive shared or language-specific embeddings. Using psycholinguistic stimuli from bilingual reading studies, we evaluate the models through surprisal and embedding similarity analyses. The models largely maintain language separation, and cross-lingual effects arise primarily when embeddings are shared. In these cases, both friends and false friends show facilitation relative to controls. Regression analyses reveal that these effects are mainly driven by frequency rather than consistency in form-meaning mapping. Only when just friends share embeddings are the qualitative patterns of bilinguals reproduced. Overall, bilingual language models capture some cross-linguistic activation effects. However, their alignment with human processing seems to critically depend on how lexical overlap is encoded, possibly limiting their explanatory adequacy as models of bilingual reading.

Keywords: bilingualism, cognate, interlingual homograph, surprisal, semantic similarity

1. Introduction

Compared to speakers of a single language, bilingual speakers process some linguistic phenomena differently during reading, a notable example being words with the same surface form in multiple languages. For cognate words, the surface form as well as the meaning is equivalent or very similar across languages (e.g. *winter* in Dutch and English). For interlingual homographs (henceforth, false friends) the surface form is the same, but the meaning is not (e.g. *brand* means 'fire' in Dutch).

Empirical studies of human reading find that bilinguals process cognate words faster than non-cognate control words (Libben and Titone, 2009; Bultena et al., 2014; Lauro and Schwartz, 2017). This cognate facilitation effect has not been observed in monolinguals, suggesting that it arises specifically from the coexistence of two languages within a single speaker.

For reading of false friends, studies report either slower reading of false friends than control words or no difference in bilingual speakers (Libben and Titone, 2009; Titone et al., 2011; Pivneva et al., 2014; Hoversten and Traxler, 2016). Slower reading of false friends in bilinguals is thought to reflect interference caused by diverging form-meaning mappings across languages.

Both facilitation and interference effects have been attributed to cross-language activation (e.g. Dijkstra and van Heuven, 2002; Kroll et al., 2012). Explaining these cross-lingual interactions requires explicit and testable computational models of how the two languages interact within a single cognitive system (Frank, 2021). Neural language models

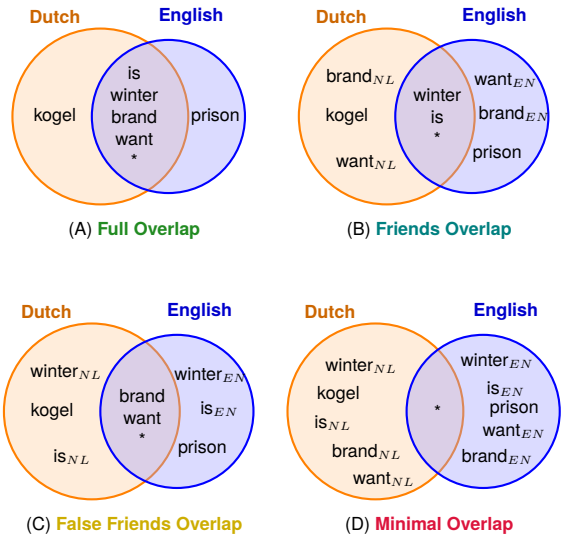

Figure 1: Vocabulary conditions. **Full overlap (A)**: All identical surface forms that appear in Dutch and English are shared between the two languages, each represented by a single embedding. **Friends Overlap (B)**: Only cognates and loan words are shared. **False Friends Overlap (C)**: Only false friends are shared. **Minimal Overlap (D)**: Only punctuation and named entities are shared, while other tokens have language-specific embeddings. Across all conditions, punctuation and named entities are shared (denoted as $*$).

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 110–123
16 May 2026.

(LMs) are strong candidates as computational models of second language learning and bilingual reading (Yadavalli et al., 2023; Oba et al., 2023; Aoyama and Schneider, 2024; Constantinescu et al., 2025).

In Natural Language Processing, cross-lingual transfer is often desired when structural or lexical similarities or simply lexical overlap between languages can be leveraged to improve downstream performance (e.g., Kallini et al., 2025; K et al., 2020; Wu and Dredze, 2019; Pires et al., 2019). In contrast, this work evaluates bilingual LMs for cross-lingual transfer between Dutch and English and explores under what conditions this transfer seems human-like (if at all). We test LMs for lexical processing via vocabulary manipulation in four conditions (Figure 1), controlling how overlapping word forms are treated in the vocabulary. In the Friends Overlap condition, each word that is a *friend* (a cognate or a loan word) is assigned a single, language-unspecific embedding allowing both Dutch and English to inform it. In the False Friends Overlap setting, the same holds for words that are false friends (but not friends). The other two conditions enforce either sharing of each form-overlapping word (Full Overlap) or a nearly complete separation of languages in the embedding space (Minimal Overlap).

We evaluate each vocabulary condition through the lens of two signals important in LM training as well as cross-lingual activation in humans: word frequency and language context (i.e. the language of the sentence).

Our results show that bilingual models generally separate the two languages, except when they are explicitly tied by a shared embedding of a word. In that case, the models show cross-lingual effects, specifically facilitation effects for friends as well as false friends compared to their respective control words. We find that this facilitation is driven by word frequency, and is not influenced by form-meaning mapping across languages. The only condition that can explain human data is Friends Overlap (i.e. facilitation for friends, but not for false friends). This suggests that while LMs can reproduce certain cross-lingual activation patterns, their behavior aligns with human bilingual reading only under specific vocabulary conditions[1].

2. Background and Related Work

2.1. Vocabulary Design

A large number of related studies use an equivalent of the Full Overlap condition when training a bilingual LM, so each overlapping surface form is shared between languages (Winther et al., 2021; Roslund and Matusevych, 2022; Oba et al., 2023; Constantinescu et al., 2025). Our vocabulary conditions are inspired by Kallini et al. (2025), who experiment with vocabulary manipulations to find that, for downstream tasks such as natural language inference and question answering, any sharing is beneficial (even that of false friends) in contrast to no overlap. Aoyama and Schneider (2024) reset the embedding layer before starting L2 acquisition, which essentially does not allow cross-lingual transfer between overlapping word forms.

In terms of the share of first (L1) and second (L2) language, some studies used a balanced proportion where each language has the same budget of training tokens (Oba et al., 2023; Kallini et al., 2025; Constantinescu et al., 2025), while others simulated larger L1 exposure compared to L2 (Roslund and Matusevych, 2022; Constantinescu et al., 2025). Winther et al. (2021) observe that the cognate facilitation effect in LMs trained on Dutch-English or Norwegian-English depends on the portion of L1/L2 and the presentation order of the languages, whereby an LM that is first trained on the L1 (75% training samples), and then on the L2 at the end of each epoch exhibits lower surprisal for cognates relative to non-cognates, mirroring the facilitation effect observed in bilinguals.

2.2. Role of Frequency and Context

Language models learn more frequent words earlier and better (Chang and Bergen, 2022; Razeghi et al., 2022). This provides a relevant link to cross-lingual transfer in bilinguals. According to the cumulative frequency hypothesis (Voga and Grainger, 2007; Strijkers et al., 2010; Midgley et al., 2011), the cognate facilitation effect in bilinguals stems from exposure to cognates regardless of the language: exposure strengthens the word's lexical representation due to the same form-meaning mapping. It is unclear how this fares with false friends, where the form-meaning mapping diverges across languages, i.e. whether we can expect cross-lingual transfer to be influenced by frequency.

Aside from word frequency, context is an important cue in LM training and inference. LMs learn how to use context information to disambiguate lexically ambiguous words within one language (Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020). One question that arises is whether a bilingual LM can use the language of the preceding context as a cue to process an overlapping word form without influence/transfer from the other language, or whether the fact that an overlapping form occurs in two language contexts makes it more ambiguous and less predictable.

[1]Our code is made available at https://github.com/izaskr/cross_lingual_transfer_dutch_english_forms.

3. Manipulation of Vocabulary Sharing

We design 4 conditions for the LM vocabulary to test the role of sharing/separation of embeddings of words with the same surface form across Dutch and English. When a word form is *shared* (see intersections in Figure 1), it has a single word embedding. If this form happens to appear in both Dutch and English, then samples from both languages contribute to the embedding during training. In case of *separation* (see the areas outside of the intersection in Figure 1), each word form is encoded for the language of the sentence it is in. Their embeddings are informed only by the samples in the respective language.

In condition Full overlap (A), there is no active intervention: each word form that appears in both languages has a single, language-unspecific embedding. In other conditions, we manipulate what is shared and what is separated.

In condition Friends Overlap (B), only friends (cognate and loan words) across the two languages are in the intersection. This condition is motivated by hypotheses that cognate words might share orthographic and (partially) semantic representations in a bilingual's mental lexicon, while false friends are represented by two different orthographic representations (Dijkstra and van Heuven, 2002; Lemhöfer and Dijkstra, 2004). All other words have language-specific embeddings.

Only false friends are placed in the intersection in condition False Friends Overlap (C). Each false friend has a single entry in the vocabulary despite different meanings across languages (e.g., *brand* means 'fire' in Dutch).

Finally, in Minimal Overlap (D), the word forms from the two languages are completely separated given the sentence language. A language-specific processing is assumed, regardless of overlapping form/meaning across languages.

Note that punctuation and named entities are placed in the intersection across all conditions. They are often not language-specific, refer to the same extra-linguistic entity and are thus not a part of our manipulation. Named entities were identified by spaCy's xx_ent_wiki_sm model[2] prior to tokenizer and LM training.

4. Implementation

4.1. Training Data

To imitate human language exposure, we train our LMs on a corpus of diverse genres: 49% of tokens

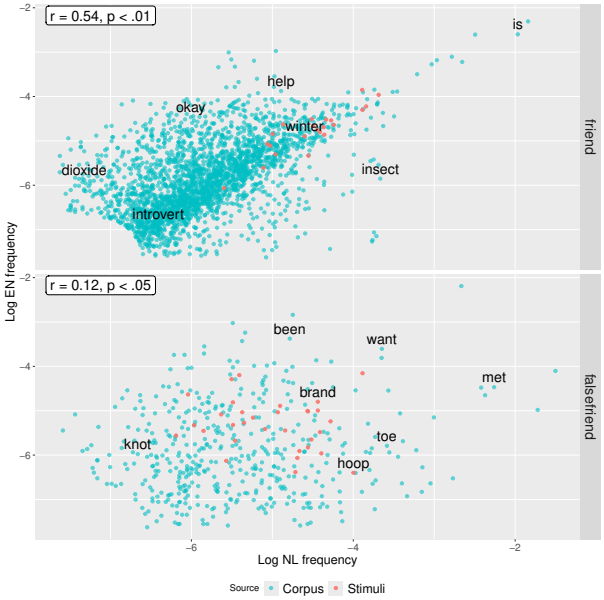

Figure 2: The frequency of each word in the Dutch and English portion of the LM training data. The frequencies are generally positively correlated: a word form that is frequent in Dutch is frequent in English, but more so for *friends* than *false friends*. The plot includes (false) friend words from psycholinguistic stimuli (Bultena et al., 2014; Huisman, 2025) as well as other compiled lists and our training data (here labeled as Corpus).

in the data come from non-fiction text (Wikipedia[3]), 26% from transcribed scripted speech (OpenSubtitles[4] and TedTalks[5]), and 25% from web-crawled data (CC100, Conneau et al., 2020)[6]. In total, each training corpus had about 400 million tokens. To simulate an unbalanced late Dutch-English bilingual, the training data consists mostly of Dutch (75% tokens), while the rest is English. In each epoch, Dutch samples were presented first, followed by English ones.

4.2. Overlapping Word Forms

To obtain a broad coverage of words sharing only form, or both form and meaning between Dutch and English, we annotated a list of words that appear in both the Dutch and English corpora as friends (cognates or loan words) and false friends (interlingual homographs). We joined our list with existing ones (Bultena et al., 2014; Poort and Rodd, 2019;

[2] https://spacy.io/models/xx

[3] https://dumps.wikimedia.org/nlwiki/latest

[4] https://opus.nlpl.eu/OpenSubtitles/nl&en/v2024/OpenSubtitles

[5] https://object.pouta.csc.fi/OPUS-NeuLab-TedTalks/v1/tmx/en-nl.tmx.gz

[6] https://data.statmt.org/cc-100

Lefever et al., 2020; Huisman, 2025) to obtain a set of 2,806 friends and 511 false friends, both referring to word types without named entities. The Dutch and English frequencies of each word have a significant positive correlation (Figure 2).

Some words are homographs within one language and can be both a friend and a false friend (e.g. *monster* in Dutch can have the English meaning as well as 'sample'). These words are excluded from our vocabulary manipulation.

Class annotations are available for a subset of the items, comprising of 1,598 friends and 379 false friends: all cognate and loan words are annotated as content words in both languages. Among false friends, 4% have different classes in Dutch and English (e.g. *toe* and *met* are content words in English, but function words in Dutch).

4.3. Psycholinguistic Stimuli

We evaluate the LMs on stimuli from two reading studies, in which the participants were late bilinguals with Dutch as their first language and English as their second. In the *friend study* (Bultena et al., 2014), participants read English sentences with target words that were either Dutch-English friends (cognates) (1a) or non-cognate controls (1b)[7].

1a. The residents dislike the *winter* for the ...

1b. The residents dislike the *prison* for the ...

The friend and control targets were matched with respect to word length and English word frequency. In total, the stimuli consisted of 22 item sentences (with two conditions per item). The context preceding the target words was designed to not be semantically constraining and biased towards the meanings of target words. The analysis of reading times by Dutch-English bilinguals revealed a cognate facilitation effect, with faster reading for cognates than the corresponding control words.

The stimuli in the *false-friend study* (Huisman, 2025) include word pairs of false friends and controls that were presented to participants in Dutch sentences, that is, in their first language, in a self-paced reading paradigm. Again, each sentence had two conditions: with a false friend (2a) or with a language-unique control word (2b):

2a. De beelden van de *brand* zullen hen ...

2b. De beelden van de *kogel* zullen hen ...

All target words were nouns and all false friends were form-identical between Dutch and English. Control words were paired with false friends to match them on word length and Dutch and English word frequency. The stimuli by Huisman (2025) consists of 32 item sentences, all designed to be semantically non-constraining in the context preceding the target words. While this study did not find significant reading time differences between false friends and control words in bilinguals (and thus no cross-language activation), the manipulation provides a valuable evaluation of the LMs tested here, particularly due to the lack of prior work on false-friend reading in Dutch-English bilinguals.

[7]We only use items where the cognate is a form-identical noun.

4.4. Tokenizers

The tokenization process included multiple components. All words annotated as *friends* or *false-friends* (Section 4.2) as well as the language-unique control words (Section 4.3) were tokenized as single subwords across all conditions to avoid biases from multi-subword segmentation (Lesci et al., 2025). We train a byte-level BPE tokenizer on combined Dutch-English train set (64k vocabulary, minimum frequency of 2) and used it to tokenize all remaining words. We also train a byte-level BPE tokenizer on named entities (10k vocabulary, minimum frequency of 2).

4.5. Transformer Language Models

For each vocabulary manipulation, we train a separate Transformer model with the causal language modeling task, following the GPT2-small configuration. In a single epoch, the model was first presented Dutch samples (300 million tokens), followed by English ones (100 million tokens) creating a sequential regime. Each LM was trained for 2 epochs. See Appendix C for vocabulary sizes and parameter counts under different conditions. Appendix D shows loss curves on the training and test sets.

5. Vocabulary Interventions in Bilingual LMs

5.1. Language Context

We first explore how (dis)similar the preceding contexts of the same surface form are across Dutch and English. If the LM "understands" that a friend word carries the same meaning in Dutch and English, this should be reflected in the similarity of its contextual and word embeddings across languages. In the same vein, if the LM represents the language-specific meanings of false friends, their context and word embeddings should be dissimilar as well. In conditions that assign each overlapping word form a shared embedding (Friends Overlap for friends; False Friends Overlap for false friends; Full Overlap for both), it might be easier for the LM

to learn friends as cross-lingual meaning equivalents, but it might be more difficult to tease apart the language-specific meanings of false friends.

Method. For each word t in the stimuli set of friends and false friends T, we sample a set of 500 sentences containing that word from the training data from each language (yielding 1,000 sentences per word in total). For each sentence in the sample, we calculate the mean-pooled embedding of the preceding context of word t (without the word itself). For each language-specific subcorpus, we average across these embeddings, obtaining $\mu_C^{NL}(t)$ and $\mu_C^{EN}(t)$.

For each sentence in the sample, we also obtain the contextualized embedding of the target word t itself. We then average across the sentences per language, yielding $\mu_W^{NL}(t)$ and $\mu_W^{EN}(t)$ for Dutch and English sentences, respectively.

We measure the similarity of these embeddings between languages using cosine similarity[8]. The similarity between $\mu_C^{NL}(t)$ and $\mu_C^{EN}(t)$ tells us how similar the preceding contexts of the same word are across languages. In contrast, cosine similarity between $\mu_W^{NL}(t)$ and $\mu_W^{EN}(t)$ estimates the similarity of the target word embedding across languages. See Appendix E for illustrative diagrams. We use the embeddings from layer 12. Results from layer 5 show very similar patterns.

Results. Figure 3 shows that the contexts for the same word in Dutch and English are generally estimated as dissimilar and vary little across vocabulary conditions, i.e. manipulating what is shared in the embedding space does not lead to more similarities or differences for the context preceding target words (upper panel). Focusing on the embeddings of the target words themselves (lower panel), the results indicate that separating the languages for overlapping word forms leads to lower similarity for the word affected by separation: in the Friends Overlap condition, false friends have language-specific vocabulary entries. The distance between Dutch and English sentences containing these words is larger than when each false friend has a single embedding. Friends, likewise, show a lower similarity when their embeddings are language-specific (False Friends Overlap condition). In the Minimal Overlap condition, the contexts as well as word embeddings stay apart and dissimilar across Dutch and English. Generally,

[8]Due to known issues with anisotropy in contextualized representations (Ethayarajh, 2019), we standardize the embeddings following Timkey and van Schijndel (2021). We estimate a mean and standard deviation embedding from a sample of 30k instances and standardize the embeddings before pooling.

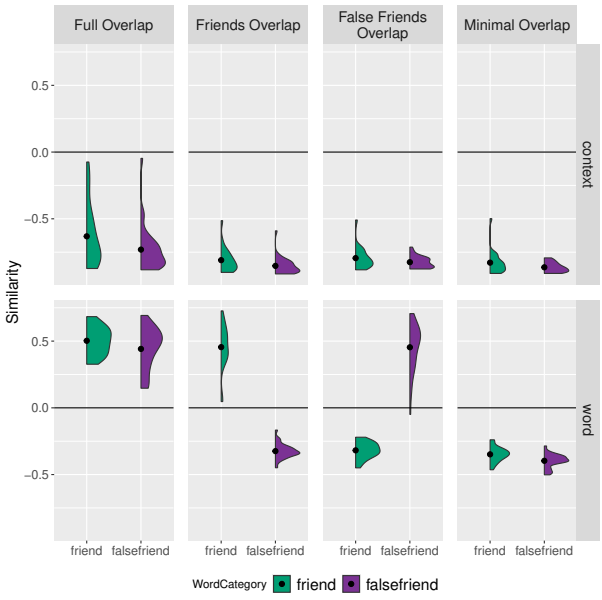

Figure 3: Cosine similarity between Dutch and English across the context and word-only embeddings for training data samples for target words (friends or false friends). A larger similarity means the LM considers Dutch and English embeddings more similar.

this result shows that the contexts are coded as dissimilar between Dutch and English, and the target word embeddings are similar when each is represented by a single, language-unspecific embedding, no matter the form-meaning match in Dutch and English.

5.2. Processing Effort via Surprisal

We further focus on the predictability of overlapping word forms in comparison to language-unique words given the same preceding sentence context by estimating their surprisal. We use word surprisal, $surp(w) = -\log_2 p(w|context)$, as a computational correlate of processing effort as observed in humans with lower surprisal corresponding to easier processing (e.g. Hale, 2001; Levy, 2008; Shain et al., 2024).

Bilingual speakers tend to show different behavior for language-unique words and words with overlapping surface form. Specifically, reading a word with form overlap across languages might show patterns of cross-language activation: facilitation for friends (faster reading in comparison to controls), and interference for false friends (slower reading in comparison to controls). We expect lower surprisal for friends than controls, reflecting the ease of processing of friends over controls. In contrast, we expect lower surprisal for controls than false friends.

With respect to the vocabulary configurations, we examine how sharing/separation of embeddings affects the form-meaning correspondence of overlapping word forms. On the one hand, in conditions where each overlapping surface forms has a single embedding, this word might be more ambiguous as it appears across Dutch and English contexts, possibly reducing its predictability in a given sentence. If so, shared forms should have higher surprisal than language-unique controls. On the other hand, a shared embedding receives more training signal than two single ones and more than a frequency-matched control, which could result in reduced surprisal compared to controls. We use $\alpha = .05$ in all statistical tests.

Method. Using sentence stimuli from the *friend study* and the *false-friend* study (see Section 4.3), we estimate surprisal for each target word. Importantly, in each item, the target word (either overlapping or control) is preceded by the same context. We test for effects of processing effort by comparing the surprisal of overlapping words to that of controls.

We use mixed-effects regression in `lme4` (Bates et al., 2015) with surprisal as the outcome, and word category and word position in sentence (index) as fixed effects, and a random intercept for sentence pair (item)[9]. We fit a regression model for each vocabulary condition separately for the two studies, resulting in 8 regression models in total.

Results. Figure 4 shows surprisal estimates across the two word categories and vocabulary conditions. For *friends*, we find that they have significantly lower surprisal than control words in two conditions: Full Overlap, and Friends Overlap. In both conditions, each friend surface form has a single embedding, which increases the word's predictability relative to controls. For *false friends*, we find that in the Full Overlap and False Friends Overlap conditions, their surprisal is significantly lower than that of controls. Again, in these conditions, the false friend's surface form has a single embedding.

Taken together, both types of overlapping word forms (friends and false friends) show smaller surprisal relative to language-unique controls when their surface forms are represented by a single shared embedding. These results indicate that embedding sharing increases the predictability of overlapping surface forms regardless of their form-meaning correspondence. Surprisingly, neither word category show lower predictability, either because of language context ambiguity or form-meaning divergence.

[9]Model structure: $surp \sim WordCategory + WordIndex + (1|Item)$. WordCategory is sum-coded (1 friend/falsefriend, −1 control).

5.3. Frequency Benefit across Languages

In bilingual training, overlapping surface forms receive additional exposure relative to language-unique controls due to their occurrence in both languages. This raises the question whether increased frequency drives improved predictability of words with overlapping surface forms across languages (i.e. as per the cumulative frequency hypothesis) or whether their predictability might depend on the language of the sentence. This analysis builds on the facilitation effects observed in Section 5.2 focusing on the effect of a word's frequency in Dutch vs. English on its surprisal.

The two frequency measures may play different roles. In the *friend study*, the target words appear in English sentences, making Dutch the "other" language. In the *false-friend study*, the sentences are in Dutch, so English serves as the "other" language. This analysis examines whether both the sentence-language frequency ($\text{freq}_S(w)$) and the other-language frequency ($\text{freq}_O(w)$) influence surprisal of word w.

Method. Both friends and false friends exhibited facilitation relative to controls in vocabulary conditions where their embeddings were shared between Dutch and English. Thus, to compare the two frequency measures ($\text{freq}_S(w)$ and $\text{freq}_O(w)$), we consider the Full Overlap and Friends Overlap conditions for the *friend study*, and Full Overlap and False Friends Overlap for the *false-friend study*. The frequency estimates of the target words are based on the Dutch and English portions of the LM training data; they are log-transformed and normalized. For each study, we start by fitting a regression model with surprisal estimated by word category, normalized word index and vocabulary condition[10]. In the next step, we add freq_S to the model[11], and finally freq_O[12]. We compare these regression models using likelihood ratio tests to determine whether the addition of the frequency terms provides a better fit.

Results. The results for the *friend study* show a main effect of word category as facilitation for friends over controls ($\beta = -.5$, $p < .01$). Upon adding word frequency in English (freq_S), the word category effect was still negative, but not a reliable effect ($\beta = -.2$, $p = .06$). There is a main

[10]The maximal model that would converge was $surp \sim WordCategory + Vocabulary + WordIndex + (1|Item)$.

[11]$surp \sim freq_S + WordCategory + Vocabulary + WordIndex + (1|Item)$

[12]$surp \sim freq_O + freq_S + WordCategory + Vocabulary + WordIndex + (1|Item)$. Interaction terms did not improve the model fit for any regression models.

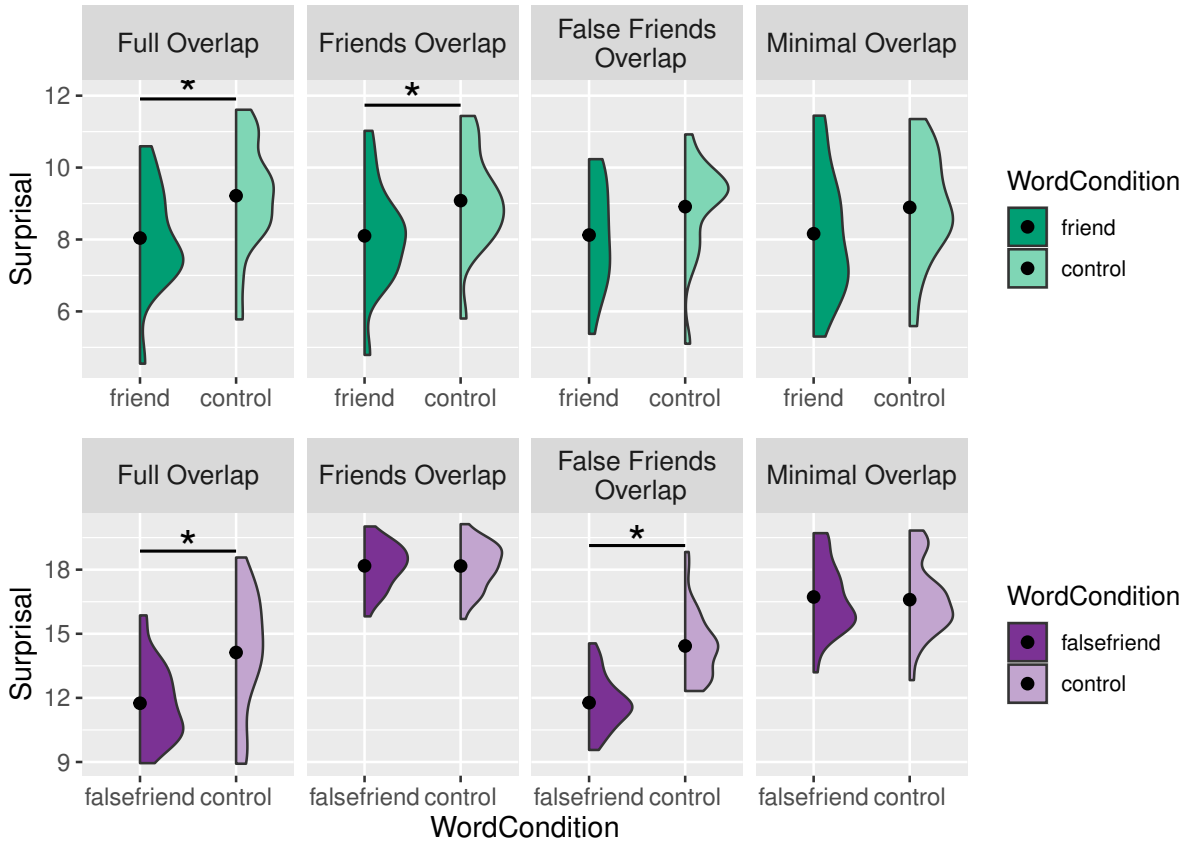

Figure 4: Surprisal estimates for friends/controls and false friends/controls from LMs trained under different vocabulary conditions. The symbol $*$ indicates a statistically significant difference ($p < .05$) between the two word types.

effect of English frequency ($\beta = -1.1$, $p < .01$), indicating that higher frequency leads to smaller surprisal. Finally, including Dutch word frequency (freq_O), we observe that English frequency is still a significant predictor ($\beta = -1.2$, $p < .01$), but Dutch frequency is not ($p = .9$) and neither is word category ($p = .4$). The model comparison revealed that including Dutch word frequency in the model does not contribute to the fit ($\chi^2(1) = .002, p = .97$).

For the *false-friend study*, we find a facilitation effect for false friends compared to controls ($\beta = -1.3$, $p < .01$). When Dutch frequency (freq_S) is included in the model, there is still a significant facilitation effect for false friends ($\beta = -1.2$, $p < .01$) as well as a significant main effect of Dutch frequency ($\beta = -.5$, $p < .01$). Finally, when English frequency (freq_O) is added to the model, Dutch frequency remains a significant predictor ($\beta = -.5$, $p < .05$), and English frequency is significant as well ($\beta = -.7$, $p < .05$), but the difference between false friends and controls is no longer significant ($\beta = -.6$, $p = .06$). Model comparison further indicates that including English word frequency significantly improves model fit ($\chi^2(1) = 4.49, p = .034$).

Overall, these results indicate that effects of bilingual exposure differ between L1 and L2. In the *friend study*, the facilitation effect for friends over controls is accounted for by sentence-language frequency (English), with seemingly no contribution from the other-language frequency (Dutch). However, correlation of fixed effects sheds light on this: there is a large correlation between word category and Dutch frequency ($r = -.9$), essentially signaling to the regression model that words that do not appear in Dutch (i.e. have a Dutch frequency of 0) are English control words. For friends, the English and Dutch frequencies are highly correlated in the model ($r = -.7$), as well as in the data (see Figure 2). This means that English frequency sufficiently predicts variance between control and friend words. The variance that could be explained by Dutch frequency is already explained by the English one, making Dutch frequency redundant.

In contrast, the facilitation effect found for the *false-friend study* seems to be driven by both the frequency in the language of the sentence and frequency in the "other" language. In this model, word category and English frequency are correlated as

well ($r = -.9$), again pertaining to Dutch control words that have a 0 frequency in English. However, the model correlation between Dutch and English false friends frequency is weak ($r = -.15$, see also Figure 2 for data correlation). This is not surprising since in the case of false friends, the same surface form is mapped to different meanings and possibly to different parts of speech, with different distributional patterns across Dutch and English. We hypothesize that this is the reason both Dutch and English frequencies remain significant in the *false-friend study*.

Generally, our results provide support for the cumulative frequency hypothesis for both friend and false friend words. Both languages contribute to the facilitation effect, but for friends it is harder to disentangle the effects of Dutch and English as frequency sources due to their high correlation, especially across highly similar languages (Schepens et al., 2013).

6. Discussion

This work asked whether cross-lingual transfer in bilingual Transformer LMs produces patterns that resemble cross-lingual activation in bilingual readers, focusing on words with overlapping surface forms across Dutch and English. To test this, we trained LMs on L1 Dutch and L2 English, and evaluated the models on experimental stimuli from studies of Dutch-English bilinguals, which included words with overlapping surface forms between the two languages. Designing a cognitively plausible vocabulary in bilingual LMs is an open question. We addressed it by manipulating lexical sharing directly by controlling which form-overlapping words received an embedding shared between languages versus language-specific embeddings.

We find that the bilingual LM keeps the two languages apart in the embedding space, but cross-lingual effects do occur via embedding sharing: friends (i.e. cognate and loan words) show lower surprisal than language-unique control words in the conditions where either each form overlapping word is shared between Dutch and English, or only friends are shared between the two languages. Similarly, false friends have lower surprisal than controls either in conditions where sharing holds for all form overlapping words or just for false friends. These facilitation effects are largely explained by frequency-based signals.

The vocabulary condition that matches human data is one where only friends are shared between Dutch and English: the LM in this condition exhibits facilitation effects for friends without the corresponding facilitation for false friends. Other conditions fail to resemble human bilinguals with respect to processing the different word types: 1) the conditions with full overlap or just for false friends show facilitation for false friends, 2) the condition with minimal overlap (named entities and punctuation) does not show cross-lingual effects. Crucially, simulations of bilingual reading with LMs typically assume full surface overlap (see Section 2.1), which may bias conclusions about cross-lingual transfer and L2 learning.

Among the conditions tested, the one sharing only friends has the best *descriptive adequacy* with respect to empirical findings of bilingual reading of overlapping words. However, it lacks *explanatory adequacy* (Jacobs and Grainger, 1994; Frank, 2021): the tokenizer and LM are based on explicitly coded rules about which forms are friends (and should each have a single embedding) and which are not. We cannot assume this is the mechanism behind cognate facilitation effect found in bilinguals. Importantly, humans learn meaning from grounded language learning, sensorimotor and social cues, and not (just) streams of text (Warstadt and Bowman, 2022; Chang and Bergen, 2022). Yet, when the LM was left to train without explicit coding of (false) friends, it resulted in behavior that does not reflect that of bilingual humans.

One of the open questions is why shared embeddings result in facilitation for false friends, rather than interference. One possibility is that surprisal in unconstrained sentence contexts mostly reflects the LM's ability to predict the surface form given the few semantic and lexical cues, and is as such mostly informed by frequency and distributional patterns rather than semantic competition between meanings. Another possibility might lie in the stimuli language (L1, so Dutch) and exposure. The LMs were largely trained on Dutch, while only a quarter of the data was in English. Perhaps the LM did not have a chance to properly learn the English meaning, and therefore does not exhibit interference effects when processing a sentence in Dutch. Generally, bilinguals tend to show cross-lingual effects in L2 reading and less so in L1 (Titone et al., 2011).

One of the main findings confirms the role of frequency in overlapping word forms specifically in the conditions with shared embeddings. When an embedding is shared, it is trained on data from both languages, which increases its effective frequency and typically reduces surprisal. This frequency benefit appears regardless of whether the word is a friend or a false friend. This suggests that overlap primarily provides a learning advantage through shared counts, not through semantic consistency.

The present study uses the GPT2-small Transformer architecture. Cross-lingual activation patterns in cognate facilitation have been observed across a range of architectures, including LSTM models (Winther et al., 2021), Transformers (albeit

shallower than GPT2-small), and simple recurrent networks (Roslund and Matusevych, 2022). This suggests that cognate facilitation is driven less by architectural choice than by training data and procedure: specifically, unbalanced exposure (with greater L1 than L2 input) consistently gives rise to a cognate facilitation effect (Winther et al., 2021; Roslund and Matusevych, 2022).

Furthermore, Winther et al. (2021) report that the order of L1 and L2 presentation plays a crucial role: the cognate effect emerges either when the model is first pretrained on L1 and subsequently trained on a mix of L1 and L2, or (in the absence of pretraining) when L1 data comprises the first 75% of the epoch and L2 data the remainder. We follow the latter training procedure, but it would be valuable to test the former with different vocabulary designs as well.

Our results can be situated within the framework of the BIA+ model (Dijkstra and van Heuven, 2002), which is one of the most influential computational accounts of bilingual word recognition. The model includes orthographic, phonological and semantic lexical representations with bidirectional connections between the representation types. In contrast, in our LMs, word representations are encoded at the embedding layer, where a word form either has a single (shared) or two language-specific embeddings. With this difference in mind, the vocabulary condition that best matches human bilingual data is broadly consistent with the BIA+ assumptions that cognates have a special shared representation due to their overlap in both form and meaning, while interlingual homographs (false friends) are represented by two separate orthographic entries. The role of form frequency in our results relates to the BIA+ model as well, where frequency is implemented as resting activation levels. Importantly, in BIA+, frequency and form overlap are separable sources of facilitation, while they are harder to disentangle in our LMs, where sharing an embedding both increases a word's effective frequency and enables cross-lingual transfer at the same time. Further comparisons between BIA+ and bilingual LMs, for instance using tasks that more directly probe semantic competition or that vary sentence context constraints, could shed more light on how well these two types of models align in explaining cross-lingual activation.

7. Conclusion

We presented a controlled study of vocabulary sharing in bilingual Dutch-English Transformer language models, targeting the question of whether bilingual LMs exhibit human-like cross-language activation. The results show that bilingual LMs generally keep Dutch and English contexts distinct, and cross-lingual effects arise primarily when the model shares embeddings. Under embedding sharing, overlapping forms become more predictable than language-unique controls, resulting in facilitation for both friends and false friends. This facilitation is largely explained by frequency and does not depend on whether the form-meaning mapping is shared across languages.

Among the evaluated vocabulary conditions, only the condition with shared embeddings for cognate and loan words reproduces the qualitative human pattern of facilitation for these words. Overall, these findings suggest that neural language models can reproduce some cross-lingual activation patterns, however, their alignment with bilingual reading behavior depends critically on how lexical overlap is treated in the encoding step.

8. Limitations

Our findings are currently based on a small-scale manipulation, which affected between 2.3% and 4.3% of the vocabulary depending on the condition. We did not explore how lexical interventions affect other aspects of cross-lingual transfer, e.g. syntactic processing.

We tested the models on two sets of psycholinguistic stimuli, but only the *friend study* (Bultena et al., 2014) has previously shown effects in bilinguals. The stimuli from the *false-friend* experiment (Huisman, 2025) did not produce cross-lingual effects, limiting the extent to which model behavior can be compared to behavioral data. This points to a general challenge of scarcity of bilingual human responses to test the human-likeness of LMs.

To measure cross-lingual effects in LMs, we use surprisal and similarity between representations, but other measures might capture patterns that correspond to human behavior, for example attention or the activation from the feed-forward networks (Oh and Schuler, 2022; Kuribayashi et al., 2025).

9. Acknowledgements

We thank Martin van Harmelen and Mayank Jobanputra for their help in implementing tokenizer and language model training. We also thank two anonymous reviewers for their constructive feedback.

Iza Škrjanec is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

Iza Škrjanec benefited from a Short-Term Scientific Mission funded by COST Action MultiplEYE (CA21131), supported by COST (European Cooperation in Science and Technology).

10. Bibliographical References

Tatsuya Aoyama and Nathan Schneider. 2024. Modeling nonnative sentence processing with L2 language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4927–4940, Miami, Florida, USA. Association for Computational Linguistics.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Sybrine Bultena, Ton Dijkstra, and Janet G. van Hell. 2014. Cognate effects in sentence context depend on word class, L2 proficiency, and task. *Quarterly Journal of Experimental Psychology*, 67:1214–1241.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2025. Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, 13:96–120.

Ton Dijkstra and Walter van Heuven. 2002. The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5:175 – 197.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Stefan L. Frank. 2021. Toward computational models of multilingual sentence processing. *Language Learning*, 71:193–218.

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

Liv J. Hoversten and Matthew J. Traxler. 2016. A time course analysis of interlingual homograph processing: Evidence from eye movements*. *Bilingualism: Language and Cognition*, 19(2):347 – 360.

Merit Huisman. 2025. The role of phonology on recognising interlingual homographs in L1 sentence context. Master's thesis, Radboud University, Nijmegen, the Netherlands.

Arthur M. Jacobs and Jonathan Grainger. 1994. Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, 20(6):1311–1334.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Julie Kallini, Dan Jurafsky, Christopher Potts, and Martijn Bartelds. 2025. False Friends are not foes: Investigating vocabulary overlap in multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21138–21154, Suzhou, China. Association for Computational Linguistics.

Judith F. Kroll, Paola E. Dussias, Cari A. Bogulski, and Jorge R. Valdes Kroff. 2012. Juggling two languages in one mind: What bilinguals tell us about language processing and its consequences for cognition. volume 56 of *Psychology of Learning and Motivation*, pages 229–262. Academic Press.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *Transactions of the Association for Computational Linguistics*, 13:1743–1766.

Justin Lauro and Ana I. Schwartz. 2017. Bilingual non-selective lexical access in sentence contexts: A meta-analytic review. *Journal of Memory and Language*, 92:217–233.

Els Lefever, Sofie Labat, and Pranaydeep Singh. 2020. Identifying cognates in English-Dutch and

French-Dutch by means of orthographic information and cross-lingual word embeddings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4096–4101, Marseille, France. European Language Resources Association.

Kristin Lemhöfer and Ton Dijkstra. 2004. Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition*, 32:533–550.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. 2025. Causal estimation of tokenisation bias. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28325–28340, Vienna, Austria. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106:1126–1177.

Maya Libben and Debra A. Titone. 2009. Bilingual lexical access in context: evidence from eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(2):381–90.

Katherine J. Midgley, Phillip J. Holcomb, and Jonathan Grainger. 2011. Effects of cognate status on word comprehension in second language learners: An erp investigation. *Journal of Cognitive Neuroscience*, 23(7):1634–1647.

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Irina Pivneva, Julie Mercier, and Debra Titone. 2014. Executive control modulates cross-language lexical activation during l2 reading: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(3):787–796.

Eva D. Poort and Jennifer M. Rodd. 2019. A Database of Dutch-English Cognates, Interlingual Homographs and Translation Equivalents. *Journal of Cognition*, 2(1):15.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rasmus Roslund and Yevgen Matusevych. 2022. Modeling sentence processing effects in bilingual speakers: A comparison of neural architectures. In *Annual Meeting of the Cognitive Science Society*.

Job Schepens, Ton Dijkstra, Franc Grootjen, and Walter J. B. van Heuven. 2013. Cross-language distributions of high frequency and phonetically similar cognates. *PLOS ONE*, 8(5):1–15.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Kristof Strijkers, Albert Costa, and Guillaume Thierry. 2010. Tracking lexical access in speech production: Electrophysiological correlates of word frequency and cognate effects. *Cerebral Cortex*, 20(4):912–928.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Debra A. Titone, Maya Libben, Julien Mercier, Veronica Whitford, and Irina Pivneva. 2011. Bilingual lexical access during l1 sentence reading: The effects of l2 knowledge, semantic constraint, and l1-l2 intermixing. *Journal of experimental psychology. Learning, memory, and cognition*, 37 6:1412–31.

Madeleine Voga and Jonathan Grainger. 2007. Cognate status and cross-script translation priming. *Memory & Cognition*, 35(5):938–952.

Alex Warstadt and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*.

Irene Elisabeth Winther, Yevgen Matusevych, and Martin J. Pickering. 2021. Cumulative frequency can explain cognate facilitation in language models. In *Annual Meeting of the Cognitive Science Society*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.

Appendix

A. Translations

English translations of Dutch sentences in Section 4.3.

2a The images of the fire ...

2b The images of the bullet ...

B. Language Model Hyperparameters

The context window size was reduced to 256. We used gradient accumulation with the effective batch size of 512. Weight decay was set to 0.1, the learning rate to 5e-4 with a cosine learning rate scheduler. The initial 1k steps of training were used as warm-up. All models were trained under the same random seed.

C. Vocabulary and LM Sizes

Condition	Vocabulary size	Num. parameters
A	77,369	144,672,000
B	141,877	194,214,144
C	144,170	195,975,168
D	144,681	196,367,616

Table 1: Vocabulary size and number of trainable parameters for the language model trained in each condition.

D. Training and Evaluation Loss

Figure 5 shows the loss values on the training and test sets at the end of each of 6 epochs in each condition. The training loss combines the values for the Dutch and English portion of the training set. We show separate values for the test set of each language. At the very end of each epoch, the LMs have been exposed to English more recently, so the test loss for English is quite low in all conditions, but high for Dutch, which indicates some level of attrition of Dutch. Compared to the other conditions, condition A (Full Overlap) shows a relatively higher training loss, but also lower test loss for Dutch.

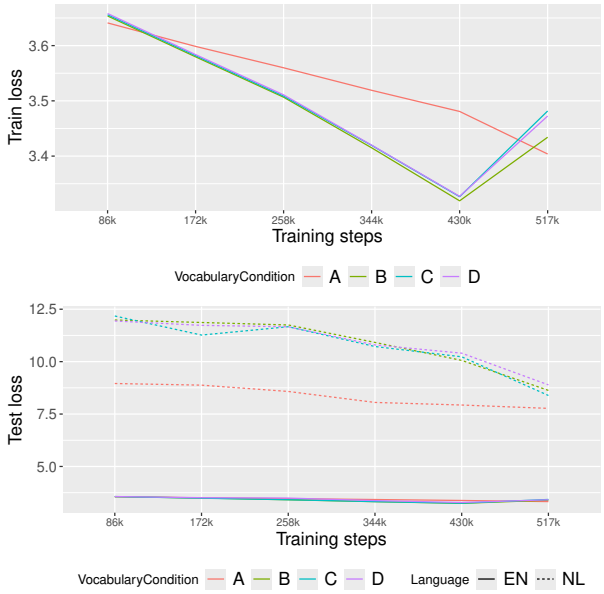

Figure 5: Training and evaluation loss.

E. Diagrams for 5.1

Figure 6 illustrates mean pooling over the preceding context of target words to obtain μ_C. Figure 7 shows how we obtained the mean-pooled token representation μ_W.

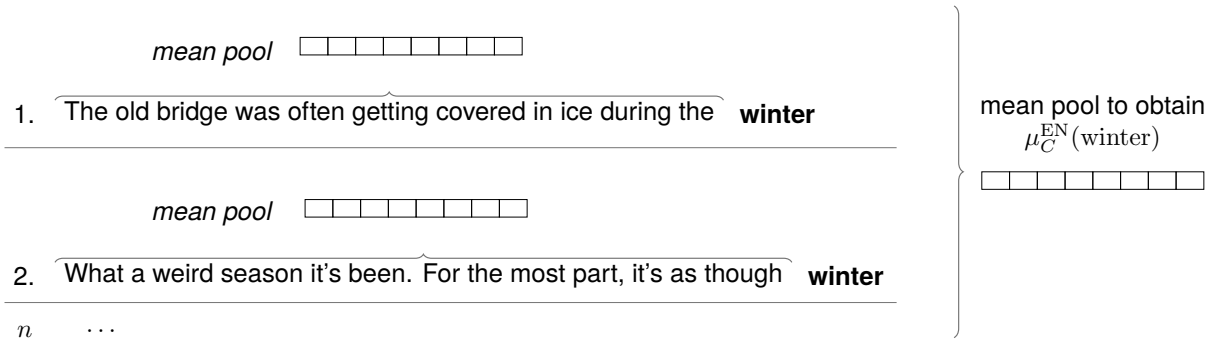

Figure 6: Mean pooling over the embeddings in the preceding tokens withing each sentence ($n = 500$). Mean pooling over these to obtain a context embeddings from English samples μ_C^{EN} for the token *winter*. The same was performed over Dutch sentences to obtain $\mu_C^{NL}(winter)$.

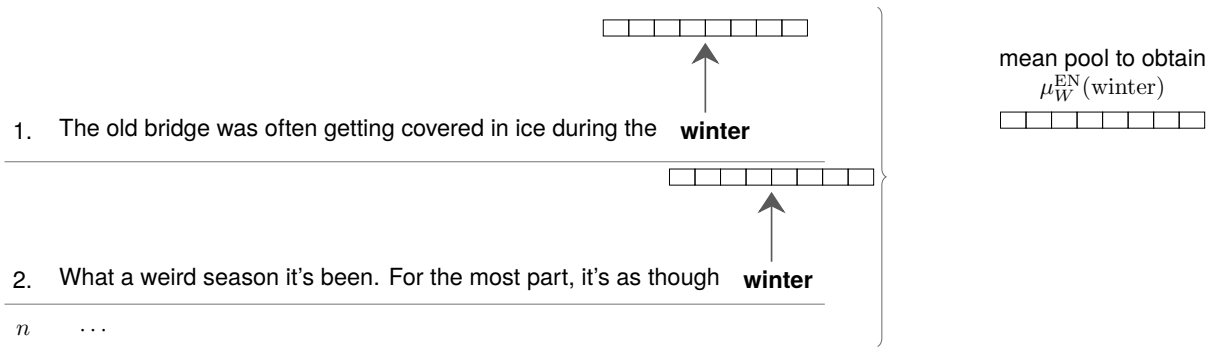

Figure 7: Mean pooling over the token representations of the target word in n English sentences ($n = 500$) to obtain $\mu_W^{EN}(winter)$. Similarly, n Dutch sentences are used to calculate $\mu_W^{NL}(winter)$.