Stroop Effect in Multi-Modal Sight Translation

Sunit Bhattacharya, Vilém Zouhar, Věra Kloudová, Ondřej Bojar
Charles University, Faculty Of Mathematics and Physics
Insititute of Formal and Applied Linguistics
{bhattacharya,zouhar,kloudova,bojar}@ufal.mff.cuni.cz

Abstract

This study investigates the human translation process from English to Czech in a multi-modal scenario (images) using reaction times. We make a distinction between ambiguous and unambiguous sentences where in the former, more information would be needed in order to make a proper translation (e.g. gender of the subject). Simultaneously, we also provide visual aid to help in disambiguation, which is necessary for the ambiguous sentences. We confirm that ambiguous sentences take longer to translate and the provision of disambiguating visual aid slows the translation process. When provided with an unrelated visual aid, humans are able to recognize and spend less time on it but still significantly more than in other conditions. These findings are a clear manifestation of the Stroop effect (longer processing times for incongruent combinations).

1 Introduction

Understanding how language is processed in the human brain is a challenging task that started out a philosophical speculation and has eventually evolved into a well structured program of scientific enquiry. Psycholinguistics is one such discipline that attempts to merge a variety of inter-disciplinary concepts to understand the psychological and cognitive mechanisms behind language. The approaches to psycholinguistics have changed significantly since its beginnings in the 1960s (Traxler, 2011). But still, psycholinguistic investigation using behavioural data has mostly relied on the “eye-mind hypothesis” by Just & Carpenter (1976), the core of which states that the eye fixates on whatever is at the “top of the stack” of cognitive processes. The application of psycholinguistic experimental methods, reliant mostly on key-logging and verbal protocols have also been extended to translation process research (summarized by Hvelplund (2011)). Eye-tracking methods to some extent have also utilized the visual world paradigm (Allopenna et al., 1998) to study how people integrate linguistic information and visual information for various tasks (Huettig et al., 2011).

We analyze of the Eyetracked Multi-Modal Translation (EMMT) corpus (Bhattacharya et al., 2022). It consists of observations from a set of psycholinguistic experiments involving translation under multi-modal settings with a number of Czech native speakers with an advanced level of English language proficiency. In the absence of any previous analysis of the data collected as part of the corpus, we focus on the very basics, i.e. looking at the impact of congruent and incongruent image stimuli on the translation process. For this, we focus on response-time metrics to evaluate processing difficulty during different stages corresponding to each stimuli and the respective image congruence conditions.

Section 2 provides a brief introduction to the experiment setup, Section 3 shows quantitative and statistical analysis of the phenomena and Section 4 concludes interpretation of the results. The contributions of this paper are the following:[1]

- Comparison of the reading and the sight translation process (Gile, 2009) for ambiguous and unambiguous sentences through event log files.
- Evaluation of how stimulus congruency (in the form of visual and linguistic parts of the stimuli) affects the translation process.

[1] The code is available open-source github.com/ufal/eyetracked-multi-modal-translation.

2 Data collection setup

The experiment design setup implemented during the collection of data used a combination of sight translation, reading aloud and thinking aloud (Tirkkonen-Condit, 1990) protocols. Bhattacharya et al. (2022) intend to compare the behavioural data of the participants when they *read out* a textual stimulus (sentence) or *looked* at a combination of text and visual (image) stimulus versus when they actually *translated* the textual stimulus. The experiment thus followed four stages corresponding to each stimulus: Stage 1 (READ: reading of the source English sentence), Stage 2 (TRANSLATE: translating the English sentence into Czech), Stage 3 (SEE: observing the corresponding image) and Stage 4 (UPDATE: producing the final translation of the English sentence given the image). See Figure 1 for a vizualization of the experiment setup.

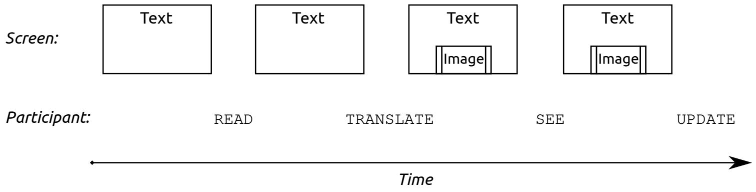
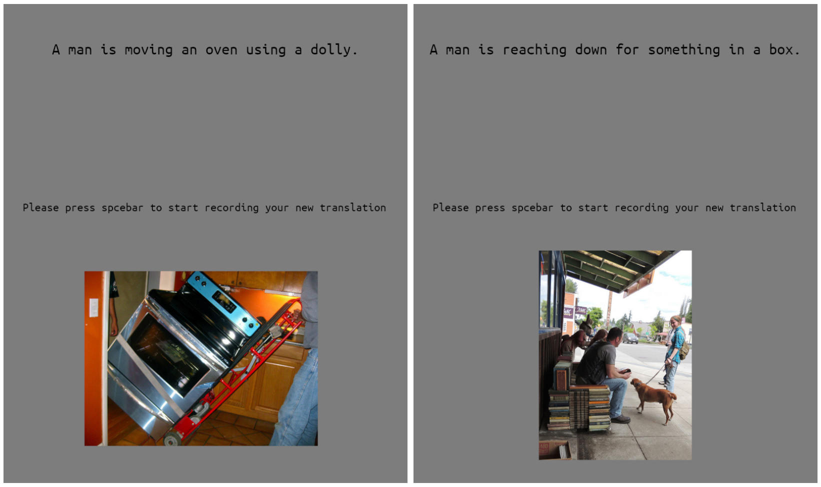


Figure 1: Visualization of the four experiment stages.


A man is moving an oven using a dolly.

Please press spcebar to start recording your new translation

A man is reaching down for something in a box.

Please press spcebar to start recording your new translation

(a) Related image

(b) Unrelated image

Figure 2: Overview of stimulus screens (cropped horizontally) for sight translation of an English sentence into Czech given a related (Figure 2a) or an unrelated (Figure 2b) image.

The original experiment design used two sentence types (unambiguous and ambiguous) with three image stimuli types (related, unrelated and no image) in a within-subjects design, i.e., every participant was exposed to all conditions (but never on the same stimulus). This resulted in the following six configurations:

- **UR** (unambiguous sentence + related image)
- **UU** (unambiguous sentence + unrelated image)
- **UN** (unambiguous sentence + no image)
- **AR** (ambiguous sentence + related image)
- **AU** (ambiguous sentence + unrelated image)
- **AN** (ambiguous sentence + no image)

The related images (congruent stimuli) matched the content of the text. The unrelated images (incongruent stimuli) were not relevant to the text. The "no image" condition (referred to as neutral stimuli) in the original experiment served as a control condition that consisted of an image with white background and a text saying *"No visual clue for this case"*. Apart from these configurations, there was a contrastive pair in each probe (set of stimuli) labelled as: **AR** (ambiguous sentence + related image): *"A person in **a** blue ski suit is racing two girls on skis."* and **UR** (unambiguous sentence + related image): *"A person in **her** blue ski suit is racing two girls on skis."*

Our study is centred around two main research questions which we answer using an analysis of previously collected data in image-supported human translation:

- **RQ1:** Do people translate ambiguous sentences differently than unambiguous sentences?
- **RQ1:** How does visual information impact the translation process?

3 Reaction times analysis

The average reaction times for each stage are shown in Figure 3. In Stage 3, when the image was first presented, the difference in observed reaction times across conditions is very prominent. The highest time is spent with the related images, followed by the unrelated images and finally the "no image" case. The same trend, with a less clear distinction, is repeated in Stage 4.

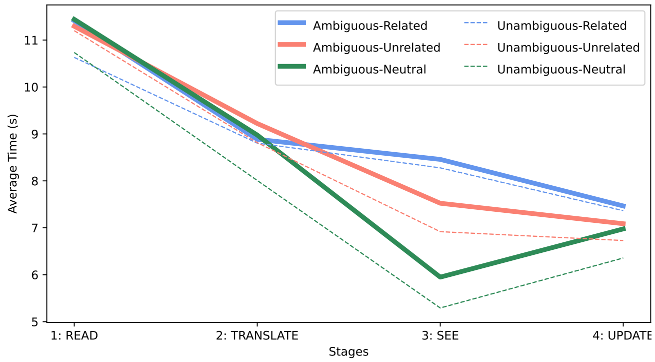


Figure 3: Average stage-wise reaction times across all conditions.

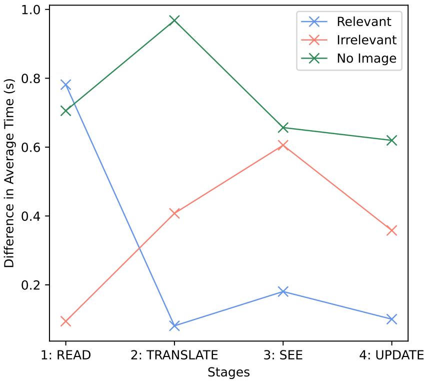


Figure 4: Difference in durations (ambiguous – unambiguous) across stages and congruency.

Stages 1 & 2. The comparison of reaction times for Stages 1 (READ) and 2 (TRANSLATE) is akin to the comparison of reading vs. production times for sentences. The results are shown in Table 1 and visualized in Figure 3. The values for Stages 1 & 2 show very little variance: 0.12 and 0.17, respectively. It is clear that participants spent more time reading than translating.

The difference between ambiguous sentences and unambiguous sentences was significantly higher during translation ($t = 1.928, p = 0.054$) than during reading ($t = 1.372, p = 0.170$). To better visualize the condition-wise differences across the two classes of sentences (ambiguous and unambiguous), we calculate the difference (ambiguous vs. unambiguous) in the time taken for each category (Figure 4). The positive differences between them show that the ambiguous sentences took more time to be processed than the unambiguous ones (**RQ1**).

Stage 3. In the SEE stage, the participants were shown the image stimuli. Across all sentences, the congruent stimuli (relevant image) took more time ($t = 2.710, p = 0.007$) than the *incongruent* (irrelevant image): variances 1.61 and 2.24 for the ambiguous and unambiguous case, respectively (**RQ2**). Also, from Figure 4, it is apparent that across all images, ambiguous sentences took more time than unambiguous sentences (**RQ1**). The greatest difference between ambiguous sentences and unambiguous sentences was noticed for the case with no images ($t = 1.746, p = 0.082$), followed by the case with an unrelated image ($t = 1.340, p = 0.181$), followed by the case with related images ($t = 0.573, p = 0.567$).

In this context, it should be noted that the setup with three image stimuli conditions in the original experiment can be thought of as a variant of the classic Stroop task (Stroop, 1935) that involved naming of coloured words (MacLeod, 1992). It was observed that incongruent color and text took longer to process. In other words, when naming the font colour, *RED* printed in red would take less time than *RED* printed in blue. This can be generalized to the effect of discrepancy (incongruency) between information in a stimulus.

In this experiment, the stimuli in categories Condition 1 (AR, UR), Condition 2 (AU, UU) and Condition 3 (AN, UN) correspond to congruent, incongruent and neutral stimuli respectively. From Figure 4 we see that in both Stage 3 (SEE) and Stage 4 (UPDATE), the difference in reaction times between ambiguous sentences and unambiguous sentences was greater for the incongruent (and neutral) visual stimuli in comparison to congruent stimuli, therefore confirming the Stroop effect of visual information (**RQ2**).

Condition	Stage 1	Stage 2	Stage 3	Stage 4
AR	11.41	23.97	8.17	7.47
AU	11.29	22.99	7.52	7.09
AN	11.44	21.63	5.95	6.98
UR	10.63	24.18	8.28	7.37
UU	11.20	21.90	6.92	6.73
UN	10.73	19.67	5.29	6.36

Table 1: Average durations of all stages and conditions.

Condition	t	p
AR-AU	1.441	0.150
AR-AN	4.085	<0.001
AU-AN	3.318	0.001
UR-UU	2.725	0.007
UR-UN	7.046	<0.001
UU-UN	4.618	<0.001

Table 2: T-Test results of case-wise comparisons in times in Stage 3. Value p is probability of means being the same and not different.

Stages 2 & 4. Both the `TRANSLATE` and `UPDATE` stages required participants to translate the text. While participants were shown only the text stimuli in Stage 2, they also had access to the visual clue (relevant or not) when they were asked to either repeat their translation or update it in Stage 4. Table 1 shows that the time taken to complete Stage 4 was always lower ($t = 10.305, p < 0.001$) than the time it took to complete Stage 2 across both categories of sentences and across all conditions. Participants also spent longer time translating the sentences with related images for both classes of sentences (ambiguous and unambiguous) in Stage 4. This leads us to claim that there was significant cognitive effort required to integrate the relevant visual information into the translation that they already had in their memory. For the incongruent stimuli, participants chose to disregard the visual information, which hence resulted in shorter timings in Stage 4.

Another interesting observation concerns the reaction times corresponding to the neutral visual stimulus in Stages 3 and 4. While the reaction times for the congruent and incongruent stimuli decrease, the reaction times for neutral stimuli show an increase with both classes of textual stimuli. Also, the difference in reaction times between ambiguous and unambiguous sentences does not show the same rate of reduction as observed with the other image stimuli categories.

4 Discussion and Conclusion

From the response-time based metrics, it can be easily seen that ‘reading‘ ambiguous sentences took slightly more time than unambiguous sentences. Interestingly, it took almost the same time to translate both types of sentences. We also observe that participants took significantly more time to look at the congruent visual stimuli in comparison to the other two categories. However, they also took almost the same time during the final translation in Stage 4 across all conditions. In fact, the final translation (Stage 4) took less time than the initial translation (Stage 2). In other words, production time was almost the same across all categories and decreased with time.

We also observe that the presence of incongruent and neutral image stimuli makes it harder to process ambiguous sentences (as reflected in the difference of reaction times) than unambiguous sentences. And hence, with the experiment setup, we notice effects associated with the classic Stroop effect. Finally we observe that, in the absence of **any** image stimulus (neutral), the difference in reaction times between ambiguous and unambiguous sentences is the greatest. This leads us to conclude that the inclusion of image modality actually helped people in the process of disambiguation and eventual translation of ambiguous sentences. However, the visual incongruence impacted the processing of ambiguous sentences more than unambiguous sentences. In other words, Stroop effect was more prevalent in ambiguous sentences. But still, even the interference caused by the Stroop effect helped in translating ambiguous sentences faster than in situations with neutral image stimulus (image stimulus with no usable visual clues).

Future work. Future work should analyze also the eye-tracking data and EEG data collected as part of the corpus and investigating any relationship between them. Importantly, the findings should be put in context of translator’s work (both professional and layman) to further facilitate research in human-computer-interaction in this domain.

5 Acknowledgement

This project has received funding from the 19-26934X (NEUREM3) grant of the Czech Science Foundation. The work has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2018101 LINDAT/CLARIAH-CZ.

References

Paul D Allopenna, James S Magnuson, and Michael K Tanenhaus. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, 38(4):419–439, 1998.

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

Daniel Gile. *Basic concepts and models for interpreter and translator training*, volume 8. John Benjamins Publishing, 2009.

Falk Huettig, Joost Rommers, and Antje S Meyer. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, 137(2):151–171, 2011.

Kristian Tangsgaard Hvelplund. *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. Frederiksberg: Copenhagen Business School (CBS), 2011.

Marcel Adam Just and Patricia A Carpenter. Eye fixations and cognitive processes. *Cognitive psychology*, 8 (4):441–480, 1976.

Colin M MacLeod. The stroop task: The" gold standard" of attentional measures. *Journal of Experimental Psychology: General*, 121(1):12, 1992.

J Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6): 643, 1935.

Sonja Tirkkonen-Condit. A think-aloud protocol study. In *Learning, Keeping and Using Language: Selected papers from the Eighth World Congress of Applied Linguistics, Sydney, 16 21 August 1987*, volume 2, pp. 381. John Benjamins Publishing, 1990.

Matthew J Traxler. Introduction to psycholinguistics: Understanding language science. 2011.

A Appendix

Condition	Stage 1	Stage 2	Stage 3	Stage 4
AR	11.41	8.89	8.46	7.47
AU	11.29	9.22	7.52	7.09
AN	11.44	8.98	5.95	6.98
UR	10.63	8.80	8.28	7.37
UU	11.20	8.81	6.92	6.73
UN	10.73	8.01	5.29	6.36

Table 3: Average response times across stages and conditions.