Human and Machine: Language Processing in Translation Tasks

Hening Wang[1], **Leixin Zhang**[2], and **Ondřej Bojar**[3]

[1]University of Tübingen, Seminar für Sprachwissenschaft
[2]University of Twente, High-tech Business and Entrepreneurship
[3]Charles University, Faculty of Mathematics and Physics, ÚFAL

Abstract

The present study analyzes the influence of linguistic factors (sentence ambiguities) and non-linguistic factors (visual cues) on online language processing in translation tasks. Moreover, it also offers an attempt at relating machine and human translation in a multimodal setting, an aspect that has received less attention before. We qualitatively evaluated translation outputs between subjects across different experimental conditions, as well as between human and machine translation processes. We observed a positive correlation between humans' reading time and models' next token prediction, with a higher similarity score for the translation of unambiguous sentences compared to translations of ambiguous sentences. We also found that a context-relevant image has a significant influence on translation updates.

1 Introduction

Translation is an important aspect of language use. A vast number of machine translation models have been developed over the last decades trying to assist and automatize this task. However, less attention has been paid to the architecture and mechanism of language processing during translation tasks and the relation between these processes in humans and in machines.

We attempt to provide a new perspective by taking translation as the task in assessing language processing and comparing human and machine processing in it. It is clear that the mechanisms are fundamentally different between the human brain and machine translation (MT) systems. However, according to the three levels of analysis proposed by Marr (1982), studying human translation processes can reveal how people handle ambiguity, context, and non-linguistic information. This knowledge can inform the development of more sophisticated and human-like MT systems at the computational level.

In real-world scenarios, human language processing is further compounded by external stimuli such as images or sounds, which can either assist, hinder or distract human comprehension. By examining how humans process sentences in real-time, we can identify strategies to improve machine translation algorithms, making them more adaptable and contextually aware. While some studies focus on specific aspects, such as using eye-tracking to evaluate MT systems (Doherty et al., 2010; Stymne et al., 2012) or using EEG to measure effort during human translation (Hansen-Schirra, 2017), it remains challenging to unravel how the working mechanisms of machines differ from those of humans and to what extent they are comparable (Wang et al., 2023; Lakretz et al., 2021).

In this study, we combined eye-tracking data to analyze human language processing and used surprisal obtained from GPT-2 to represent the processing of models. Our experiments examine language processing for both ambiguous and unambiguous sentences, presented with or without relevant visual cues. Furthermore, we compared machine translation outputs based solely on textual input and human translations performed under three different visual stimuli. We assess the potential influence of visual cues on human comprehension and evaluate whether multi-modal machine translation is necessary for reaching human-like performance in our setting. This inquiry is particularly pertinent due to the inconclusive results in integrating visual stimuli to enhance machine translation (Specia et al., 2016; Elliott, 2018; Caglayan et al., 2019). This attempt also allows us to relate human cognitive processes to artificial systems in future research.

The objective of the current study is to assess whether machine processing can be numerically correlated with human language processing in translation tasks. The main research question can be formulated as the following two respects:

- **Research Question 1**: Do machines and humans exhibit comparable difficulties in processing ambiguous vs. unambiguous sentences?

  Hypothesis: Higher processing complexity should be shown for both humans and machines.

- **Research Question 2**: Do visual cues impact human translation outcomes, and which visual condition in human translation aligns better with the machines' outcomes that rely solely on text?

  Hypothesis: Visual conditions affect human translation, and machine, text-only processing should be more similar to human translation results when no additional visual cues are provided.

The following sections are organized as follows: Section 2 provides an overview of previous studies on language processing and highlights the research gap in language processing, particularly in human-machine comparison that we are addressing in this study. Section 3 introduces the corpus we used in our study. Section 4 focuses on the input processing in humans and models (machine), while Section 5 analyzes the output of language processing by human and machine processing.

2 Background in Language Processing

In human language processing studies, reading time serves as a crucial measure for assessing language processing difficulty. In psycholinguistic research, there has been a comprehensive study of the correlation between processing difficulty and longer reading duration (Underwood et al., 2000; Juhasz and Rayner, 2003; Rayner and Raney, 1996). In the studies of eye-tracking techniques and language processing, fixation duration can be an indicator of processing complexity. Specifically, shorter fixation durations have been associated with more predictable words, whereas longer durations have been linked to unpredictable words (Ehrlich and Rayner, 1981).

For statistical models, surprisal theory provides a measure of the difficulty of language processing (Hale, 2001; Levy, 2008; Boston et al., 2011). Surprisal estimates how surprising or unlikely the next word appears based on the partially established structure of the sentence. For instance, the processing difficulty of garden path sentences can be captured by surprisal (Hale, 2001). In previous studies, surprisal shows a positive correlation with reading time (Smith and Levy, 2013; Monsalve et al., 2012; Goodkind and Bicknell, 2018). Roger (2008) proposes that the word surprisal is proportional to the negative log probability of words.

Another method to investigate processing difficulty can be translation output from a source language to a target language. It is found that the source text is one factor that affects translation (Campbell, 1999; Dragsted, 2012). Tokowicz and Degani (2010) state that ambiguity slows translation and can reduce translation accuracy due to the competition of potential target translation choices. Heilmann (2020) and Hvelplund (2014) study the language processing in the setting of translation tasks and state that the focus (longer gazing duration) on the source text corresponds to more translation options in the target language system. Dragsted also (2012) found that high variability of translation output is related to higher reading duration and self-corrections.

In previous studies, the complexity of language processing has rarely been examined under the human-machine comparison setting. We attempt to provide a new perspective by taking translation as the principal task in assessing language processing and comparing human and machine processing.

3 Corpus: EMMT

We use Eyetracked Multi-Modal Translation (EMMT) corpus (Bhattacharya et al., 2022) for our research. The corpus comprises 200 sentences, categorized into two types, ambiguous and unambiguous, with 100 sentences in each category.

In this corpus, source sentences are in English and they were translated into Czech. Each participant went through two rounds of reading and translating phases. In the first round, only a plain sentence was shown and the subjects were expected to say its translation into Czech aloud. In the subsequent phase, one of three visual conditions was provided: a relevant picture, an irrelevant picture, or no image. Subjects were expected to confirm their previous translation, or say an updated version. Both ambiguous and unambiguous sentences were distributed equally among the participants and across three visual conditions.

4 Input Processing

This section studies the input processing of humans and models. In Section 2, we discussed the surprisal theory and its correlation to human language processing, however, it is not yet confirmed whether surprisal also correlates with text reading specifically for translation purposes. Our study aims to fill the existing gap.

In our experiments, we also test whether an intrinsic factor (sentence ambiguity) has an impact on the language processing of humans (measured by reading duration) and the model (measured by surprisal obtained from GPT-2), and investigate whether the model's surprisal correlates with human's reading duration.

We compute the reading duration for each sentence based on eye-tracking data. The eye tracker collects data with an interval of approximately 0.5 milliseconds between each two adjacent time points. The overall reading duration of a

As the machine counterpart to human processing duration, we take the model's surprisal: the method of negative logarithm of probability proposed by Levy (2008) is adopted to compute surprisal. In addition, we view human language processing as an incremental procedure, where meaning is obtained as words are encountered in a sequential manner (Brouwer et al., 2010). Guided by this premise, we utilize the generative model GPT-2 (large) to derive word probabilities.

The probability of the next word is obtained one at a time with previous words in the sentence serving as a prompt. The predicted difficulty of a sentence is computed as the sum of negative logarithms of the conditional probabilities of the words in the sentence (excluding the first word of the sentence, which only serves as the prompt). For example, we calculate the surprisal of the sentence 'The stand is stable' as Equation (1).

$$\begin{aligned}\text{Surprisal} = &-\log\left(\text{P}\left(\textit{stand} \mid \textit{The}\right)\right) \\ &- \log\left(\text{P}\left(\textit{is} \mid \textit{The stand}\right)\right) \\ &- \log\left(\text{P}\left(\textit{stable.} \mid \textit{The stand is}\right)\right) \quad (1)\end{aligned}$$

Table 1 presents the results of fixation duration in two groups (ambiguous and unambiguous) when reading source texts. The results indicate a slightly longer duration that participants dedicated to reading ambiguous sentences as opposed to unambiguous sentences. However, it is noteworthy that this difference between the two groups is not statistically significant (T-test: $p = 0.161$).

	Ambiguous	**Unambiguous**
Reading (sec)	7.637	7.334

Table 1: Reading time during sentence reading phrase.

	Ambiguous	**Unambiguous**
Surprisal value	51.21	49.56

Table 2: Sentence surprisal value obtained from GPT-2.

The surprisal values for both the ambiguous group and unambiguous group are displayed in Table 2. The table demonstrates that GPT-2 perceives ambiguous sentences to be marginally more surprising than unambiguous sentences. Similarly to human processing, the difference between the two groups is not statistically significant as indicated by T-test ($p = 0.162$).

We further analyze the correlation between sentence reading duration and surprisal using Pearson's correlation coefficient (r). The results indicate a moderate positive correlation between the two ($r = 0.507$). Analyzing the ambiguous and unambiguous sentence groups individually, we find correlations of $r = 0.58$ for the unambiguous group is higher than $r = 0.43$ for the ambiguous group. This suggests that the alignment between human reading time and the model's surprisal is more pronounced in the case of unambiguous sentences.

5 Translation Outputs

In this section, we analyze the translation outputs as the results of language processing. Three experiments were implemented to investigate two factors (ambiguity and visual cues): (1) a comparison between the initial translation and subsequent updated version by the same subjects (Section 5.1); (2) a comparison of the translation outputs across different subjects (Section 5.2); and (3) a comparison of the translation outputs between humans and machine translation systems (Section 5.3).

On the one hand, we explore the effect of sentence ambiguity on translation outputs. Our study builds on previous research (Tokowicz and Degani, 2010; Heilmann, 2020; Hvelplund, 2014), which suggests that translation results exhibit greater vari-

ance for sentences that are more challenging to process. Our hypothesis is that ambiguous sentences can be interpreted in different ways, and as a result, their translations should undergo more updates when accompanied by an image in the second translation phase. Moreover, we anticipate that the translation outputs from humans and machines would exhibit greater dissimilarity for ambiguous sentences than for unambiguous ones.

On the other hand, we intend to analyze the influence of visual cues on human translations (Sections 5.1 and 5.2). Specifically, we aim to explore the conditions under which translation outputs demonstrate greater similarity across subjects when considering three different visual cues (a related image, no image, and an unrelated image) (Section 5.2). Additionally, we aim to identify the visual conditions under which human translations exhibit greater similarity to machine-generated translations that rely solely on textual inputs[1] (Section 5.3).

5.1 Translation Updates

This section analyzes translation updates, a comparison between subjects' initial translations (relying solely on source sentences) and their subsequent updated versions (when one of the image conditions is presented).

The similarity of sentence pairs is measured using the Levenshtein distance over words, which is further normalized into a similarity ratio using Equation (2) in order to minimize the impact of varying sentence lengths. This normalized similarity ratio ranges between 0 and 1, where 0 indicates no word overlap in the sentence pair and 1 indicates two sentences are identical. The analysis of translation updates is conducted considering two factors and 6 conditions in total: 2 [AMBIGUITY] × 3 [VISUAL CUES] setting.

$$\text{Ratio} = \frac{\text{len}(\text{Sen}_1) + \text{len}(\text{Sen}_2) - \text{distance}}{\text{len}(\text{Sen}_1) + \text{len}(\text{Sen}_2)} \quad (2)$$

We utilize a two-way ANOVA (with factor interaction considered) to assess the influence of the factors. The initial results confirm that there is no significant interaction between the two factors: AMBIGUITY and VISUAL CUES ($F = 1.385$, $p = 0.251$). The results also reveal that the difference between ambiguous and unambiguous sentences is not statistically significant ($F = 0.273$, $p = 0.251$), suggesting that sentence ambiguity has a minimal effect on translation updates during the second translation phase. We explain this by the nature of ambiguity types observed in EMMT data: depending on the source of the image and sentence, the sentences exhibit syntactic ambiguity (like "I saw a man with the telescope"), for which however the translation into Czech does not need to resolve the ambiguity, or lexical ambiguity (like "court", which is ambiguous between the court of justice and a tennis court), where however the remaining words in the sentence typically provide enough non-visual context for ambiguity resolution. In either case, there is no need to update the translation into Czech. The last common ambiguity type, gender ambiguity (male vs. female tennis player) is not very frequent and its visual resolution is often in line with the stereotypical solution chosen by the translators in the absence of other information.

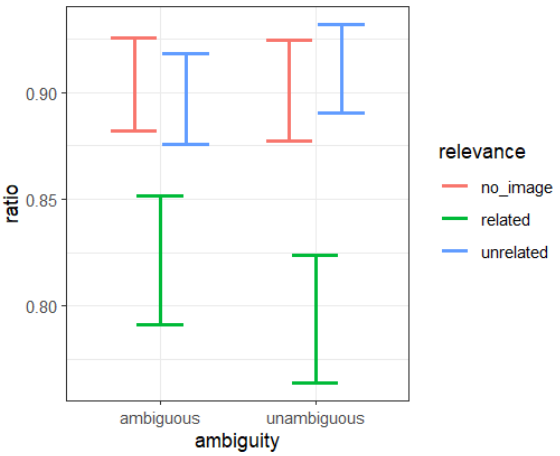

Figure 1: Similarity ratio between the initial translation and the updated translation (error bar plots).

More significantly, the test indicates a notable influence of visual cues ($F = 38.141$, $p < 2e^{-16}$). Figure 1 illustrates that the lowest similarity ratio occurs when a related image is provided in the second round of translation. This suggests that subjects tend to make more updates to their translation when provided with a relevant picture. Further Welch-Satterthwaite t-test shows that 'related images' exhibit a statistically significant effect on the similarity ratio of translation updates ($t = -4.588$, $p = 5.36e^{-06}$) compared to the visual condition

[1]Given the restriction that multi-modal machine translationreadily available, we only compare all visual conditions from humans with one condition from machines, which is only with textual input.

of 'no images'. Our explanation here is based on the observation that the text is often *vague*. The provided image allows the translators consider the general setting in which the sentence was used, and rephrase the translation to be appropriate for this setting.

Finally, there's no significant distinction observed between the 'unrelated image' condition and the 'no image' condition ($\beta = -0.007$, $t = -0.482$, $p = 0.630$).

5.2 Translation Comparison across Subjects

This section presents the analysis of translation similarity across subjects, specifically examining the extent to which translations of the same source sentence, produced by different subjects are similar.

Unlike the previous subsection (Section 5.1), which focuses on updates at the word or lexical level, we now evaluate the similarity of translations across subjects in terms of meaning. For this purpose, we employ the BLEURT metric (Bilingual Evaluation Understudy with Representations from Transformers, Sellam and Parikh, 2020a; Sellam et al., 2020b) to evaluate the similarity of translation pairs.

BLEURT leverages contextualized word representations from BERT to provide a score aligning better with human assessment of translation similarity (Sellam and Parikh, 2020a; Sellam et al., 2020b). The BLEURT score ranges roughly between 0 and 1, with 1 indicating more similar translation pairs and 0 less similar (the score occasionally goes below 0).

We computed BLEURT scores for all translation pairs of the same sentence across the visual conditions and subjects. More specifically, we compare sentence translations in various scenarios, such as when both subjects saw no image (written as 'no-no' for short); when one saw an unrelated image, and the other a related image ('unrelated-related); etc, resulting in a total of 6 visual cues combination conditions. Overall, the study demonstrates a 2 [AMBIGUITY] × 6 [VISUAL CUE COMBINATION] setting.

Considering the repeated sampling when establishing pairwise comparisons and the potential interplay between factors, we employ a linear mixed model (with interaction and random effect structures considered) to examine the impact of ambiguity and visual conditions on cross-subject translation similarity. The linear mixed model was fitted using the REML method, and t-tests using Satterthwaite's method.

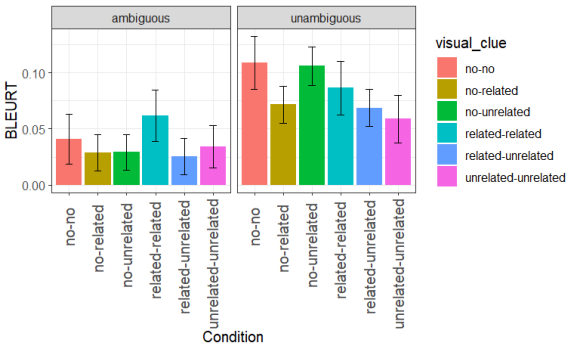

Figure 2: Cross-Subject Translation Similarity

The results are visualized in Figure 2. We observed that translations of unambiguous sentences exhibit higher cross-subject BLEURT scores than those of ambiguous sentences. It implies that unambiguous sentences are translated by humans with less variance, although this result is not statistically significant ($\beta = 6.249e^{-2}$, $t = 1.768$, $p = 0.774$).

Regarding the influence of visual conditions, the linear mixed model demonstrates that the 'related-related' condition is the only one demonstrating a significant effect, compared with the 'no-no' condition ($\beta = 4.908e^{-02}$, $t = 1.963$, $p = 0.0498$). It means that when both subjects are provided a relevant image as translation support, their translation outputs tend to be more similar compared to other visual conditions, supporting also our argument about the related image reducing the information vagueness about the described situation. This finding also implies that a relevant image may help to resolve ambiguity in the ambiguity group.

In the unambiguous group, subjects exhibited the greatest translation similarity when no image was provided for both subjects. The provision of unrelated images ('unrelated-unrelated' condition) results in the least similarity between translation pairs. This suggests that unrelated images might serve as distractions for subjects. However, these findings aren't statistically significant and need further examination to verify.

5.3 Human-Machine Translation Comparison

Following the exploration of translation comparison across subjects, this section compares human translations with translations generated by four machine translation systems: Google, Lindat,[2] DeepL,

[2] https://lindat.mff.cuni.cz/services/translation/

and chatGPT.[3]

Firstly, we investigate which translation systems exhibit greater similarity to human translations. Prior research (Popel et al., 2020) suggests that the Lindat translation model (also known as CUBBITT) demonstrates higher fluency and accuracy levels than other systems, and even surpasses human translation quality. We will test the performance with our sentences and experiment settings.

Secondly, we examine visual conditions under which human translations exhibit greater similarity to machine-generated ones that rely solely on textual inputs. We use BLEURT to measure translation similarity, as in Section 5.2. Concerning ambiguity, we hypothesize that human and machine translations should exhibit greater similarity when translating unambiguous sentences. Our hypothesis regarding visual conditions is that machine translation relying solely on texts should exhibit greater similarity (higher BLEURT scores) to human translations with no images. The linear mixed model is used again (as in Section 5.2) to test the factors.

To better assess the performance of the four models, we additionally established a worst-case baseline by shuffling the Lindat translations which leads to translation pairs without association. The BLEURT score in this case is negative (-0.62). The results show that all four systems show significantly better results than the baseline ($p < 2e^{-16}$). Moreover, t-tests from the linear mix model reveal that chatGPT scores significantly lower than the other three systems ($p < 0.01$). This result might indicate a lower translation quality, but it can also be an artifact due to considerable dissimilarity between LLM-based translation outputs and standard MT outputs.

Additionally, Figure 3 demonstrates that the performance of Lindat stands out as the best among the models, although the difference from Google ($\beta = 3.63e^{-02}$, $t = 1.309$, $p = 0.191$) and DeepL ($\beta = 0.043$, $t = 1.502$, $p = 0.133$) is not statistically significant. This result verifies the performance of Lindat in prior studies.

Regarding the influence of visual cues, the t-tests conducted in the linear mixed model suggest no significant effect is observed. Nevertheless, Figure 3 provides additional insights. It shows that within the unambiguous sentence group, all four translation systems exhibit the highest BLEURT scores

[3]Translations from the respective systems were obtained in March 2023.

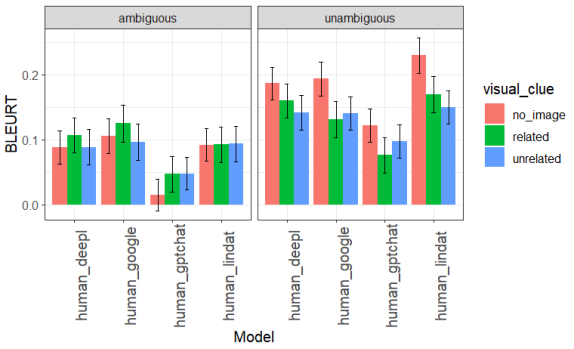

Figure 3: Similarity between human and machine translation estimated by BLEURT score taking the human translation as the reference and each of the MT outputs as candidates.

when their translations are compared to human translation under the 'no image' condition. This observation supports our assumption that machine translation aligns better with human translations without image assistance (though statistically insignificant, $p > 0.05$). Additionally, within the unambiguous group, translations of four systems exhibit lower BLEURT similarity scores with human translation under the condition of 'irrelevant image', compared to the condition of 'no image'. It implies that irrelevant images might distract human translators, resulting in a lower correlation between machine and human translations. However, further research is needed to confirm this hypothesis.

For ambiguous sentence groups, the visual conditions do not show a consistent influence on translation similarity. Translations from Google and DeepL correlate better with human translations when related images are included. However, this pattern is not apparent in Lindat and chatGPT, and the effect remains statistically insignificant.

6 Conclusion

Our study analyzes the language processing of humans and machines in translation tasks and examines the impact of sentence ambiguity and visual cues on sentence processing in translation tasks.

Section 4 suggests that processing from humans and machines correlates with each other: humans exhibit a slightly longer fixation duration, and the model reveals slightly higher surprisal values (showing higher degree of processing complexity) during the processing of ambiguous sentences.

Given the restriction that we cannot assess the machine's translation ability when providing a visual condition, we compared the machine's textual

translation outputs with the human's translation under three visual conditions to see which condition correlates better with the machine's textual translation results. We noted that translations generated by machines tend to exhibit a higher degree of similarity to human translations when subjects are provided only with plain texts. For the unambiguous sentence group, we also observe that machine translations are more similar to human translations with only plain texts provided (without visual cues) compared to conditions with a relevant or irrelevant image.

In the examination of the effect of visual cues on human language processing, we discovered that image conditions display an influence on subjects' translation updates. In particular, when related images are provided, there is a tendency for more word updates in the later translation correction phase. In the context of translation comparison across subjects, we observed that translations tend to be more similar when both subjects are exposed to a related image. Irrelevant images might distract human translators, resulting in a lower similarity between machine and human translations.

7 Acknowledgements

This work has been supported by the GAČR EXPRO grant NEUREM3 (19-26934X), by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ and by the CEDMO 2.0 NPO project.

References

Sunit Bhattacharya, Věra Kloudová, Vilém Zouhar, and Ondřej Bojar. 2022. Emmt: A simultaneous eye-tracking, 4-electrode eeg and audio corpus for multi-modal reading and translation scenarios. *arXiv preprint arXiv:2204.02905*.

Marisa Ferrara Boston, John T Hale, Shravan Vasishth, and Reinhold Kliegl. 2011. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3):301–349.

Harm Brouwer, Hartmut Fitz, and John Hoeks. 2010. Modeling the noun phrase versus sentence coordination ambiguity in Dutch: Evidence from surprisal theory. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–80, Uppsala, Sweden. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal 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 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Stuart Campbell. 1999. A cognitive approach to source text difficulty in translation. *Target. International Journal of Translation Studies*, 11(1):33–63.

Stephen Doherty, Sharon O'Brien, and Michael Carl. 2010. Eye tracking as an MT evaluation technique. *Machine Translation*, 24(1):1–13.

Barbara Dragsted. 2012. Indicators of difficulty in translation—correlating product and process data. *Across Languages and Cultures*, 13(1):81–98.

Susan F Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. 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.

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*.

Silvia Hansen-Schirra. 2017. Eeg and universal language processing in translation. *The handbook of translation and cognition*, pages 232–247.

Arndt Heilmann. 2020. *Profiling effects of syntactic complexity in translation: a multi-method approach*. Ph.D. thesis, Dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 2020.

Kristian Tangsgaard Hvelplund. 2014. Eye tracking and the translation process: Reflections on the analysis and interpretation of eye-tracking data.

Barbara J Juhasz and Keith Rayner. 2003. Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6):1312.

Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699.

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

David Marr. 1982. Vision: A computational investigation into the human representation and processing of visual information.

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.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):4381.

Keith Rayner and Gary E Raney. 1996. Eye movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review*, 3(2):245–248.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020b. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Thibault Sellam and Ankur P Parikh. 2020a. Evaluating natural language generation with bleurt. *Google AI Blog*.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.

Sara Stymne, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, and Martin Wester. 2012. Eye Tracking as a Tool for Machine Translation Error Analysis.

Natasha Tokowicz and Tamar Degani. 2010. Translation ambiguity: Consequences for learning and processing. *Research on second language processing and parsing*, pages 281–293.

Geoffrey Underwood, Alice Binns, and Stephanie Walker. 2000. Attentional demands on the processing of neighbouring words. In *Reading as a perceptual process*, pages 247–268. Elsevier.

Shaonan Wang, Nai Ding, Nan Lin, Jiajun Zhang, and Chengqing Zong. 2023. Language cognition and language computation–human and machine language understanding. *arXiv preprint arXiv:2301.04788*.