Modeling Nonnative Sentence Processing with L2 Language Models

Tatsuya Aoyama   **Nathan Schneider**
Georgetown University
{ta571, nathan.schneider}@georgetown.edu

Abstract

We study LMs pretrained sequentially on two languages ("L2LMs") for modeling nonnative sentence processing. In particular, we pretrain GPT2 on 6 different first languages (L1s), followed by English as the second language (L2). We examine the effect of the choice of pretraining L1 on the model's ability to predict human reading times, evaluating on English readers from a range of L1 backgrounds. Experimental results show that, while all of the LMs' word surprisals improve prediction of L2 reading times, especially for human L1s distant from English, there is no reliable effect of the choice of L2LM's L1. We also evaluate the learning trajectory of a monolingual English LM: for predicting L2 as opposed to L1 reading, it peaks much earlier and immediately falls off, possibly mirroring the difference in proficiency between the native and nonnative populations. Lastly, we provide examples of L2LMs' surprisals, which could potentially generate hypotheses about human L2 reading.

1 Introduction

It has been widely shown that one's first language (L1) affects both second language (L2) production (e.g., Murakami and Alexopoulou, 2016) and processing (e.g., Clahsen and Felser, 2006), a phenomenon called *L1 transfer*. In computational linguistics, most studies of LMs as models of acquisition (e.g., Huebner et al., 2021) have considered monolingual settings. To date, Yadavalli et al. (2023) and Oba et al. (2023) seem to be among the rare exceptions that have investigated LMs as models of second language acquisition (SLA), which we refer to as L2 Language Models (**L2LMs**). Both of these studies establish that the typological distance between the model's first language and its second language (English) correlates with its English performance as measured by a morphosyntactic benchmark. However, these L2LMs are yet to be tested against human L2 speakers. As one way of testing if the L1 of L2LMs affects L2LMs' performance on English in a humanlike manner, we study their sentence processing.

Sentence processing is widely used to study (L)LMs' cognitive plausibility (e.g., Oh et al., 2022; Oh and Schuler, 2022, 2023a; Kuribayashi et al., 2021, 2022; Wilcox et al., 2023, *inter alia*). By comparing the LM *surprisal* (Hale, 2001; Levy, 2008) to human reading time, we determine whether humans and models process given texts in a similar manner. Studies in this area have investigated the impact of data size, model size, and model architecture, among other variables (see §2).

In this study, we train autoregressive L2LMs from scratch to investigate the following questions (§3): (1) Does the L1 effect on L2LMs' L2 grammaticality discrimination, previously demonstrated for encoder-only models, extend to decoder-only models, namely GPT2? We hypothesize that, as in previous studies, the L2LMs trained with L1s closer to the L2 (English) will perform better. (2) Do L2LM surprisals predict reading time of human English speakers of different L1 backgrounds? We hypothesize that L2LM best predicts L2 reading time when the L2LM and human L1s match. Our findings (§4) are, in brief:

1. The L1 chosen for pretraining does impact the L2LM's English perplexity and performance on the morphosyntactic benchmark (BLiMP), with closer-to-English languages generally helping more, echoing prior findings about encoder-only models.
2. Contrary to our hypothesis, matching L1s between L2LMs and humans has little effect on the accuracy of models' human reading time predictions, which are largely dominated by the main effect of human L1 alone.
3. As we show with selected examples, L2LM surprisals could potentially generate hypothe-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4927–4940
November 12-16, 2024 ©2024 Association for Computational Linguistics

ses about human L2 reading behaviors.[1]

2 Related Work

2.1 LMs and (Second) Language Acquisition

With the ever-growing training data and model sizes of modern large language models (LLMs), their cognitive plausibility has been garnering attention. Huebner et al. (2021) is among the earlier attempts to train a more cognitively plausible LM, and they show that BabyBERTa, a downsized RoBERTa (Liu et al., 2019b) trained on a substantially smaller amount of data, achieves competitive performances on various linguistic tasks. Warstadt and Bowman (2024) point out that many LLMs are trained on data that are orders of magnitude larger than the realistic input humans are exposed to. A shared task called the BabyLM Challenge (Warstadt et al., 2023; Choshen et al., 2024) promotes evaluation of models trained with data quantities on par with child exposure.

Another factor that plays an important role in understanding the language acquisition of LMs is inductive bias. For example, McCoy et al. (2020a) investigated the effect of various architectural factors (e.g., choice of recurrent unit, attention type, and explicit tree structure in the model) on the way LMs process ambiguous input, and found that, among the various factors they studied, the presence of an explicit tree structure in the encoder and decoder was the only factor that consistently led to LMs' preference for hierarchical generalization.

Other works study inductive biases as a trainable set of parameters (e.g., McCoy et al., 2020b). Of particular relevance to this study is the work by Papadimitriou and Jurafsky (2020), where they propose a method called TILT (test of inductive bias via language transfer). Training LMs on various first "languages", including music scores, artificial and natural languages, and then on the second language (Spanish), they find that all of them improved learning of the second language, with natural language pretraining showing the best result.

Yadavalli et al. (2023) used TILT to test positive and negative language transfer by comparing how pretraining an LM on various L1s affects the performance on L2 (English). They find that the effect of the L1 training on L2 performance largely correlated with the L1's linguistic distance from L2, with the pretraining on English exhibiting the best performance on English, followed by German, French, Polish, Japanese, and Indonesian, in descending order. Oba et al. (2023) report similar results, where the order was German, French, Japanese, and Russian. Both of these studies rely on BLiMP for evaluation, comparing the experimental results against typological/linguistic distance between L1s and L2. While this experimental setup is reasonable and has its pros (e.g., no noise from human data), our study focuses on adding human performance as a reference to study L2LMs and the inductive biases added through the L1 pretraining. For this purpose, we use sentence processing as the primary evaluation, which we now turn to.

2.2 Sentence Processing

Levy (2008) posits that any realistic theory of human sentence comprehension must account for processing difficulty. He proposes the *resource-allocation* account of sentence processing, which maintains that the processing difficulty corresponds to the amount of resource reallocation needed. This account is found to be the equivalence of *surprisal theory* (Hale, 2001), a probabilistic account of cognitive effort. Both Hale (2001) and Levy (2008) provide empirical support for the *surprisal theory* using probabilistic parsers, showing that the surprisal (negative log-probability) of a given word predicts human processing phenomena.

More recently, decoder-only left-to-right incremental processing models such as GPT2 (alongside simpler LMs such as n-gram models and LSTMs) have become a standard testbed for the aforementioned hypothesis. Specifically, surprisal theory can be tested by measuring how well LMs' conditional output probabilities predict human behavioral data (often referred to as *psychometric predictive power*, or *ppp*; e.g., Wilcox et al., 2020; Kuribayashi et al., 2022), such as self-paced reading time data, eye-tracking data, and brain activity data. This line of literature corroborates the surprisal theory, showing that the model quality (as measured in perplexity) correlates with the predictive power of human reading time (quality-power hypothesis); in other words, the lower the perplexity, the higher the predictive power (Goodkind and Bicknell, 2018; Wilcox et al., 2020), and that this trend holds crosslinguistically (Wilcox et al., 2023).

However, exceptions have been pointed out: the trend has not been found in Japanese (Kuribayashi et al., 2021), and similarly for English, smaller

[1] Code available at `https://github.com/t-aoyam/l2lm-sentence-processing` under a Creative Commons license.

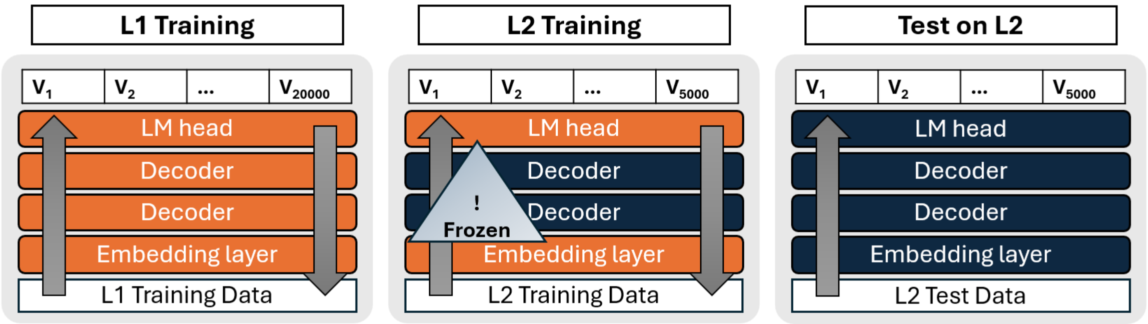

Figure 1: Training setup for L2LMs. The model is first pretrained on a given L1. We then freeze all the layers *except for* the embedding and output layers, and then continue pretraining on the L2 (English).

LMs were more predictive of human reading time than larger LMs (Oh and Schuler, 2023b). Oh and Schuler (2023a) explains this discrepancy by showing that GPT-family LMs are most predictive of human reading time after seeing around 2B–4B word tokens and their predictive power plateaus or decreases beyond that point.

Given that the surprisal theory has been empirically supported and can be used to test humanlikeness of LMs (e.g., Oh and Schuler, 2023a), and that sentence processing is susceptible to L1 transfer (e.g., Clahsen and Felser, 2006), it provides an apposite evaluation for L2LMs. In light of all these, we investigate L2LMs' predictive power of L2 English reading time, in addition to their morphosyntactic abilities.

3 Methods

To study the effect of L1 on L2LM grammaticality judgment (RQ1) and sentence processing (RQ2), we train L2LMs on various L1s and a common L2 (English), while constraining their learning during the L2 training phase, mirroring the human SLA.

3.1 Training

We adopt the TILT method from Papadimitriou and Jurafsky (2020), as well as its implementation with transformer-based LMs from Yadavalli et al. (2023). Figure 1 illustrates the approach. We first train a GPT2-based LM on a given L1 from scratch, freeze all the transformer blocks (i.e., only the embedding layers and the LM head are trainable), and then "keep pretraining" on the L2 (English).[2] The idea is that the high-level abstract linguistic knowledge has been shown to be stored in the intermediate layers (see Rogers et al., 2020 for a comprehensive review), and that freezing those parameters will force the model to acquire a new language primarily by learning new words (in the embedding layer) while relying on the L1 grammar (in the frozen decoder blocks).

All of the models were downsized for computational efficiency. Oh and Schuler (2023a) find that transformer-based LMs as small as 2 layers with 3 attention heads and an embedding size of 192 are competitive with or even better than larger models on human reading time prediction. We therefore adopt this model configuration.

3.2 L1 Training

We use the `CC100` corpus (Conneau et al. 2020; Wenzek et al. 2020), a multilingual common web crawl corpus that covers a set of 100 typologically diverse languages, available under Common Crawl term of use.[3] Since our goal is to measure the relationship between LM surprisal and human L2 reading times, the LMs' L1s were chosen based on the L1s of the human participants of the CELER corpus (Berzak et al., 2022, see §3.4.3): Arabic, Chinese (simplified), English, Japanese, Portuguese, and Spanish. We sampled a training set of 100M tokens from each of the L1 subcorpora in `CC100`.

For each L1, we first train a tokenizer. We use `SentencePieceBPETokenizer` from the `Hugging Face` library (Wolf et al., 2020), available under an Apache License. The SentencePiece algorithm (Kudo and Richardson, 2018) works well with languages not separated by white spaces, and its language-agnostic nature fits our purpose to keep the conditions as close as possible to each other across different L1s. We train each tokenizer on 5M sentences (≈500MB of data). We set the vocabulary size to 20K, based on the estimate of human L1 vocabulary size that ranges from 17K to 20K

[2] Some refer to this as a "finetuning" phase; however, because the training objective is the same, we refer to this phase as the second pretraining phase, or L2 learning phase.

[3] `https://commoncrawl.org/terms-of-use`

distinct words (Nation, 2006; Goulden et al., 1990; Zechmeister et al., 1995).

Once the tokenizer is trained, we then train the model on 4B words, saving the model at every 400M words. We focus our analyses on the first (400M words) and the last (4B words) checkpoints. We consider the 400M variant to be cognitively plausible, based on the estimate that the number of word tokens a person is annually exposed to amounts to 11M words (Hart and Risley, 1995). The 4B variant is by no means comparable to humans in terms of the amount of input; however, this variant is expected to be most predictive of human reading time based on Oh and Schuler (2023a), who find that LMs' predictive power of human reading time peaks after seeing about 2B tokens (and 4B tokens for smaller models). For both conditions, we used an effective batch size of 64 and context length of 256. This resulted in ≈24K and ≈244K training steps for each condition, respectively. The training took ≈10 hours for each L1 on a single RTX A6000 GPU with 48GB vRAM. For the subsequent L2 training, we reuse each of these L1LMs as the starting checkpoint.

3.3 L2 Training

Once the L1 training phase is complete, we freeze all of the LM's parameters except for the embedding layer and the LM head. By doing so, we allow the model to acquire a new set of vocabulary items (in the L2, namely English), as well as to adjust the classification space (we are changing both the language and the size of the classification space; $|V_{L1}| > |V_{L2}|$). For training data, we use Simple English Wikipedia,[4] available under a Creative Commons license. The preprocessed version is available under the Hugging Face library.[5]

We train the model on 30M words and save a checkpoint at every 3M words, focusing on the first (3M words) and last (30M words) checkpoints. The 3M variant is motivated by Nation (2014) and Mason and Krashen (2014), where they found that around 1M and 3M words were necessary to see the 5,000 most frequent word families and 9,000 most frequent word families, respectively, for at least 12 times. The 30M variant is to mirror the L1 training phase: we simply exposed the model to 10 times more word tokens than the cognitively plausible exposure condition (3M words), amounting to 30M word tokens. We used the same training setup, with an effective batch size and context length of 64 and 256, respectively. This resulted in 183 and 1,831 training steps for the two conditions, respectively.[6]

With 6 L1s, 1 L2, and 2 configurations for each of the two training phases, we obtain $6 \times 1 \times 2 \times 2 = 24$ L2LMs. These variants establish the baseline for the initial comparisons, although we train additional models and investigate intermediate checkpoints as follow-ups, as described in later sections. The differences in the results can be fully attributed to the inductive biases of the L2 LMs, since everything else was held constant.

3.4 Evaluation

3.4.1 Perplexity

We report the perplexity of each L2LMs, mainly for the purpose of (1) ensuring that the TILT training is properly working as expected (L2 perplexity is expected to go down) and (2) testing the quality-power hypothesis in L2 sentence processing. We obtain the perplexity on a held-out validation set of Simple English Wikipedia of 3M tokens, or 10% in size of the training set, using the sliding window strategy. Because each L2LM has a context length of 256, we set the sliding window size to 128, with each token going through the forward pass twice.

3.4.2 Morphosyntax

We also evaluate L2LMs on morphosyntactic knowledge, namely the Benchmark of Linguistic Minimal Pairs (BLiMP; Warstadt et al., 2020). As discussed earlier, this is mainly because the preceding studies (Yadavalli et al., 2023; Oba et al., 2023) used this benchmark as the main evaluation, and we aim to test whether (1) the results in the literature can be replicated with a decoder-only model (GPT2), and (2) the L2LM is properly infused with an inductive bias based on the L1 training.

3.4.3 Reading Time

LMs are considered more "humanlike" in terms of sentence processing if their per-word surprisal estimates are more predictive of the per-word human reading times. We need 3 key ingredients to test this: (i) a left-to-right incremental processing model, (ii) per-word surprisal estimates obtained from such a model, and (iii) human reading time data. For (i), we use the GPT2-based L2LM described earlier in this paper. For (ii), based on previous work (e.g., Kuribayashi et al., 2021; Oh and

[4] https://simple.wikipedia.org/wiki/Main_Page

[5] https://huggingface.co/datasets/rahular/simple-wikipedia

[6] See Appendix A for the list of hyperparameters.

Schuler, 2023b; Clark et al., 2023; Wilcox et al., 2023, *inter alia*), we take the negative log probability of a given word conditioned on all of the preceding words to obtain the model's surprisal value of a given word:

$$S_{w_i} = -\log P(w_i \mid \boldsymbol{w}_{<i}). \tag{1}$$

This represents how "surprised" the model is given the word of interest and the preceding context. The L2LM's predictive power of human reading behaviors is then measured by how much improvement we see in a linear regression model's fit for human reading time data when the surprisal is added to the baseline model. This is operationalized as the difference in the model fit between the 2 regression models (delta log-likelihood; ΔLL):

$$\Delta LL = LL_{\phi_{bl+S}} - LL_{\phi_{bl}}, \tag{2}$$

where $LL_{\phi_{bl+S}}$ and $LL_{\phi_{bl}}$ are log-lilekihood (measure of model fit) of the baseline model with and without surprisal estimates, respectively. The intuition behind this operationalization is as follows: The fit of the first regression model ($LL_{\phi_{bl}}$) represents how well human reading time can be predicted *without* an LM, whereas the fit of the second regression model ($LL_{\phi_{bl+S}}$) represents how well human reading time can be predicted *with* an LM. By taking the difference between these two, we can measure how much improvement the addition of LM surprisals makes on the fit on human reading time. This difference in the model fit, or ΔLL, is the operationalization of the LM's predictive power of human reading time. Following the previous studies, we include the word length and position as fixed effects, and subject ID as a random effect in baseline features (Φ_{bl}) to predict gaze duration. We report the per-word average ΔLL.

For (iii), we use CELER (Berzak et al., 2022), a corpus of English reading times collected from a total of 365 L1 and L2 English speakers. For the L2 speakers, L1s include Arabic, Japanese, Mandarin Chinese, Portuguese, and Spanish, as discussed earlier. We hypothesize that the surprisals obtained from L2LM trained on the same L1 as the human learners will produce the greatest ΔLL.

4 Results

4.1 Perplexity

Figure 2 summarizes the perplexities of L2LMs trained on 400M (dotted lines) and 4B (solid lines) tokens on 6 L1s, throughout the L2 training phase. The color of each bar corresponds to different L1s of L2LMs as shown in the legend (ordered in descending order based on the linguistic distance from English; Littell et al., 2017). As expected, regardless of the L1 and L1 training amount, the perplexity decreases monotonically throughout the L2 training.

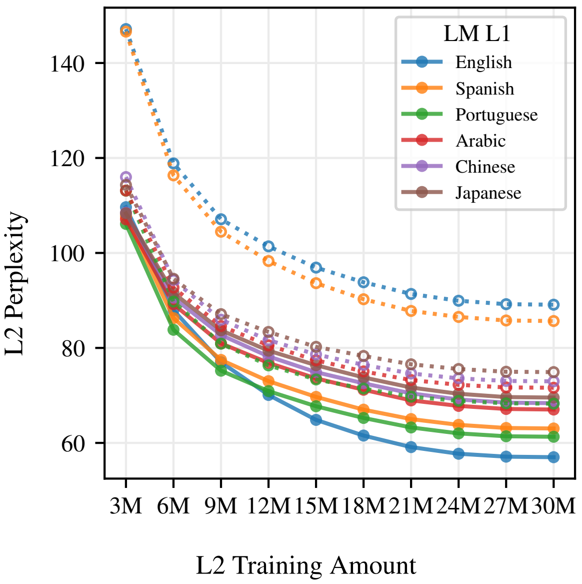

Figure 2: L2LMs' perplexities on the L2 validation set. The color indicates the L1 of the L2LM as shown in the legend. Dotted and solid lines represent the L1 training amount of 400M and 4B tokens, respectively.

More importantly, the results support Papadimitriou and Jurafsky's (2020) finding that L1s typologically closer to the L2 result in lower perplexities. The expected order was (from lower to higher in perplexity) English, Spanish, Portuguese, Arabic, Chinese, and Japanese, and the observed order was identical except for the flipped order between Spanish and Portuguese. In addition, this order was only observed when the L2LMs were sufficiently trained on the L1 (4B tokens), and not when they were trained less (400M tokens).

4.2 Morphosyntax

Figure 3 summarizes each L2LM's performance on the BLiMP dataset. Each of the 4 blocks of 6 bars corresponds to one of the 4 possible combinations of L1 and L2 training configurations. For example, 400M→3M means that all of the 6 L2LMs in that block were trained on their respective L1 for 400M tokens, and then on L2 (English) for 3M tokens. A few observations were made.

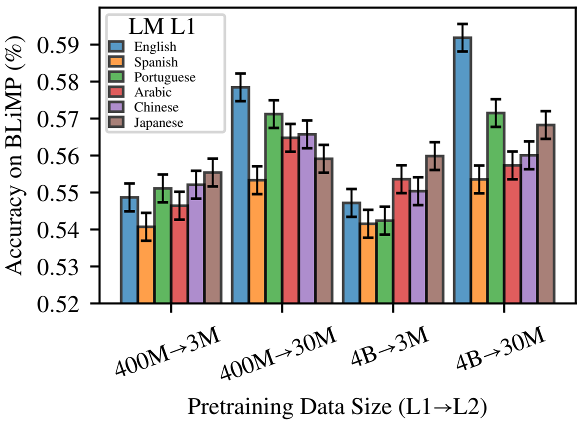

Figure 3: L2LMs' performance on BLiMP. The color indicates the L1 of the L2LM as shown in the legend, and each block indicates the respective data sizes in L1 and L2 training. Error bars represent 95% confidence intervals, obtained from binomial distributions.

	Effect of L2 Input		Effect of L1 Input	
	Δ3M→30M		Δ400M→4B	
L2LM	L1=400M	L1=4B	L2=3M	L2=30M
en→en	0.0298	0.0447	−0.0015	0.0134
es→en	0.0126	0.012	0.0008	0.0002
po→en	0.0201	0.0291	−0.0087	0.0003
ar→en	0.0184	0.0037	0.0072	−0.0075
zh→en	0.0136	0.0097	−0.0017	−0.0057
jp→en	0.0037	0.0084	0.0044	0.0091

Table 1: The difference in BLiMP accuracy scores when L2 training amount is varied (left), and when L1 training amount is varied (right).

First, with enough exposure to L2, it appears that L2LM's performance on BLiMP negatively correlates with the corresponding L1's typological distance from English, largely replicating the results from Yadavalli et al. (2023) and Oba et al. (2023). That is, models in both 400M→30M and 4B→30M (but not *→3M) configurations tend to perform better on BLiMP when their L1 is typologically closer to English.

Second, more L2 training always led to better BLiMP performance. As shown in the left half of Table 1, although the degree of improvement varied based on the L1, without exception, L2LMs' performance improves with more L2 training when the L1 training amount is held constant. That is to say, the performance improved from the 400M→3M to the 400M→30M setting, and from the 4B→3M to the 4B→30M setting.

Third, the effect of the amount of L1 training on BLiMP performance varied by L1, as shown in the right half in Table 1. This is in stark contrast with the second observation that more L2 training invariably led to higher BLiMP performance. More concretely, the L2LM's performance improves for some L1s and degrades for others, when trained on more L1 data with the L2 training amount held constant. Specifically, more L1 training on Japanese always led to better BLiMP performance, whereas more L1 training on Chinese always led to poorer BLiMP performance. More L1 training on English and Portuguese led to better BLiMP performance when L2 training was sufficient (30M) but to poorer BLiMP performance when L2 training was limited (3M), and the opposite was true about Arabic L1 training. L1 training on Spanish had virtually no effect on BLiMP performance, which may explain the anomaly we observe in Spanish L2LM in Figure 3. As we saw in §4.1, that each L1 differently affects the outcome of identical L2 training, confirming the idea that the L1 training infuses the model with different inductive biases (Papadimitriou and Jurafsky, 2020).

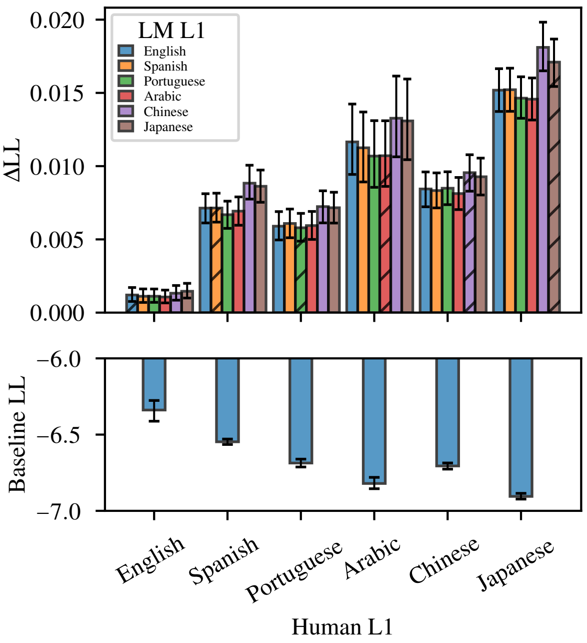

Figure 4: ΔLL of each L2LM when adding its surprisal estimates to the baseline linear regression model (top), and the corresponding log-likelihood of the baseline linear regression model for each human L1 (bottom). Each L2LM's L1 is indicated by the color of the bar, and the L1 of the human participants of the CELER corpus is indicated on the x-axis. Error bars represent 95% confidence intervals, obtained from fitting regression models on bootstrapped samples 1,000 times.

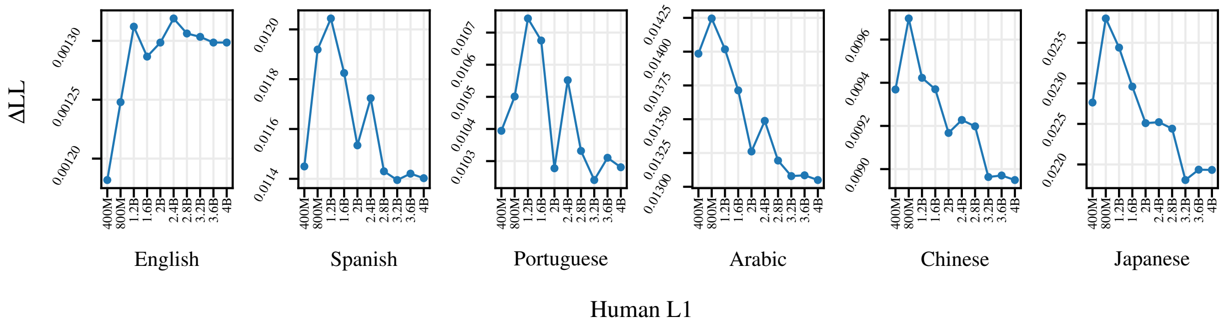

Figure 5: Reading time predictivity (ΔLL) of a monolingual English LM over the course of training. Plots reflect evaluation on data from different human L1 groups.

4.3 Reading Time

4.3.1 Effect of L1 on ΔLL

Figure 4 summarizes the per-word average ΔLL of each of the L2LM (indicated by the color of the bar) when its surprisals are added to the baseline regression model to predict the reading time of human participants of the CELER corpus (whose L1 is indicated on the x-axis). The shaded bars represent the ΔLLs that were expected be the highest within the same block (i.e. within the 6 regression models that predict the reading time of human participants of the same L1), based on the hypothesis that L2 humans and L2LMs behave similarly if they share the same L1. Figure 4 shows that this is not the case: within-block differences are small, and the predicted pattern is not consistently supported.

Rather, differences across blocks are more pronounced. Two-way analysis of variance (ANOVA) reveals that the main effect for human L1 (F = 628.64, df = 5, $p < .001$) is much stronger than that of LM L1 (F = 16.67, df = 5, $p < .001$), although both are statistically significant. In other words, regardless of the L1 of the L2LM, adding the L2LM surprisals leads to the greatest improvement in the regression model's reading time predictions when the human L1 is Japanese, followed by Arabic, Chinese, Spanish, Portuguese, and English, in descending order. We suspect that this is due to the fit of the baseline models, which rely on the word length and the word position alone. In the bottom of Figure 4, we plot the LL of the baseline regression model for each human L1. Indeed, there is a strong correlation between the mean ΔLL of 6 L2LMs for a given human L1, and the baseline LL for the same human L1 ($r = -.92$, $p < .01$), meaning that, the higher the LL of the baseline model is (the more predictable human reading time is, based on the baseline variables i.e. word length and position), the lower the ΔLL is (the less the improvement from adding L2LM surprisals).

Additionally, it is also worth noting that the order of baseline LL roughly follows the linguistic distance between L1s and English (see §4.2). The linguistic distances are in the following order: English (trivially), Spanish, Portuguese, Arabic, Chinese, and Japanese; while the baseline LL is in the following order: English, Spanish, Chinese, Portuguese, Arabic, and Japanese. This suggests that, given an L2 English reader, **the more distant their L1 is from English, the less predictable their English reading behavior is solely based on word length and word position**. This also implies that, between L1 speakers and L2 speakers, and among L2 speakers of different L1s, different strategies are employed for online English processing, which is also reported in applied psycholinguistics (e.g. Clahsen and Felser, 2006).

4.3.2 Effect of L1 and L2 Training on ΔLL

In this section, we provide analyses of the developmental trajectory of each L2LM's reading time predictions. We first show the equivalent plot of a monolingual English LM to (1) show that our methods replicate the previously reported observations on the relationship between the predictive power and training amount when tested on L1 English reading time, and (2) determine whether a similar observation can be made about L2 reading time. Figure 5 summarizes how the predictions for human L1 English reading times change throughout the pretraining process. The LM is trained on 4B tokens from scratch as described in §3.2, and each of the 6 plots differs from each other only in the L1 of the humans whose reading times the LM is predicting. Notably, on the one hand, for English speakers' reading time, ΔLL peaks at around 2.4B tokens and plateaus for the most part after-

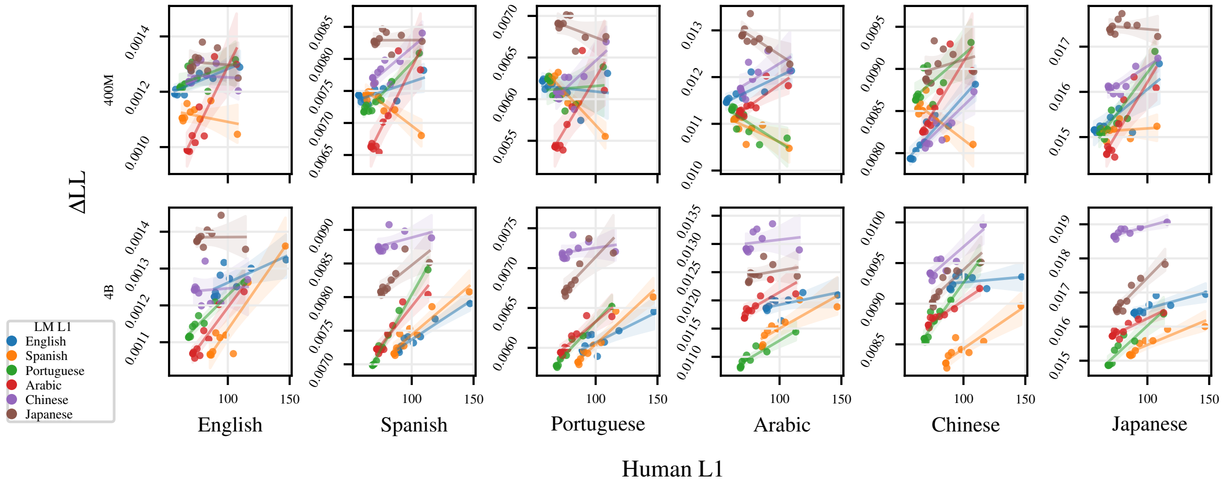

Figure 6: The relation between L2LMs' ΔLL (y-axis) and L2 perplexity (x-axis) at every 3M tokens during the L2 training phase. Each line represents an L2LM trained on the L1 of the corresponding color for 400M tokens (top) and 4B tokens (bottom), respectively. The shaded region around each line represents the 95% confidence interval. Human L1s are indicated on the x-axis.

ward. This is congruent with previously reported results that LMs' predictive power peaks after 2B–4B tokens (Oh and Schuler, 2023a), confirming (1). **On the other hand, when a monolingual English LM predicts L2 English reading time, ΔLL peaks at around 800M–1.2B and plummets beyond that point.** One potential account for this tendency is that L2 English speakers' proficiency is comparable to LMs trained on 800M–1.2B words, while LMs reach "nativelike" proficiency at around 2B–4B words.

Figure 6 plots the ΔLLs and L2 perplexities colored by LM L1 for each human L1, trained on 400M (top) and 4B (bottom) L1 tokens (see Figure 9 in Appendix C for a plot similar to Figure 5 but for L2LMs). Importantly, when L2LMs are trained on 4B L1 tokens (bottom half of the figure), regardless of LM L1 or human L1, L2 perplexity and ΔLL are positively correlated, meaning that the higher the LM quality (lower the perplexity) is, the less they are predictive of L2 human reading time. This is in contrast with the *quality-power hypothesis* (Wilcox et al., 2023), where they find a positive correlation between LM quality and LM psychometric predictive power. However, Oh and Schuler (2023a) show that 2B tokens is the tipping point where the quality-power correlation changes from positive to negative. Given that all of the L2LMs in the bottom half of the figure are trained on 4B L1 tokens, they may be at the phase where quality-power correlation is negative.

It is important to note, however, that the L2 training amount is on the order of 3M to 30M tokens, which is far from the aforementioned tipping point (2B tokens). Taken together with the observation that the choice of LM L1 had little effect on the L2LM's L2 reading time predictions (see §4.3.2), it appears that, regardless of the choice of L1, when an LM is trained on *some* L1 for a sufficient amount (say, 2B tokens), it reaches the maximum predictive power for human sentence processing, even when the target language (on which sentence processing is measured) is different from the language the LM is trained on. Considering the small model (192-2-3 architecture) and training data (≤30M tokens) sizes, there may be an alternative account for the inverse quality-power relation besides what has been proposed to date, such as larger model size allowing for the memorization of training data (Oh and Schuler, 2023b), and the learning of infrequent words later in pretraining (Oh et al., 2024). Clearly, these hypotheses have been proposed with respect to monolingual English LMs, and the observations we made on the L2LMs trained with the TILT-based (Papadimitriou and Jurafsky, 2020) unique pretraining setup may not be straightforwardly applicable.

4.3.3 Qualitative Examples

Since human data are noisy, it is not surprising that L2LMs' behaviors are not in alignment with those of human L2 speakers. Needless to say, multiple hypotheses can be given to explain these results, with the obvious one being that the L2LMs introduced in this study are simply not good models of human L2 speakers. However, in this section, we show a sample sentence from the CELER corpus where L2LMs' predictions were maximally different from each other (i.e. their surprisals were most divergent based on the L1 they were trained on). With these, we aim to show that these L2LMs do exhibit some behaviors that could potentially help us generate hypotheses about human behaviors.

In Figure 7, we see that each L2LM shows a similar level of surprisal until faced with the word *occupied*. Interestingly, L2LMs first trained on Spanish and Portuguese are more than 1.5 times more "surprised" than the other three L2LMs (first trained on Arabic, Chinese, and Japanese, respectively). To reiterate, their L2 English training and English tokenizers are identical. The word *occupied* was tokenized into *['oc', 'c', 'up', 'ied']*, suggesting that the model recognizes its part of speech (past or past participle of a verb based on the rightmost suffix). We speculate that this is because Arabic, Portuguese, and Spanish place relative clauses after noun phrases while Chinese and Japanese languages place them before. Therefore, because the decoder blocks were frozen during the L2 training phase, the model may be more likely to have a strong expectation for a determiner or a noun after a preposition for the latter group.

This speculation is based on the idea that the TILT method allows for the preservation of L1 structural information in the frozen middle layers learned during the L1 pretraining phase (Papadimitriou and Jurafsky, 2020). It has been widely observed in the BERTology literature that structural information (including POS information, which is critical to an expectation of given word class as hypothesized above) is encoded in middle layers of the transformer architecture has been widely shown (e.g., Tenney et al., 2019; Liu et al., 2019a, *inter alia*). Aoyama and Schneider (2022) also corroborate this hypothesis directly through a language modeling task, showing that the model learns to predict a word with 'correct' (i.e., same as the target) POS most actively in middle layers.

Given this literature, we suspect that our TILT-based L2LMs have L2 vocabulary and L1 structural knowledge, potentially resulting in the observed preference in certain word orders reflective of L1 structure. However, it is important to note that this hypothesis calls into question why the Arabic L2LM patterns with Chinese and Japanese (see Figure 8 in Appendix B for a similar grouping effect).

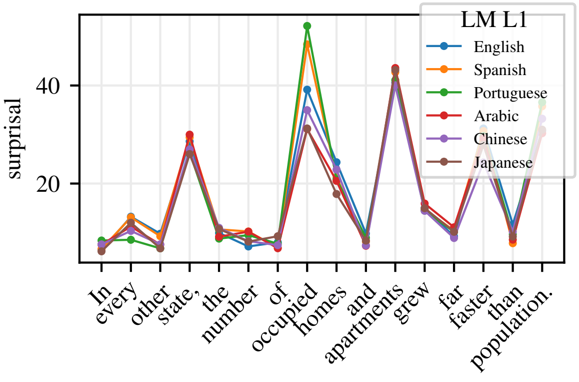

Figure 7: Per-word surprisals of a sample sentence from the CELER corpus. L2LMs maximally differ from each other at the 8th word *occupied*.

5 Conclusion

In this study of L2LMs, we trained GPT2–style decoder-only LMs on 6 L1s (Arabic, Chinese, English, Japanese, Portuguese, and Spanish) and then on a common L2 (English), freezing the decoder blocks the L2 training phase, with all of the variables including data size and hyperparameters held constant. We replicate findings from previous literature that the linguistic distance between the L1 and L2 negatively correlates with the LM's performance on the L2 (as measured in BLiMP). Our novel findings include that a monolingual English LM is most predictive of L1 English reading time at around 2.4B word tokens, and of L2 English reading time at around 800M–1.2B word tokens; that L2LMs' sentence processing was not shown to correlate with that of human L2 speakers of English; and that the overall predictability of human L2 speakers' sentence processing largely depended on their L1. We also showed that, despite the lack of overall correlation between L2LMs surprisals and L2 English speakers' reading times, qualitative examples could be conducive to generating hypotheses of L2 reading behaviors.

6 Limitations

First, this work relied solely on the TILT method (Papadimitriou and Jurafsky, 2020) to simulate the L2 learning process; of course, it is not realistic to assume that all the neurons are "frozen" in human brain once an L1 is acquired, and testing other methods of simulating human L2 (and L1) acquisition with LMs is an important avenue for future research. More broadly, simulating language learning with text inputs alone is both unrealistic and inadequate, and incorporating other input modalities, such as vision, remains an important and exciting direction.

Second, since an extensive comparison of LMs of different sizes was not feasible, we only trained L2LMs using the 192-2-3 architecture (hidden size of 192, 2 layers, and 3 attention heads), as discussed in §3.1. Although the efficacy of small LMs has been widely shown (e.g., Huebner et al., 2021) and although this particular architecture was reported as a variant that had a predictive power of human reading time similar to or even better than larger variants (Oh and Schuler, 2023a), it remains possible that testing L2LMs of larger sizes would yield new insights. In a similar vein, we only tested GPT2-based LMs in this study, and results may vary for other decoder-only models.

Third, human reading time data are scarce, let alone human L2 reading time data. The data used in this study, CELER (Berzak et al., 2022), is a rare exception; however, the total of 365 participants means that each of the 6 L1s is represented by ≈60 participants. While this is an impressive number given the cost of collecting such behavioral data, we acknowledge that findings based on these participants may not generalize to broader speaker populations.

Ethics Statement

We only studied L2 English and 6 L1s (Arabic, Chinese, English, Japanese, Portuguese, Spanish), all of which are well-resourced languages. This is because we needed to match the L1s available in the CELER corpus (Berzak et al., 2022), and not because particular languages are more "important" than others. We acknowledge that it is important to study additional first and second languages, and expanding the availability of datasets in other languages remains an important goal of future research. In addition, although certain L2LMs seemed to perform better than other L2LMs, this is not to be taken as an indication of superiority of a certain L1 population in mastering English as a second language.

Lastly, this work involved training multiple deep learning models, which is energy-intensive and could contribute to carbon emissions. However, as already described in §3, we minimize the number of models by first training 6 monolingual models while saving checkpoints and reusing them for L2LM training. Therefore, we assess the climate impacts to be modest.

Acknowledgements

We thank Ethan Wilcox and anonymous reviewers for their valuable feedback and insightful discussions. We also thank the organizers and participants of Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL 2024) for an opportunity to present this work and receive valuable feedback at an earlier stage of this work. The research was supported in part by NSF award IIS-2144881 (Nathan Schneider, PI).

References

Tatsuya Aoyama and Nathan Schneider. 2022. Probeless probing of BERT's layer-wise linguistic knowledge with masked word prediction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 195–201, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. CELER: A 365-Participant Corpus of Eye Movements in L1 and L2 English Reading. *Open Mind*, 6:41–50.

Leshem Choshen, Ryan Cotterell, Michael Y. Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. [Call for Papers] the 2nd BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2404.06214*.

Harald Clahsen and Claudia Felser. 2006. Continuity and shallow structures in language processing. *Applied Psycholinguistics*, 27(1):107–126.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A Cross-Linguistic Pressure for Uniform Information Density in Word Order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

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.

Adam Goodkind and Klinton Bicknell. 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)*, pages 10–18.

Robin Goulden, Paul Nation, and John Read. 1990. How Large Can a Receptive Vocabulary Be? *Applied Linguistics*, 11(4):341–363.

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*.

Betty Hart and Todd R Risley. 1995. *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context limitations make neural language models more human-like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 8–14.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Beniko Mason and Stephen Krashen. 2014. Can second language acquirers reach high levels of proficiency through self-selected reading? an attempt to confirm nation's (2014) results. *International Journal of Foreign Language Teaching*, 10(2):10–19.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020a. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

R Thomas McCoy, Erin Grant, Paul Smolensky, Thomas L Griffiths, and Tal Linzen. 2020b. Universal linguistic inductive biases via meta-learning. *arXiv preprint arXiv:2006.16324*.

Akira Murakami and Theodora Alexopoulou. 2016. L1 influence on the acquisition order of english grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, 38(3):365–401.

Paul Nation. 2006. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review-revue Canadienne Des Langues Vivantes - CAN MOD LANG REV*, 63:59–81.

Paul Nation. 2014. How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language*, 26:1–16.

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, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5.

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.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023b. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Alex Warstadt and Samuel R. Bowman. 2024. What artificial neural networks can tell us about human language acquisition. *arXiv preprint arXiv:2208.07998*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511, Singapore. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, page 1707–1713.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. 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.

Eugene B. Zechmeister, Andrea M. Chronis, William L. Cull, Catherine A. D'Anna, and Noreen A. Healy. 1995. Growth of a functionally important lexicon. *Journal of Reading Behavior*, 27(2):201–212.

A Hyperparameters

	L1	L2
vocab_size	20_000	5_000
context_length	256	256
train_layers	all	[wte.weight, wpe.weight, ln_f.weight, ln_f.bias]
n_embed	192	192
n_layer	2	2
n_head	3	3
batch_size	8	1
grad_acc_steps	8	8
weight_decay	0.1	0.1
warmup_steps	1_000	30
lr_scheduler	cosine	cosine
learning_rate	5e-4	5e-4

Table 2: List of hyperparameters used to train L2LMs.

Table 2 summarizes the hyperparameters used to train L2LMs. The keys are shortened in the table for readability and for space reasons. Please refer to the `configs` directory in our Github repo to see the exact setup.

B More Qualitative Examples

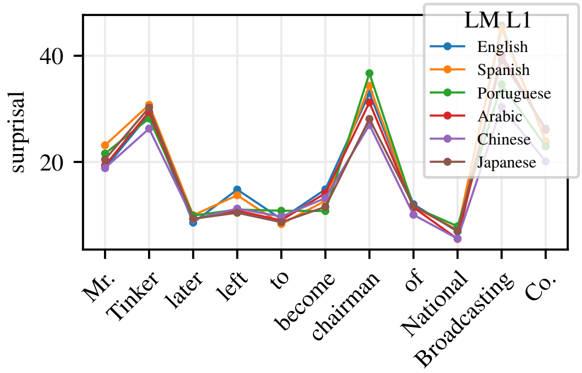

Figure 8: Per-word surprisals of a sample sentence from the CELER corpus. L2LMs maximally differ from each other at the 7th word *chairman*.

Figure 8 plots the per-word suprisal obtained from each of the 6 L2LMs. Here again, Chinese and Japanese L2LMs pattern together, showing lower surprisals compared to other L2LMs. This is perhaps due to the fact that Chinese and Japanese are the only 2 languages among the 6 that do not have the article system, and that an article-less singular noun is less surprising.

C Training Steps and L2LM ΔLL

Figure 9 summarizes each L2LM's predictive power of each L1 group's L2 English reading time at every 3M tokens they saw in the L2 training phase. Because L2 perplexity was monotonically decreasing in relation to L2 training amount (see Figure 2), Figure 6 looks similar in shape to Figure 9, which summarized the trajectory of ΔLL as a function of L2 perplexity. We find a few general patterns.

First, comparing the L2LMs trained on 400M words in the L1 (top) and those trained on 4B words in the L1 (top), even though the L2 input and size are identical, the development of the predictive power seems to be different. When L2LMs are exposed to 4B words during the L1 training phase, with very few exceptions, L2 exposure does not improve L2LMs' predictive power, as observed in the almost monotonically decreasing trend in all 6 plots in the bottom half of Figure 9.

Second, when L2LMs are exposed to 400M words during the L1 training, L2 exposure sometimes improves the predictive power, for some combinations of L2LMs' L1s and human L1s (e.g., the Spanish L2LM predicting L2 English reading time of Spanish, Portuguese, and Chinese speakers).

Lastly, it is worth noting that this overall downward trend in the development of predictive power stands in sharp contrast to the monotonically upward trend in BLiMP performance (left half of Table 1; see §4.2), where L2 exposure *always* led to higher BLiMP performance, regardless of the number of words L2LMs have seen during the L1 training phase (400M and 4B). This suggests that, with the decoder blocks frozen and only embedding and output layers left trainable, L2LMs adapted to the English inputs during the L2 training phase by learning to distinguish likely and unlikely (grammatical and ungrammatical) sequences of inputs, using novel strategies that seem to deviate from both humans and regular LMs.

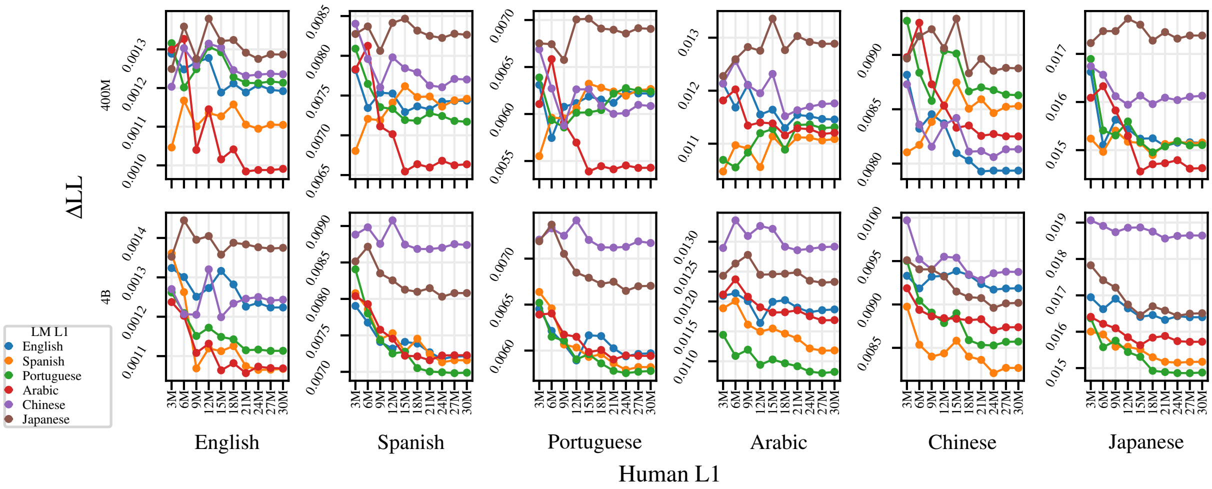

Figure 9: L2LMs' ΔLL on CELER corpus at every 3M tokens during the L2 training phase. Each line represents an L2LM trained on the L1 of the corresponding color for 400M tokens (top) and 4B tokens (bottom), respectively. Human L1s are indicated on the x-axis.