Predicting Human Translation Difficulty with Neural Machine Translation

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and **Trevor Cohn***
The University of Melbourne, Australia
z.lim4@student.unimelb.edu.au
{vylomovae,c.kemp,t.cohn}@unimelb.edu.au

Abstract

Human translators linger on some words and phrases more than others, and predicting this variation is a step towards explaining the underlying cognitive processes. Using data from the CRITT Translation Process Research Database, we evaluate the extent to which surprisal and attentional features derived from a Neural Machine Translation (NMT) model account for reading and production times of human translators. We find that surprisal and attention are complementary predictors of translation difficulty, and that surprisal derived from a NMT model is the single most successful predictor of production duration. Our analyses draw on data from hundreds of translators operating across 13 language pairs, and represent the most comprehensive investigation of human translation difficulty to date.

1 Introduction

During the Nuremberg trials, a Soviet interpreter paused and lost her train of thought when faced with the challenge of translating the phrase ‘‘Trojan Horse politics’’, and the presiding judge had to stop the session (Matasov, 2017). Translation difficulty rarely has such extreme consequences, but the process of translating any text requires a human translator to handle words and phrases that vary in difficulty. Translation difficulty can be operationalized in various ways (Sun, 2015), and one approach considers texts to be difficult if they trigger translation errors (Vanroy et al., 2019). Here, however, we focus on difficulty in cognitive processing, and consider a word or phrase to be difficult if it requires extended processing time. Our opening example (i.e., ‘‘Trojan Horse politics’’) illustrates translation difficulty at the phrase level, and Figure 1 shows how our notion of translation difficulty varies at the level of individual words. Across this sample, words like ‘‘societies’’ and ‘‘population’’ are consistently linked with longer production times than words like ‘‘result’’ and ‘‘tend’’.

Processing times have been extensively studied by psycholinguists, but the majority of this work is carried out in a monolingual setting. Within the literature on translation, analysis of cognitive processing is most prominent within a small but growing area known as Translation Process Research (TPR). Researchers in this area aim to characterize the cognitive processes and strategies that support human translation, and do so by analyzing eye movements and keystrokes collected from translators (Carl, 2012). Here we build on this tradition and focus on three variables at the word and segment level derived from the CRITT TPR-DB database (Carl et al., 2016b): source reading time (`TrtS`), target reading time (`TrtT`), and translation duration (`Dur`). Our analyses are relatively large in scale by the standards of previous work in this area, and we draw on data from 312 translators working across 13 language pairs.

A central goal of our work is to bring translation process research into contact with modern work on Neural Machine Translation (NMT). Recent work in NLP has led to dramatic improvements in the multilingual abilities of NMT models (Kudugunta et al., 2019; Aharoni et al., 2019), and these models can support tests of existing psycholinguistic theories and inspire new theories. Our work demonstrates the promise of NMT models for TPR research by testing whether surprisal and attentional features derived from an NMT model are predictive of human translation difficulty. Two of these predictors are shown in the right panels of Figure 1, and both are correlated with translation duration for the example sentence shown.

*Now at Google.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1479–1496, 2024. https://doi.org/10.1162/tacl_a_00714
Action Editors: Liang Huang and Stefan Riezler. Submission batch: 9/2023; Revision batch: 5/2024; Published 11/2024.

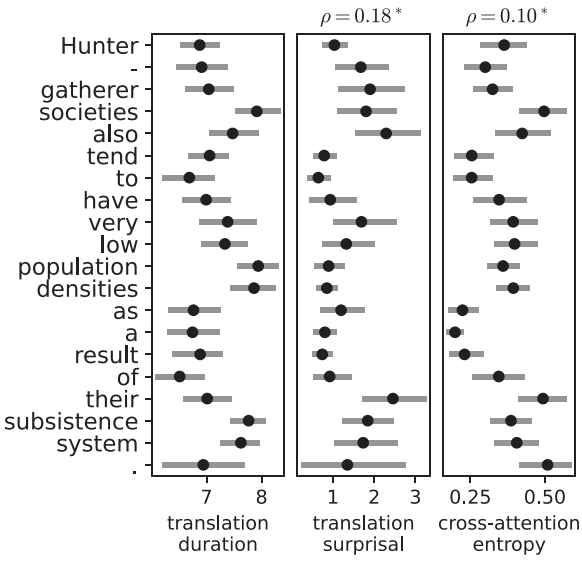


Figure 1: Translation surprisal and cross-attention entropy are predictive of average word translation duration, which is the time taken (in ms, log scale) to produce an aligned segment on the target side, divided by the number of source words aligned with this segment. Error bars show 95% confidence intervals over 47 translations in five target languages: Chinese, Japanese, Hindi, German, and Spanish. '*' marks significant Pearson correlations ($p < .05$) of predictors with translation duration (leftmost panel).

In what follows we introduce the surprisal and attentional features (Sections 3 and 4) that we consider then evaluate the extent to which they yield improvements over baseline models of translation difficulty. We find that surprisal is a strong predictor of difficulty, which supports and extends previous psycholinguistic findings that surprisal predicts both monolingual processing (Levy, 2008; Wilcox et al., 2023) and translation processing (Teich et al., 2020; Wei, 2022; Carl, 2021a). The attentional features we evaluate predict difficulty less well, but provide supplementary predictive power when combined with surprisal.

2 Related Work

An extensive body of monolingual work has demonstrated that the pace of human reading is based on next-word predictability—more contextually surprising words incur higher cognitive costs and are slower to process (Hale, 2001; Levy, 2008). The phenomenon is observed across a wide range of language usage, including reading (Monsalve et al., 2012; Smith and Levy, 2013; Meister et al., 2021; Shain et al., 2024; Wilcox et al., 2021, 2023), listening and comprehension (Russo et al., 2020, 2022; Kumar et al., 2023; Yu et al., 2023), speech (Jurafsky, 2003; Levy and Jaeger, 2006; Cohen-Priva and Jurafsky, 2008; Demberg et al., 2012; Malisz et al., 2018; Dammalapati et al., 2019; Pimentel et al., 2021), typing (Chen et al., 2021), and code-switching (Calvillo et al., 2020).

Surprisal has also been proposed as a predictor of translation difficulty in Translation Processing Research (TPR) (Schaeffer and Carl, 2014; Teich et al., 2020; Carl, 2021a; Wei, 2022; Deilen et al., 2023; Lim et al., 2023), but existing evaluations of this idea are limited in two respects. First, previous measures of translation surprisal are based on overly simplistic probability estimates, which are inferior to modern NMT models. Prior work on TPR mostly relies on probabilities extracted from manual word alignments, and these probabilities are noisy because they are not sensitive to target context and because the corpora are relatively small. Second, relying on manual alignment means that most previous approaches do not scale well and cannot be applied to novel unaligned texts. Lim et al. (2023) address this second limitation using automatic alignment of existing parallel corpora, but their probability estimates are again insensitive to context and therefore less accurate than probabilities derived from NMT models.

Along with surprisal, prior TPR research has also proposed contextual entropy as a predictor of translation difficulty (Teich et al., 2020; Carl, 2021a; Wei, 2022). Entropy is the expected surprisal of a translation distribution, and indicates the effort in resolving the uncertainty over all possible translation choices. The results of Wei (2022) and Carl (2021a) suggest, however, that entropy is weaker than surprisal as a predictor of translation duration. Entropy is also a relatively weak predictor of monolingual reading difficulty, despite being hypothesized to indicate a reader's anticipation of processing effort (Hale, 2003; Linzen and Jaeger, 2016; Lowder et al., 2018; Wilcox et al., 2023).

In the broader NLP literature, several lines of work suggest that attentional measures derived from language models and NMT models can capture processing difficulty. Attention weights in language models (Li et al., 2024) and NMT models (Ferrando and Costa-jussà, 2021) are known to contribute to next-token prediction, and previous

studies have explored whether attentional weights in language models can account for monolingual processing time (Ryu and Lewis, 2021; Oh and Schuler, 2022). For translation, NMT studies suggest that the contextualization of ambiguous tokens occurs in encoder self-attention (Tang et al., 2019; Yin et al., 2021), and NMT models show higher cross-attention entropy with increasingly difficult tokens (Dabre and Fujita, 2019; Zhang and Feng, 2021; Lu et al., 2021). NMT attentional weights also reflect whether a figurative expression is handled by paraphrasing as opposed to literal translation (Dankers et al., 2022). All of these results suggest that attentional weights are worth exploring as predictors of human translation difficulty.

In the TPR literature, prior work has explored whether errors made by MT models tend to predict translation difficulty and post-editing difficulty for humans (Carl and Báez, 2019). To our knowledge, however, previous work has not used probabilities and attentional measures derived from NMT models as predictors of human translation difficulty. Building on previous work, we focus on surprisal rather than entropy, and provide a comprehensive evaluation of the extent to which surprisal predicts translation difficulty. We also propose several attentional features based on previous literature and test whether they contribute predictive power that goes beyond surprisal alone.[1]

3 Surprisal: An Information-theoretic Account of Translation Processing

In contrast with traditional monolingual surprisal, translation surprisal is based on a context that includes a complete sequence in a source language (SL) and a sequence of previously translated words in a target language (TL). Target words with high translation surprisal are hypothesized to require extended cognitive processing because they are relatively unpredictable in context.

Relative to prior work, we aim to provide a more comprehensive and rigorous evaluation of the role of surprisal in translation processing. We consider predictors of translation difficulty based on both monolingual and translation surprisal, which we estimate using a large language model and a neural translation model respectively.

[1]Code available at https://github.com/ZhengWeiLim/pred-trans-difficulty-NMT.

3.1 Monolingual Surprisal

The monolingual surprisal of a word or segment can be estimated from an autoregressive language model (LM). Let $\boldsymbol{w} = [w_1, \ldots, w_m]$ be a complete sequence of tokens (e.g., a full sentence), and let $\boldsymbol{w_i}$ denote a segment of $\boldsymbol{w}$ where $\boldsymbol{i} \subseteq \{1, \ldots, m\}$ denotes the token indices. The surprisal of $\boldsymbol{w_i}$ is the sum of negative log-probabilities of each token w_i conditioned on the preceding context $\boldsymbol{w}_{<i}$:

$$s_{\text{lm}}(\boldsymbol{w_i}) = \sum_{i \in \boldsymbol{i}} -\log p_{\text{lm}}(w_i | \boldsymbol{w}_{<i}), \qquad (1)$$

where p_{lm} denotes the LM distribution. Note that this definition of surprisal applies to both words and segments, which consist of a sequence of tokens defined under a model tokenization scheme.[2]

Given the well-established link between s_{lm} and reading time (Monsalve et al., 2012; Smith and Levy, 2013; Shain et al., 2024; Wilcox et al., 2021), we test whether s_{lm} predicts reading times for source and target texts when participants are engaged in the act of translation. Previous work also establishes links between surprisal and language production in tasks involving speech (Dammalapati et al., 2021), oral reading (Klebanov et al., 2023), code-switching (Calvillo et al., 2020), and typing (Chen et al., 2021). Building on this literature, we also test whether monolingual surprisal predicts the time participants take to type their translations.

3.2 Translation Surprisal

The surprisal of a translation can be obtained from the distribution of a neural machine translation model, p_{mt}, conditioned on a source sequence and previously translated tokens. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be a pair of parallel sequences, and let $\boldsymbol{y_j}$ be a segment of $\boldsymbol{y}$ with token indices $\boldsymbol{j}$ and $\boldsymbol{y}_{<j}$ be the preceding context of each token y_j. The translation surprisal of segment $\boldsymbol{y_j}$ is defined as

$$s_{\text{mt}}(\boldsymbol{y_j}) = \sum_{j \in \boldsymbol{j}} -\log p_{\text{mt}}(y_j | \boldsymbol{x}, \boldsymbol{y}_{<j}). \qquad (2)$$

We will compare translation surprisal and monolingual surprisal as predictors of target-side

[2]$\boldsymbol{w_i}$ is also defined in a way that allows difficulty prediction of non-contiguous segments.

reading time and production duration. We expect translation surprisal to be the more successful predictor because this measure incorporates context on both the source and target sides, and because NMT models are trained specifically for translation. Translation surprisal, however, cannot be used to predict source-side reading time because the model must encode the entire source sequence before providing an output distribution.

4 Predicting Translation Difficulty using NMT Attention

Modern encoder-decoder NMT models rely on the transformer architecture and incorporate three kinds of attention: encoder self-attention, cross-attention, and decoder self-attention (Vaswani et al., 2017).[3] We consider all three sets of attention weights as potential predictors of translation difficulty. By some accounts, reading, transferring, and writing are three distinct stages in the human translation process (Shreve et al., 1993; Macizo and Bajo, 2004, 2006; Shreve and Lacruz, 2017), and we propose that the three attentional modules roughly align with these three stages.

Let $\boldsymbol{x} = [x_1, \ldots, x_m]$ and $\boldsymbol{y} = [y_1, \ldots, y_n]$ denote parallel source and target sequences, with $\boldsymbol{m} = \{1, \ldots, m\}$ and $\boldsymbol{n} = \{1, \ldots, n\}$ as their token indices. Let $\boldsymbol{u}$ and $\boldsymbol{v}$ denote segments of $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively such that $\boldsymbol{u} = \boldsymbol{x_i}$ and $\boldsymbol{v} = \boldsymbol{y_j}$, where the indices $\boldsymbol{i} \subseteq \boldsymbol{m}$ and $\boldsymbol{j} \subseteq \boldsymbol{n}$. Note that $\boldsymbol{x}$ and $\boldsymbol{y}$ do not include special tokens (e.g., the end-of-sequence tag `eos`) added to the sequences under the NMT's tokenization scheme. We additionally define $\overline{\boldsymbol{u}} = \boldsymbol{x}_{\bar{\boldsymbol{i}}}$ and $\overline{\boldsymbol{v}} = \boldsymbol{y}_{\bar{\boldsymbol{j}}}$ as the contexts where the corresponding segments are excluded, i.e., $\bar{\boldsymbol{i}} = \boldsymbol{m} \setminus \boldsymbol{i}$ and $\bar{\boldsymbol{j}} = \{1, \ldots, \max(\boldsymbol{j})\} \setminus \boldsymbol{j}$. In contrast to $\bar{\boldsymbol{i}}$, $\bar{\boldsymbol{j}}$ only includes token indices from the context preceding $\boldsymbol{v}$. We define $\bar{\boldsymbol{j}}$ in this way because the decoder typically does not have access to future tokens when generating translations.

We consider two kinds of attentional features. The first captures the total attentional flow from segment $\boldsymbol{w}$ to segment $\boldsymbol{z}$:

$$\text{flow}(A, \boldsymbol{w}, \boldsymbol{z}) = \sum_{l \in \boldsymbol{l}} \sum_{k \in \boldsymbol{k}} a_{kl}, \tag{3}$$

where A is an attention matrix, $\boldsymbol{k}$ and $\boldsymbol{l}$ are token indices of $\boldsymbol{w}$ and $\boldsymbol{z}$, and a_{kl} is the attention from the token at index k to the token at index l.

The second kind of attentional feature sums the entropies of all attentional distributions from segment $\boldsymbol{w}$ to segment $\boldsymbol{z}$:

$$H(A, \boldsymbol{w}, \boldsymbol{z}) = \sum_{k \in \boldsymbol{k}} \sum_{l \in \boldsymbol{l}} -\hat{a}_{kl} \log \hat{a}_{kl} \tag{4}$$

$$\hat{a}_{kl} = \frac{a_{kl}}{\|\boldsymbol{a}_{k \to \boldsymbol{l}}\|},$$

where again $\boldsymbol{k}$ and $\boldsymbol{l}$ are the indices of $\boldsymbol{w}$ and $\boldsymbol{z}$. We renormalize a_{kl} along $\boldsymbol{a}_{k \to \boldsymbol{l}}$, the vector including attention weights from k to token indices $\boldsymbol{l}$.[4] This step allows us to filter out attention to special tokens (e.g., `eos`) when considering the entropy of a distribution.

Using Equations 3 and 4, we define in the following six attentional features as candidate predictors of source-side reading time and five features for target-side reading time and production duration. For simplicity, we obtain the final attentional features in our analysis by averaging values computed for each attentional head across layers.

4.1 Predicting Source Text Difficulty

Encoder Attention. We propose four features extracted from encoder attention that are inspired by Dankers et al. (2022), who find that when producing a paraphrase rather than a literal translation of a figurative expression, an NMT encoder tends to allocate more attention from the phrase to itself, while reducing attention directed to and received from the context. This result is relevant to our goal of predicting translation difficulty because paraphrasing is known to be more effortful than literal translation (Balling et al., 2014; Schaeffer and Carl, 2014; Rojo, 2015; Carl and Schaeffer, 2017).

Given an encoder self-attention matrix A^e, we use Equation 3 to define features that capture the total attentional flow from $\boldsymbol{u}$ to $\boldsymbol{u}$, to its context $\overline{\boldsymbol{u}}$, and to the `eos` tag:

$$f^{\text{e}}_{u,\overline{u}} = \text{flow}(A^{\text{e}}, \boldsymbol{u}, \overline{\boldsymbol{u}}) \tag{5}$$

$$f^{\text{e}}_{u,\texttt{eos}} = \text{flow}(A^{\text{e}}, \boldsymbol{u}, \texttt{eos}). \tag{6}$$

[3] Cross-attention is sometimes known as encoder-decoder attention.

[4] ℓ_1 normalization.

In line with Dankers et al. (2022), we hypothesize that harder-to-translate segments direct more attention to themselves and less to their contexts. To examine if the model also reduces attention flow from context to $\boldsymbol{u}$, we further define $f^{\mathrm{e}}_{\bar{u},u}$:

$$f^{\mathrm{e}}_{\bar{u},u} = \mathrm{flow}(A^{\mathrm{e}}, \bar{\boldsymbol{u}}, \boldsymbol{u}). \tag{7}$$

The NMT encoder relies on attention to relevant context to disambiguate input meanings and to resolve anaphoric pronouns (Tang et al., 2019; Yin et al., 2021). These words may take longer to read, as more time is required to determine pronoun referents and word senses. Ambiguous words also contribute to low attentional entropy (Tang et al., 2019). To characterize this feature of $\boldsymbol{u}$, we compute $H^{\mathrm{e}}_{u,x}$, the overall attentional entropy based on Equation 4:

$$H^{\mathrm{e}}_{u,x} = H(A^{\mathrm{e}}, \boldsymbol{u}, \boldsymbol{x}). \tag{8}$$

We hypothesize that low $H^{\mathrm{e}}_{u,x}$ predicts longer reading time of $\boldsymbol{u}$.

Cross-attention. Cross-attention allows information to pass from encoder to decoder and establishes rough alignments between input and output tokens in an NMT model (Alkhouli et al., 2018; Li et al., 2019). However, it is unclear from previous work whether more attention weight received from the target sequence contributes to harder or easier translation. Tu et al. (2016) and Mi et al. (2016) show that increased attention received by part of the source text is related to *over-translation*, a phenomenon in which the model focuses too much on some parts of the input and neglects others when generating a translation. In contrast, Dankers et al. (2022) demonstrate that paraphrasing a figurative expression instead of literal translation reduces attention to corresponding source tokens.

On one hand, source tokens that receive more attention and stronger alignments are deemed more important by the model; on the other, the same phenomenon corresponds to literal translation, which is easier than paraphrasing. Nevertheless, both studies show that cross-attention flow to $\boldsymbol{u}$ is pertinent to our goal of characterizing translation difficulty. Given a cross-attention matrix A^{c}, we define the attentional flow from $\boldsymbol{y}$ to $\boldsymbol{u}$ as:

$$f^{\mathrm{c}}_{y,u} = \mathrm{flow}(A^{\mathrm{c}}, \boldsymbol{y}, \boldsymbol{u}), \tag{9}$$

and test $f^{\mathrm{c}}_{y,u}$ as a predictor of source reading time.

4.2 Predicting Target Text Difficulty

Cross-attention. When paraphrasing non-literal nouns, NMT also tends to shift cross-attention weight away from these nouns, focusing instead on surrounding tokens and source `eos` (Dankers et al., 2022). We therefore hypothesize that the translation difficulty associated with target segment $\boldsymbol{v}$ may be predicted by the total attention directed from $\boldsymbol{v}$ to source `eos`:

$$f^{\mathrm{c}}_{v,\texttt{eos}} = \mathrm{flow}(A^{\mathrm{c}}, \boldsymbol{v}, \texttt{eos}). \tag{10}$$

Prior work attributes cross-attention uncertainty to lack of confidence in translation outputs and less informative source inputs (Dabre and Fujita, 2019; Zhang and Feng, 2021; Lu et al., 2021). Following these studies, we hypothesize that higher uncertainty in cross-attention indicates less confident alignment with source tokens and therefore predicts increased translation difficulty. We measure alignment uncertainty between $\boldsymbol{v}$ and the source sequence $\boldsymbol{x}$ as:

$$H^{\mathrm{c}}_{v,x} = H(A^{\mathrm{c}}, \boldsymbol{v}, \boldsymbol{x}). \tag{11}$$

Decoder Attention. While previous work on NMT models has considered encoder and cross-attention in depth, decoder self-attention has received less investigation. Yang et al. (2020) demonstrate that the role of decoder self-attention is to ensure translation fluency. Relative to encoder attention and cross attention, decoder attention aligns less well with human annotations, and contributes less to improving NMT performance when regularized with human annotations (Yin et al., 2021).

For completeness, however, we consider three decoder attentional features that parallel those introduced in Section 4.1 for encoder self-attention. Despite their similarity, these attentional features are evaluated against different behavioral

	Study	Token			Segment		
		TrtS	TrtT	Dur	TrtS	TrtT	Dur
en → da	ACS08 (Sjørup, 2013), BD13, BD08 (Dragsted, 2010)	5305	5320	6176	4121	4203	4779
en → de	SG12 (Nitzke, 2019)	3691	3956	4534	2991	3243	3589
en → es	BML12 (Mesa-Lao, 2014)	4067	4020	8280	3555	3330	6072
en → hi	NJ12 (Carl et al., 2016b)	4717	4828	5205	2917	2851	2933
en → ja	ENJA15 (Carl et al., 2016a)	6806	8329	2168	4263	4299	2130
en → nl	ENDU20 (Vanroy, 2021)	0	0	7814	0	0	6318
en → pt	JLG10 (Alves and Gonçalves, 2013)	0	0	2443	0	0	2217
en → zh	RUC17, STC17 (Carl and Báez, 2019), CREATIVE (Vieira et al., 2023)	8949	7934	3876	6097	5922	3925
da → en	LWB09 (Jensen et al., 2009)	3844	4177	5327	3445	3493	4315
fr → pl	DG01 (Płońska, 2016)	0	0	17041	0	0	13283
pt → en	JLG10	0	0	2053	0	0	1876
pt → zh	MS13 (Schmaltz et al., 2016)	1011	830	203	781	755	237
zh → pt	MS13	1210	1237	1509	1101	1027	1209

Table 1: Data drawn from studies in CRITT-TPRDB with the number of valid samples ($\geq$ 20ms) per segmentation level.

measures. The encoder features are treated as candidate predictors of source reading time, and the decoder features are used to predict target reading and production time. If A^{d} is the decoder attention matrix, the following features capture attention flow from $\boldsymbol{v}$ to itself and to the preceding context, as well as self-attention entropy:

$$f^{\mathrm{d}}_{v,v} = \mathrm{flow}(A^{\mathrm{d}}, \boldsymbol{v}, \boldsymbol{v}) \quad (12)$$

$$f^{\mathrm{d}}_{v,\bar{v}} = \mathrm{flow}(A^{\mathrm{d}}, \boldsymbol{v}, \bar{\boldsymbol{v}}) \quad (13)$$

$$H^{\mathrm{d}}_{v,\overleftarrow{v}} = H(A^{\mathrm{d}}, \boldsymbol{v}, \overleftarrow{\boldsymbol{v}}). \quad (14)$$

As mentioned earlier, a decoder does not attend to future tokens, including target-side eos. Decoder features analogous to $f^{\mathrm{e}}_{u,\mathtt{eos}}$ and $f^{\mathrm{e}}_{\bar{u},u}$ are not possible for this reason. Attentional entropy $H^{\mathrm{d}}_{v,\overleftarrow{v}}$ is computed over $\overleftarrow{\boldsymbol{v}}$, which includes all target-side tokens up to the rightmost token in $\boldsymbol{v}$.

5 Data

Our empirical measures of translation difficulty are derived from the CRITT Translation Process Research Database (TPR-DB) (Carl et al., 2016b). We focus on three behavioral measures: source text reading time (TrtS) is the sum of all fixation durations on a given source segment during a session; target text reading time (TrtT) is the sum of all fixation durations on a target segment; and translation production duration (Dur) is the time taken to produce a segment. Both reading time measures are based on eye-tracking data, and the translation production measure is based on keylogging data. The data set is organized in terms of words, segments, and sentences, and we carry out separate analyses at the word and segment levels.

In CRITT TPR-DB, word and segment boundaries and alignments are provided by human annotators. For consistency, we remove alignments of words and segments that cross sentence boundaries. Following Carl (2021b), we filter values of TrtS, TrtT, and Dur lower than 20ms. The remaining values are log scaled. We analyze data from 17 public studies available from the public database.[5] These studies represent 13 different language pairs, and each study includes data from an average of 18 human translators. The studies included along with the size of each one are summarized in Table 1. For cross-validation, we divide the samples into 10 folds. To ensure that all predictions are evaluated using previously unseen sentences, we randomly sample test data at the sentence level, which means that the source sentences in train and test partitions do not overlap.

[5]The translation studies selected from the TPR database exclude data sets where many alignments cross sentence boundaries, or that contain too many errors (e.g., missing values and inconsistent sentence segmentations) across tables.

6 Models and Methods

LM and NMT Models. We follow Wilcox et al. (2023) and use mGPT (Shliazhko et al., 2024), a multilingual language model, to estimate monolingual surprisal (s_{lm}).[6] To compute translation surprisal (s_{mt}), we use NLLB-200's 600M variant, a multi-way multilingual translation model that is distilled from a much larger 54.5B Mixture-of-Experts model (Costa-jussà et al., 2022).[7] Among publicly available NMT models, NLLB-200 is a standard benchmark and achieves state-of-the-art results across many language pairs (Moslem et al., 2023; Sealess Communication et al., 2023). We compute attentional features for each of 16 heads across 12 layers, then average across heads and layers to create the final set of attentional features for our analyses.

Normalization. Sections 3 and 4 describe feature definitions that are sums over a sequence of tokens, which makes it crucial to control for segment length when predicting reading time and production duration. All surprisal values are therefore normalized by the lengths of the input segments, $\boldsymbol{w_i}$ and $\boldsymbol{y_j}$. To normalize attentional features, we first calculate dummy feature values by replacing a_{lk} in Equation 3 and 4 with uniform attention values (i.e., $a_{lk} = 1/|k|$ where $|k|$ is the length of the attention vector). A normalized attentional feature is defined as the ratio of the raw feature value (defined in Section 4) to its dummy value.

Control Features. Although we are most interested in surprisal and attentional features as predictors of translation difficulty, other simple features may also predict difficulty. In particular, longer segments, low-frequency segments, and segments towards the beginning of a sentence might be systematically more difficult than shorter segments, high-frequency segments, and segments towards the end of a sentence. We therefore include segment length, average unigram frequency[8] (log scaled) and average position quantile as control features in all models, where both averages are computed over all tokens belonging to a segment.

Linear Models. Following previous studies, we use linear models to evaluate the predictive power of both surprisal and attentional features. To predict translation difficulty for all languages, we use a mixed model that includes language pair and participant id as random effects. In addition, we fit individual linear regression models for the four language pairs (en → da, en → de, en → hi, and en → zh) for which we have most data.

[6] mGPT checkpoint.

[7] NLLB-200 checkpoint.

[8] `https://pypi.org/project/wordfreq/`.

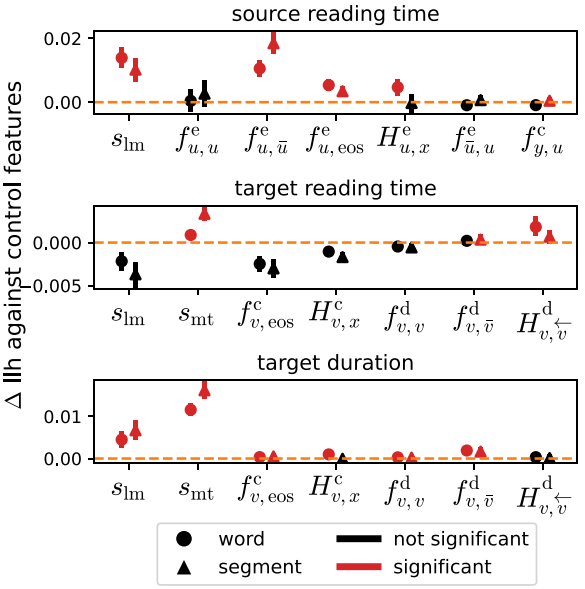


Figure 2: Δllh plotted against control features. 95% confidence intervals are estimated based on the 10 cross-validation folds.

7 Results

Δllh as a Measure of Predictive Power. Prior work on translation difficulty rarely uses held-out evaluation, but we follow previous psycholinguistic studies (Goodkind and Bicknell, 2018; Kuribayashi et al., 2021; Wilcox et al., 2020, 2023; De Varda and Marelli, 2023) and evaluate our models using log-likelihood of held-out data. To assess the predictive power of a feature, we train a mixed model with the feature of interest in addition to all control features, and compare against a baseline model which includes only the control features. The contribution of the predictor feature is then measured as the difference in log-likelihood of the held-out test data (Δllh) between the two models. A positive Δllh indicates added predictive power from the feature relative to the baseline model, whereas $\Delta\text{llh} \leq 0$ means that we have no evidence for the effect of the feature on reading and production times.[9] Like Wilcox et al.

[9] Wilcox et al. (2023) point out that Δllh may be ≤ 0 because of overfitting, or because the relationship between

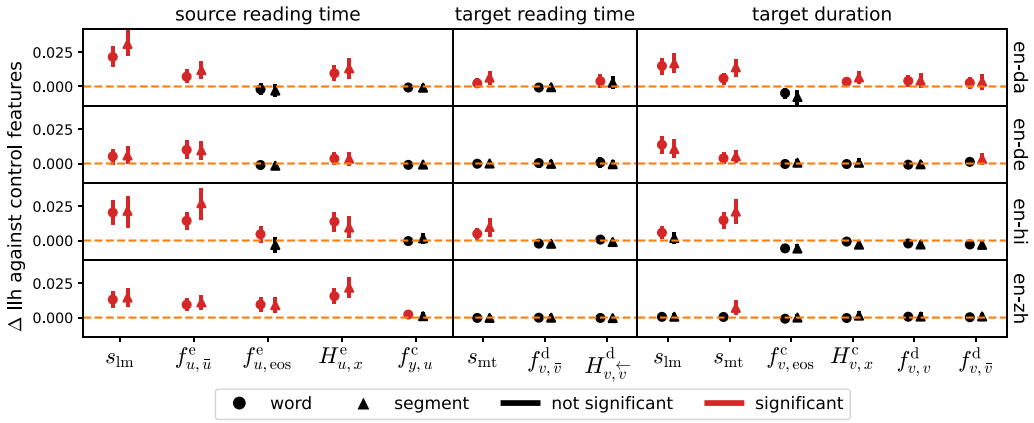


Figure 3: Δllh by language pairs with features that are significant for at least one of the word and segment levels in Figure 2.

(2023), we test if $\Delta\text{llh} > 0$ is significant across held-out samples using a paired permutation test based on 1000 random permutations.

7.1 Surprisal and Attentional Features

Figure 2 shows Δllh of surprisal predictors and attentional features at both word and segment levels. Data points shown in red indicate predictors that are statistically significant ($p < .05$) relative to the baseline model. When fitted on all language pairs, s_{lm} (Eq. 1) is a significant predictor of source text reading time and target production duration, but not target reading time. On the target side, s_{mt} (Eq. 2) is the best overall predictor of difficulty. Target reading time has fewer significant predictors than does target production duration, and therefore appears to be harder to predict. From here on, we restrict our analyses to features that are significant at at least one of the two segmentation levels.

Figure 3 shows analogous results for four individual language pairs. Surprisal features s_{lm} and s_{mt} remain strong predictors for reading time and target production duration in general. Among the attentional features, $f^{\text{e}}_{u,\bar{u}}$ and $H^{\text{e}}_{u,x}$ (Eq. 5 and 8) most consistently predict source reading time across language pairs and segmentation levels. However, the predictions of target reading time and duration by attentional features are less consistent—despite predicting en → da target difficulty, most attentional features fail to contribute in other language pairs. One possible reason is that these features overfit to small samples of individual language pairs, compared to a mixed model that is fitted on a much larger data set including all language pairs.

7.2 Attention is Supplementary to s_{lm} and s_{mt}

Our results so far confirm that both s_{lm} and s_{mt} individually predict translation difficulty, whereas attentional features on their own are less consistent. We next ask if the attentional features that proved significant in Section 7.1 provide supplementary predictive power when combined with s_{lm} and s_{mt}. To predict source reading time, we train models that include control features, s_{lm} and one attentional feature. For target difficulty, the models are trained with control features, s_{mt} and an attentional feature. We then calculate two variants of Δllh for these models; the first compares against the baseline model, and the second compares against a model that is trained on control features and either s_{lm} or s_{mt}.

We repeat the same significance tests as before, and the results are shown in Figure 4. For the entire data set (top row of Figure 4), models with the addition of individual attentional features predict translation difficulty better than those trained with surprisal and control features only (except $H^{\text{c}}_{v,x}$, Eq. 11). Again, however, these results are weaker for individual language pairs.

7.3 Predictor Coefficients

Thus far we have only demonstrated the predictive power of surprisal and attentional features. To enable conclusions about the nature of the relationship between individual features and translation difficulty, Figure 5 shows average mixed model

the predictor feature and the target variable is not adequately captured by the model class used (in our case, linear models).

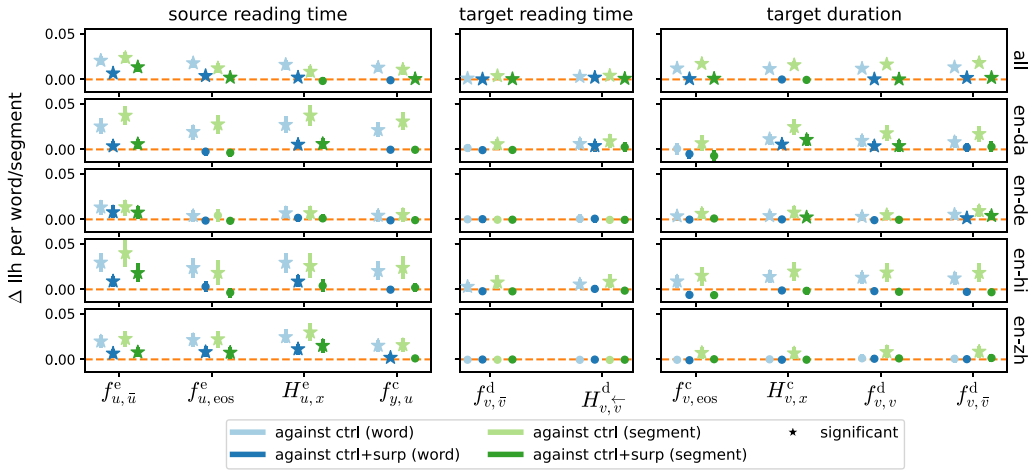
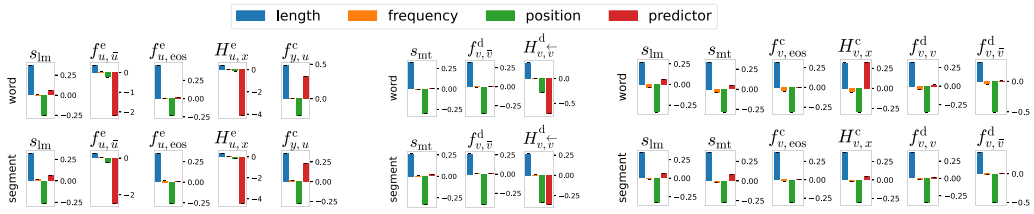


Figure 4: Δllh with one feature in addition to s_{lm}/s_{mt} and control features in training. Light and dark colored stars indicate Δllh to be significantly above zero when compared against control features and control features with surprisal respectively. For example, $f^{\text{c}}_{v,\texttt{eos}}$ significantly contributes in addition to s_{mt} in predicting target duration at both word and segment levels, but $H^{\text{c}}_{v,x}$ does not contribute predictive power beyond s_{mt}.



(a) Source reading time (b) Target reading time (c) Target production duration

Figure 5: Predictor coefficients for linear models that include all three control features (length, frequency, position) along with one additional predictor (either surprisal or an attentional feature). Confidence intervals are plotted but are hard to see because they are so narrow.

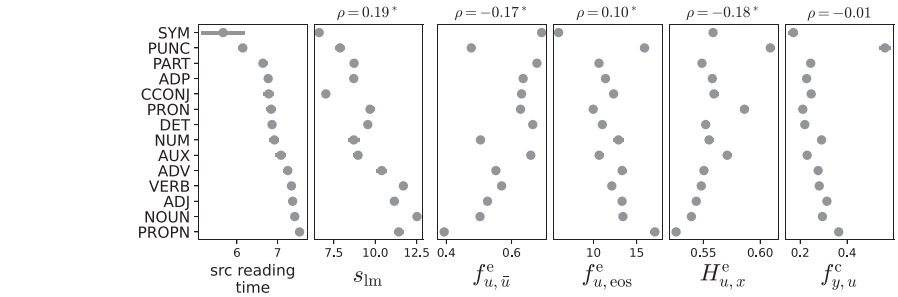
coefficients over data folds. The coefficients plotted support conclusions about the direction and effect size of the relationship between each predictor and translation difficulty, but the bar heights may not reflect predictive power, which has been indicated previously by Δllh.[10] In general, segments are more difficult when they are longer and occur earlier in the sequence. Rare words in general take longer to read and produce, as our baseline models consistently converge to negative coefficients for frequency (not shown in the figure). However, with the addition of surprisal and attentional features, the frequency effect for rare source words is reversed, whereas rare targets still require more attention and take longer to produce. As expected, increases in s_{lm} and s_{mt} are associated with increased reading time and production duration.

On the source side, the coefficients related to encoder self-attention indicate that harder-to-translate source texts direct less attention to context ($f^{\text{e}}_{u,\bar{u}}$, Eq. 5) and more to `eos` ($f^{\text{e}}_{u,\texttt{eos}}$, Eq. 6), which reduces their entropy ($H^{\text{e}}_{u,x}$, Eq. 8). Difficult source words are also singled out as important by having more incoming cross-attention from the target sequence ($f^{\text{c}}_{y,u}$, Eq. 9).

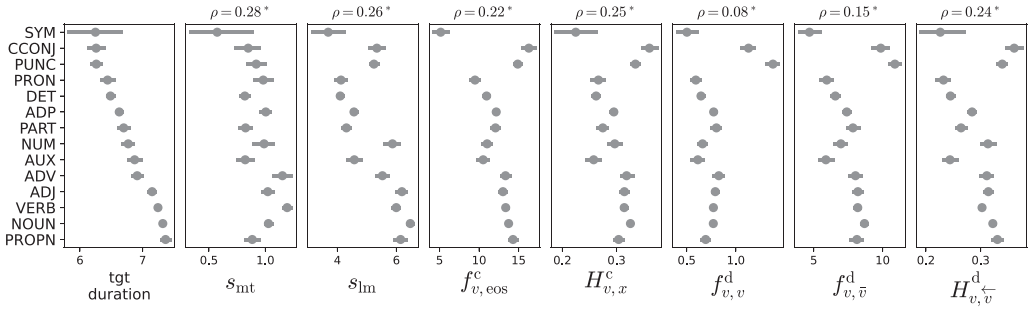
On the target side, harder translations tend to show slight increases in cross-attention to source `eos` ($f^{\text{c}}_{v,\texttt{eos}}$, Eq. 10), and show more diffuse attention across the source sequence ($H^{\text{c}}_{v,x}$, Eq. 11).[11] Our results thus support Dankers et al.'s (2022) claim that paraphrases show increased attention to `eos` and take longer to produce than literal translations.

[10] We tested for collinearity by calculating variance inflation factors (VIF) for each set of features. The highest VIF is 2.2, which is below the thresholds (2.5/3) recommended by Szmrecsanyi (2006) and Zuur et al. (2010).

[11] Mean coefficients of $f^{\text{c}}_{v,\texttt{eos}}$ for token and segment are .001 and .002, respectively.



(a) Source word reading time and predictors



(b) Average translation duration and predictors

Figure 6: Values are grouped by part-of-speech tags based on the multiLing corpus, and are sorted in order of difficulty. Compared to function words, open-class words take longer to read and translate. Pearson correlations with translation difficulty (leftmost panel) are shown at the top of the predictor panels, and '*' indicates statistical significance ($p < .001$).

Figures 5b and 5c also suggest that harder translations have more informative decoder attention ($H^{\mathrm{d}}_{v,\overleftarrow{v}}$, Eq. 14), and direct more attention to themselves ($f^{\mathrm{d}}_{v,v}$, Eq. 12) and the context ($f^{\mathrm{d}}_{v,\bar{v}}$, Eq. 13). These results imply reduced attention to `bos`, the initial token of a translation sequence that conveys the target language to the NLLB model.

8 Discussion

Section 7.1 showed that monolingual surprisal predicts source reading time, but that translation surprisal is a more consistent predictor of target reading time and production duration. On its own, NMT attention also predicts translation difficulty to some degree, but the most accurate predictions are achieved by combining surprisal and attentional features.

8.1 Psycholinguistic Implications

Our results support previous findings that surprisal predicts translation difficulty (Wei, 2022; Teich et al., 2020; Carl, 2021a). Surprisal has several justifications as a cognitive difficulty metric (Levy, 2013; Futrell and Hahn, 2022), and one approach interprets surprisal as a measure of a shift in cognitive resource allocation. On this account, higher translation surprisal indicates that more effort is needed to shift cognitive resources to the word ultimately selected (Wei, 2022).

Teich et al. (2020) suggest that translators aim for translations that are both faithful to the source ($p_{\mathrm{mt}}(t|s)$ is high) and fluent ($p_{\mathrm{lm}}(t)$ is high). These goals do not always align, and capture two different translation strategies—literal translation optimizes MT probability at the expense of LM probability, whereas figurative translation prioritizes the latter. Although increases in s_{lm} and s_{mt} both predict increased target difficulty, our data reveals that these predictors have a weak but significant negative correlation ($p < .001$) at both token ($\rho = -.053$) and segment ($\rho = -.079$) levels. We therefore find quantitative support for a trade-off between fidelity and fluency (Müller et al., 2020; Lim et al., 2024).

8.2 Translatability by Parts of Speech

To gain more insight into the aspects of translation difficulty captured by surprisal and attention, we analyzed human difficulty and model predictions for different parts of speech. All results that follow are based on a subset of studies that use the same

source texts, *multiLing*, a small sample of news articles and sociological texts in English.[12] We break down the difficulty of these English words by their part-of-speech (POS) tags, which are available in the corpus.[13]

Figure 6a shows reading time, s_{lm} and attentional features of words grouped by their POS tags. Compared to function words, open-class words, such as proper nouns, nouns and adjectives, are the most difficult to translate and have higher surprisal. These words also direct more attention to `eos` and less to the context, and attract more cross-attention from the translated sequence.

For target difficulty, the distinction between open-class and function words is also evident in Figure 6b. For each source word, translation duration is defined as the duration of the target segment aligned with the source word divided by the number of alignments between the target segment and the source sentence. Translations of coordinating conjunctions and punctuation stand out as among the easiest by humans, but are surprising for the LM and difficult for NMT. One possible reason is that conjunctions can be cross-linguistically ambiguous (Li et al., 2014; Gromann and Declerck, 2014; Novák and Novák, 2022). For example, English ‘‘but’’ and ‘‘and’’ have been shown to affect NMT fluency (Popović and Castilho, 2019; Popović, 2019). For punctuation, He et al. (2019) demonstrate that the importance of these tokens in NMT can vary by language pairs; for example, translation to Japanese often relies on punctuation to demarcate coherent groups, which is useful for syntactic reordering.

9 Conclusion

Our results support the prevailing view that current NLP models, including LM and NMT, align partially with human language usage and are predictive of language processing complexity. We evaluated surprisal and NMT attention as predictors of human translation difficulty, and found that both factors predict reading times and production duration. Previous work provides some evidence that surprisal and NMT attention capture important aspects of translation difficulty, and our work strengthens this conclusion by estimating surprisal based on state-of-the-art models and analyzing data based on 13 language pairs and hundreds of human translators.

Although the attentional features we consider are empirically successful and grounded in prior literature, they are not without limitations. These features are relatively simple and combining attention weights in more sophisticated ways may allow stronger predictions of human translation difficulty. A more theoretically motivated approach that builds on recent studies of the interpretability of attention distributions (Vashishth et al., 2019; Zhang et al., 2021; Madsen et al., 2022) is worth exploring to develop more fine-grained predictors of translation processing.

To work with as much data as possible, we focused primarily on analyses that combine data from all 13 language pairs, but analyzing translation challenges in individual language pairs is a high priority for future work. A possible next step is an analysis exploring whether the predictors considered here are sensitive to constructions in specific languages that are known sources of processing difficulty (Campbell, 1999; Vanroy, 2021). Factors such as surprisal and attentional flow are appealing in part because their generality makes them broadly applicable across languages, but understanding the idiosyncratic ways in which each pair of languages poses translation challenges is equally important.

Acknowledgments

We thank the action editor and reviewers for thoughtful feedback that improved this work. This project was supported by ARC FT190100200.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884. https://doi.org/10.18653/v1/N19-1388

Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney. 2018. On the alignment problem in multi-head attention-based neural machine

[12] Studies included from multiLing corpus are RUC17, ENJA15, NJ12, STC17, SG12, ENDU20 and BML12.

[13] POS tags are predictions of NLTK tagger converted to universal POS tags.

translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 177–185. https://doi.org/10.18653/v1/W18-6318

Fabio Alves and José Luiz Gonçalves. 2013. Investigating the conceptual-procedural distinction in the translation process: A relevance-theoretic analysis of micro and macro translation units. *Target. International Journal of Translation Studies*, 25(1):107–124. https://doi.org/10.1075/target.25.1.09alv

Laura Winther Balling, Kristian Tangsgaard Hvelplund, and Annette C. Sjørup. 2014. Evidence of parallel processing during translation. *Meta*, 59(2):234–259. https://doi.org/10.7202/1027474ar

Jesús Calvillo, Le Fang, Jeremy Cole, and David Reitter. 2020. Surprisal predicts code-switching in Chinese-English bilingual text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4029–4039. https://doi.org/10.18653/v1/2020.emnlp-main.330

Stuart Campbell. 1999. A cognitive approach to source text difficulty in translation. *Target. International Journal of Translation Studies*, 11(1):33–63. https://doi.org/10.1075/target.11.1.03cam

Michael Carl. 2012. The CRITT TPR-DB 1.0: A database for empirical human translation process research. In *Workshop on Post-Editing Technology and Practice*.

Michael Carl. 2021a. Information and entropy measures of rendered literal translation. *Explorations in Empirical Translation Process Research*, pages 113–140. https://doi.org/10.1007/978-3-030-69777-8_5

Michael Carl. 2021b. Translation norms, translation behavior, and continuous vector space models. *Explorations in Empirical Translation Process Research*, pages 357–388. Springer. https://doi.org/10.1007/978-3-030-69777-8_14

Michael Carl, Akiko Aizawa, and Masaru Yamada. 2016a. English-to-Japanese translation vs. dictation vs. post-editing: Comparing translation modes in a multilingual setting. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4024–4031.

Michael Carl and María Cristina Toledo Báez. 2019. Machine translation errors and the translation process: A study across different languages. *Journal of Specialised Translation*, 31:107–132.

Michael Carl, Moritz Schaeffer, and Srinivas Bangalore. 2016b. The CRITT translation process research database. In *New Directions in Empirical Translation process research*, pages 13–54. Springer. https://doi.org/10.1007/978-3-319-20358-4_2

Michael Carl and Moritz Jonas Schaeffer. 2017. Why translation is difficult: A corpus-based study of non-literality in post-editing and from-scratch translation. *HERMES-Journal of Language and Communication in Business*, 56:43–57. https://doi.org/10.7146/hjlcb.v0i56.97201

Robert Chen, Roger Levy, and Tiwalayo Eisape. 2021. On factors influencing typing time: Insights from a viral online typing game. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Uriel Cohen-Priva and Dan Jurafsky. 2008. Phone information content influences phone duration. In *Conference on Prosody and Language Processing*.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672v3*.

Raj Dabre and Atsushi Fujita. 2019. Recurrent stacking of layers for compact neural machine translation models. *Proceedings of the*

AAAI Conference on Artificial Intelligence, 33(01):6292–6299. https://doi.org/10.1609/aaai.v33i01.33016292

Samvit Dammalapati, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2019. Expectation and locality effects in the prediction of disfluent fillers and repairs in English speech. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 103–109. https://doi.org/10.18653/v1/N19-3015

Samvit Dammalapati, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2021. Effects of duration, locality, and surprisal in speech disfluency prediction in English spontaneous speech. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 91–101.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? Analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626. https://doi.org/10.18653/v1/2022.acl-long.252

Andrea De Varda and Marco Marelli. 2023. Scaling in cognitive modelling: A multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149. https://doi.org/10.18653/v1/2023.acl-short.14

Silvana Deilen, Ekaterina Lapshinova-Koltunski, and Michael Carl. 2023. Cognitive aspects of compound translation: Insights into the relation between implicitation and cognitive effort from a translation process perspective. *Ampersand*, 11:100156. https://doi.org/10.1016/j.amper.2023.100156

Vera Demberg, Asad Sayeed, Philip Gorinski, and Nikolaos Engonopoulos. 2012. Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 356–367.

Barbara Dragsted. 2010. Coordination of reading and writing processes in translation: An eye on uncharted territory. In *Translation and Cognition*, pages 41–62. John Benjamins Publishing Company. https://doi.org/10.1075/ata.xv.04dra

Javier Ferrando and Marta Ruiz Costa-jussà. 2021. Attention weights in transformer NMT fail aligning words between sequences but largely explain model predictions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 434–443. https://doi/10.18653/v1/2021.findings-emnlp.39

Richard Futrell and Michael Hahn. 2022. Information theory as a bridge between language function and language form. *Frontiers in Communication*, 7:657725. https://doi.org/10.3389/fcomm.2022.657725

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. https://doi.org/10.18653/v1/W18-0102

Dagmar Gromann and Thierry Declerck. 2014. A cross-lingual correcting and completive method for multilingual ontology labels. *Towards the Multilingual Semantic Web: Principles, Methods and Applications*, pages 227–242. https://doi.org/10.1007/978-3-662-43585-4_14

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 on Language technologies 2001-NAACL'01*. Association for Computational Linguistics. https://doi.org/10.3115/1073336.1073357

John Hale. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32:101–123. https://doi.org/10.1023/A:1022492123056

Shilin He, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael Lyu, and Shuming Shi. 2019. Towards understanding neural machine translation with word importance. 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 953–962. https://doi.org/10.18653/v1/D19-1088

Kristian Tangsgaard Hvelplund Jensen, Annette C. Sjørup, and Laura Winther Balling. 2009. Effects of L1 syntax on L2 translation. *Copenhagen Studies in Language*, 38:319–336.

Dan Jurafsky. 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. *Probabilistic Linguistics*, 21:1–30. https://doi.org/10.7551/mitpress/5582.003.0006

Beata Beigman Klebanov, Michael Suhan, Zuowei Wang, and Tenaha O'Reilly. 2023. A dynamic model of lexical experience for tracking of oral reading fluency. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 567–575. https://doi.org/10.18653/v1/2023.bea-1.48

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575. https://doi.org/10.18653/v1/D19-1167

Manoj Kumar, Ariel Goldstein, Sebastian Michelmann, Jeffrey M. Zacks, Uri Hasson, and Kenneth A. Norman. 2023. Bayesian surprise predicts human event segmentation in story listening. *Cognitive Science*, 47(10):e13343. https://doi.org/10.1111/cogs.13343

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. https://doi.org/10.18653/v1/2021.acl-long.405

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Roger Levy. 2013. Memory and surprisal in human sentence comprehension. In R. P. G. van Gompel, editor, *Sentence Processing*, pages 78–114. Psychology Press Hove, UK.

Roger Levy and Tim Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pages 849–856. https://doi.org/10.7551/mitpress/7503.003.0111

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Cross-lingual discourse relation analysis: A corpus study and a semi-supervised classification system. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 577–587.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303. https://doi.org/10.18653/v1/P19-1124

Yingcong Li, Yixiao Huang, Muhammed E. Ildiz, Ankit Singh Rawat, and Samet Oymak. 2024. Mechanics of next token prediction with self-attention. In *International Conference on Artificial Intelligence and Statistics*, pages 685–693. PMLR.

Zheng Wei Lim, Trevor Cohn, Charles Kemp, and Ekaterina Vylomova. 2023. Predicting human translation difficulty using automatic word alignment. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11590–11601. https://doi.org/10.18653/v1/2023.findings-acl.736

Zheng Wei Lim, Ekaterina Vylomova, Trevor Cohn, and Charles Kemp. 2024. Simpson's paradox and the accuracy-fluency tradeoff in translation. *arXiv preprint arXiv:2402.12690v2*.

Tal Linzen and Tim Florian Jaeger. 2016. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, 40(6):1382–1411. https://doi.org/10.1111/cogs.12274

Matthew W. Lowder, Wonil Choi, Fernanda Ferreira, and John M. Henderson. 2018. Lexical predictability during natural reading: Effects

of surprisal and entropy reduction. *Cognitive Science*, 42(4):1166–1183. https://doi.org/10.1111/cogs.12597

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2021. Attention calibration for transformer in neural machine translation. 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 1288–1298. https://doi.org/10.18653/v1/2021.acl-long.103

Pedro Macizo and M. Teresa Bajo. 2004. When translation makes the difference: Sentence processing in reading and translation. *Psicológica*, 25(2):181–205.

Pedro Macizo and M. Teresa Bajo. 2006. Reading for repetition and reading for translation: Do they involve the same processes? *Cognition*, 99(1):1–34. https://doi.org/10.1016/j.cognition.2004.09.012

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural NLP: A survey. *ACM Computing Surveys*, 55(8):1–42. https://doi.org/10.1145/3546577

Zofia Malisz, Erika Brandt, Bernd Möbius, Yoon Mi Oh, and Bistra Andreeva. 2018. Dimensions of segmental variability: Interaction of prosody and surprisal in six languages. *Frontiers in Communication*, 3:25. https://doi.org/10.3389/fcomm.2018.00025

Roman Matasov. 2017. Nuremberg: The trial of six million words. https://aiic.org/document/995/

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980. https://doi.org/10.18653/v1/2021.emnlp-main.74

Bartolomé Mesa-Lao. 2014. Gaze behaviour on source texts: An exploratory study comparing translation and post-editing. In *Post-editing of Machine Translation: Processes and Applications*, pages 219–245. Cambridge Scholars Publishing.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 955–960. https://doi.org/10.18653/v1/D16-1096

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408.

Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237. https://doi.org/10.18653/v1/2023.wmt-1.82

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

Jean Nitzke. 2019. *Problem Solving Activities in Post-editing and Translation from Scratch: A Multi-method Study*. Language Science Press. https://doi.org/10.4324/9780429030376-5

Attila Novák and Borbála Novák. 2022. Cross-lingual transfer of knowledge in distributional language models: Experiments in Hungarian. *Acta Linguistica Academica*, 69(4):405–449. https://doi.org/10.1556/2062.2022.00580

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. https://doi.org/10.18653/v1/2022.emnlp-main.632

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off

across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 949–962. https://doi.org/10.18653/v1/2021.emnlp-main.73

Dagmara Płońska. 2016. Problems of literality in French-Polish translations of a newspaper article. *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*, pages 279–291. https://doi.org/10.1007/978-3-319-20358-4_13

Maja Popović. 2019. Evaluating conjunction disambiguation on English-to-German and French-to-German WMT 2019 translation hypotheses. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 464–469. https://doi.org/10.18653/v1/W19-5353

Maja Popović and Sheila Castilho. 2019. Are ambiguous conjunctions problematic for machine translation? In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 959–966. https://doi.org/10.26615/978-954-452-056-4_111

Ana Rojo. 2015. Translation meets cognitive science: The imprint of translation on cognitive processing. *Multilingua*, 34(6):721–746. https://doi.org/10/multi-2014-0066

Andrea Gerardo Russo, Maria De Martino, Annibale Elia, Francesco Di Salle, and Fabrizio Esposito. 2022. Negative correlation between word-level surprisal and intersubject neural synchronization during narrative listening. *Cortex*, 155:132–149. https://doi.org/10.1016/j.cortex.2022.07.005

Andrea Gerardo Russo, Maria De Martino, Azzurra Mancuso, Giorgio Iaconetta, Renzo Manara, Annibale Elia, Alessandro Laudanna, Francesco Di Salle, and Fabrizio Esposito. 2020. Semantics-weighted lexical surprisal modeling of naturalistic functional MRI time-series during spoken narrative listening. *Neuroimage*, 222:117281. https://doi.org/10.1016/j.neuroimage.2020.117281

Soo Hyun Ryu and Richard L. Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71. https://doi.org/10.18653/v1/2021.cmcl-1.6

Moritz Schaeffer and Michael Carl. 2014. Measuring the cognitive effort of literal translation processes. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 29–37. https://doi.org/10.3115/v1/W14-0306

Márcia Schmaltz, Igor AL da Silva, Adriana Pagano, Fabio Alves, Ana Luísa V. Leal, Derek F. Wong, Lidia S. Chao, and Paulo Quaresma. 2016. Cohesive relations in text comprehension and production: An exploratory study comparing translation and post-editing. *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*, pages 239–263. https://doi.org/10.1007/978-3-319-20358-4_11

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. SeamlessM4T: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596v3*.

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. https://doi.org/10.1073/pnas.2307876121

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 11:58–79. https://doi.org/10.1162/tacl_a_00633

Gregory M. Shreve and Isabel Lacruz. 2017. Aspects of a cognitive model of translation. *The Handbook of Translation and Cognition*, pages 127–143. https://doi.org/10.1002/9781119241485.ch7

Gregory M. Shreve, Christina Schäffner, Joseph H. Danks, and Jennifer Griffin. 1993. Is there a special kind of ''reading'' for translation? An empirical investigation of reading in the translation process. *Target*, 5(1):21–41. https://doi.org/10.1075/target.5.1.03shr

Annette Camilla Sjørup. 2013. *Cognitive effort in metaphor translation: An eye-tracking and key-logging study*. Frederiksberg: Copenhagen Business School (CBS).

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Sanjun Sun. 2015. Measuring translation difficulty: Theoretical and methodological considerations. *Across Languages and Cultures*, 16(1):29–54. https://doi.org/10.1556/084.2015.16.1.2

Benedikt Szmrecsanyi. 2006. *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Mouton de Gruyter. https://doi.org/10.1515/9783110197808

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2019. Encoders help you disambiguate word senses in neural machine translation. 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 1429–1435. https://doi.org/10.18653/v1/D19-1149

Elke Teich, José Martínez Martínez, and Alina Karakanta. 2020. Translation, information theory and cognition. *The Routledge Handbook of Translation and Cognition*, pages 360–375. https://doi.org/10.4324/9781315178127-24

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85.

Bram Vanroy. 2021. *Syntactic Difficulties in Translation*. Ph.D. thesis, Ghent University.

Bram Vanroy, Orphée De Clercq, and Lieve Macken. 2019. Correlating process and product data to get an insight into translation difficulty. *Perspectives: Studies in Translation Theory and Practice*, 27(6):924–941. https://doi.org/10.1080/0907676X.2019.1594319

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks. *arXiv preprint arXiv:1909.11218v1*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lucas Nunes Vieira, Natalie Zelenka, Roy Youdale, Xiaochun Zhang, and Michael Carl. 2023. Translating science fiction in a CAT tool: Machine translation and segmentation settings. *Translation & Interpreting*, 15(1):216–235. https://doi.org/10.12807/ti.115201.2023.a11

Yuxiang Wei. 2022. Entropy as a measurement of cognitive load in translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas*

(Workshop 1: Empirical Translation Process Research), pages 75–86.

Ethan Wilcox, Pranali Vani, and Roger Levy. 2021. A targeted assessment of incremental processing in neural language models and humans. 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 939–952. https://doi.org/10.18653/v1/2021.acl-long.76

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

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470. https://doi.org/10.1162/tacl_a_00612

Yilin Yang, Longyue Wang, Shuming Shi, Prasad Tadepalli, Stefan Lee, and Zhaopeng Tu. 2020. On the sub-layer functionalities of transformer decoder. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4799–4811. https://doi.org/10.18653/v1/2020.findings-emnlp.432

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. 2021. Do context-aware translation models pay the right attention? 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 788–801. https://doi.org/10.18653/v1/2021.acl-long.65

Chi-Lin Yu, Rachel Eggleston, Kehui Zhang, Nia Nickerson, Xin Sun, Rebecca A. Marks, Xiaosu Hu, Jonathan R. Brennan, Henry Wellman, and Ioulia Kovelman. 2023. Neural processing of children's theory of mind in a naturalistic story-listening paradigm. *PsyArXiv*.

Shaolei Zhang and Yang Feng. 2021. Modeling concentrated cross-attention for neural machine translation with Gaussian mixture model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1401–1411. https://doi.org/10.18653/v1/2021.findings-emnlp.121

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. 2021. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742. https://doi.org/10.1109/TETCI.2021.3100641

Alain F. Zuur, Elena N. Ieno, and Chris S. Elphick. 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1(1):3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x