Tools Impact on the Quality of Annotations for Chat Untangling

Jhonny Cerezo¹, Alexandre Bergel¹, Felipe Bravo-Marquez¹ 2
¹Department of Computer Science, University of Chile
²Millennium Institute for Foundational Research on Data, IMFD-Chile
jcerezo@dcc.uchile.cl, abergel@dcc.uchile.cl, fbravo@dcc.uchile.cl

Abstract
The quality of the annotated data directly influences in the success of supervised NLP models. However, creating annotated datasets is often time-consuming and expensive. Although the annotation tool takes an important role, we know little about how it influences annotation quality. We compare the quality of annotations for the task of chat-untangling made by non-experts annotators using two different tools. The first is SLATE, an existing command-line based tool, and the second is Parlay, a new tool we developed that integrates mouse interaction and visual links. Our experimental results indicate that, while both tools perform similarly in terms of annotation quality, Parlay offers a significantly better user experience.

1 Introduction
Human linguistic annotation is essential for many natural language processing tasks. However, the construction of these datasets is extremely expensive, both in terms of annotator hours and financial cost (Snow et al., 2008). We know that the performance of many natural language processing tasks is limited by the quantity and quality of the data available to them (Banko and Brill, 2001; Snow et al., 2008). Many of the studies focus on the number of annotators and their experience in improving annotation quality. However, little has been studied about the role annotation tools play in producing quality data.

We study the effect of annotation tools in chat-untangling task. Chat-tangling occurs when simultaneous conversations arise in chat with multiple participants (Elser and Charniak, 2008). The goal of chat-untangling is to identify the conversations in a chat-thread. In order to perform this task, annotators must maintain a complex mental representation of the ongoing conversations. Hence, a tool should minimize the task load and facilitate the annotation process. To the best of our knowledge, there is only one public dataset sufficiently large for training modern NLP architectures for this task published by Kummerfeld et al. (2019). This dataset was annotated using SLATE (Kummerfeld, 2019), a terminal based annotation tool. In SLATE, interactions are marked by keyboard commands, messages are shown as raw-text, and annotations are presented by color coding. We argue that these characteristics may influence in quality of annotations for non-experts.

This paper presents Parlay, a new annotation tool for the chat-untangling task. The main difference between Parlay and SLATE is that it integrates mouse interaction and visual links into the annotation representation. We conducted a controlled experiment with 12 non-expert participants, in which each participant was first introduced to the annotation task and then asked to use each tool to annotate 100 messages. The data to be annotated are selected from gold standard adjudicated dataset presented by Kummerfeld et al. (2019). Next, each annotation tool is evaluated in terms of usability (SUS), task load (NASA/TLX), performance and annotation time. We define annotator performance by the quality of its annotations calculated by comparing them to the gold standard (expert) annotations (Kummerfeld et al., 2019) using Cohen’s Kappa coefficient (Cohen, 1960).

2 Background
Chat-untangling. Chat-untangling is an NLP task that aims to find existing conversations within a chat-log with two or more participants (Adams and Martell, 2008; Holmer, 2008; Kummerfeld, 2019; Shen et al., 2006; Elser and Charniak, 2010, 2008). Conversations are represented by a connected graph in which the vertices are messages and the edges are relationships between them (e.g.,
an answer to a question) (Adams and Martell, 2008; Wang et al., 2008; Kummerfeld et al., 2019; Holmer, 2008).

**Annotation process.** Manual text annotation is the process of assigning some tags to the whole text or to fragments of it. A corpus is a collection of texts on a particular topic. Annotators tag parts of the text in the corpus with labels that represent the structure or semantics of interest. The annotated data then goes through a curation process in which a curator manually resolves discrepancies and inconsistencies. Curation is primarily performed to ensure data quality (Grosman et al., 2020). One important step in curation is adjudication (Ide and Pustejovsky, 2017). In this step the curator merges the annotations from different annotators though agreement and resolves discrepancies to produce a gold standard annotation. The annotation process concludes with the release of the curated corpus to the community.

**Annotation quality.** To a large extent, the final quality of the annotation process depends on how human error is reduced and how discrepancies are consolidated without biasing the annotator’s judgment (Chau et al., 2020). It is demonstrated that the annotation quality is related to the annotator expertise (Snow et al., 2008; Burnap et al., 2015), number of annotators (Snow et al., 2008) and the process of learning the annotation task (Teruel et al., 2018). On the other hand, annotation tools are used to assist the annotation process (e.g., file loading, user annotation interaction, data curation) (Yimam et al., 2013; Grosman et al., 2020; Kummerfeld, 2019; Yordanova et al., 2018).

Although there is evidence that user-friendly annotation tools can benefit the annotation process and reduce the annotation time (Yimam et al., 2013; Kummerfeld, 2019; Grosman et al., 2020), little is known how can they can influence to the annotator’s performance. To the best of our knowledge, Grosman et al. (2020) is the only study that report changes in the inter-agreement evaluating different tools for annotation. However, there is no further understanding how this occurs at annotation time.

## 3 Methodology

In this section we introduce Parlay, a new chat-untangling annotation tool, and SLATE as baseline. Finally, we provide the evaluation methodology, criteria and our hypotheses during this study.

---

1Pharo is a pure object-oriented programming language and offers a flexible programming environment (Bergel et al., 2013).

2GT, which stands for Glamorous Toolkit, is a “moldable programming environment”. https://gtoolkit.com.
This study aims to assess the usability, task load, performance and annotation time of the SLATE and Parlay annotation tools. In order to create fair conditions for evaluating both tools, we conducted a controlled experiment.

Participants. We gathered 12 participants among university graduates (4), master students (4), master graduate (1), doctoral students (1), postdoctoral (2). None of the participants had a background in text annotation. All of them were experienced in the use of Linux, and none of them were native English speakers, although all of them declared to have a proficient level of English.

<table>
<thead>
<tr>
<th>Tool</th>
<th>SLATE</th>
<th>Parlay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select message</td>
<td>Move messages by pressing arrow key one line at the time to select a message.</td>
<td>Scroll down or up, then left-click on a message’s green button at the left part of each message to select it.</td>
</tr>
<tr>
<td>Create annotation</td>
<td>Move to select both messages then press key “D” to link them.</td>
<td>Select the first message, then scroll down/up and left-click in a message’s grey button at the left of the message. It eliminates only the link between those messages.</td>
</tr>
<tr>
<td>Remove annotation</td>
<td>Select a message, then press key “U”. It eliminates all links from the selected messages.</td>
<td>i) Select a message, then press key “B”. It eliminates all links from the selected message. ii) Select the first message, then scroll down/up and left-click on the grey button at the left of the message. It eliminates only the link between those messages.</td>
</tr>
<tr>
<td>Validate annotation</td>
<td>Select a message that has already been annotated. This action will change the color of messages that are linked to the selected message.</td>
<td>i) Select a message that has already been annotated. ii) Hover over the left green button of a message. Both options will highlight the color of links and messages that are linked to the selected message.</td>
</tr>
</tbody>
</table>

Table 1: Functional differences between Parlay and SLATE.

<table>
<thead>
<tr>
<th>Setting</th>
<th>First session</th>
<th>Second session</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>2015-08-10</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
<td>2015-08-10</td>
</tr>
</tbody>
</table>

Table 2: Experiment settings.

SLATE shows a change in the color of messages. Second, Parlay allows one to use the mouse to define annotations, whereas in SLATE the annotations are made by keyboard commands. At first, these two differences may look superficial. However, it is reasonable to think that they may significantly impact the user experience.

3.2 Evaluation Design

This study aims to assess the usability, task load, performance and annotation time of the SLATE and Parlay annotation tools. In order to create fair conditions for evaluating both tools, we conducted a controlled experiment.

Data. We selected two files from Kummerfeld et al. (2019) adjudicated dataset that was developed to calculate inter-agreement. The adjudicated file is considered as the gold standard annotation to which we compare our non-expert participants’ annotations. Then, we selected a pair of files with high similarity by two criteria: a) the agreement between annotators against the adjudicated file, and b) the number of annotated conversations.

Design. The controlled experiment was divided in three sessions, one introductory and two evaluation sessions. In the introductory session each participant had to i) answer a demographic questionnaire, ii) read an introductory document to the chat-untangling task, iii) read the annotation guidelines developed by Kummerfeld et al. (2019), and iv) answer a chat-untangling annotation exercise. The exercise consists of annotating three 5-messages long conversations without the help of Parlay or SLATE. The moderator then discusses the participants’ annotations. The annotation exercise ensures the participants’ maximum understanding of the chat-untangling task.

In the consequent evaluation sessions the participants had to i) annotate a file using one of the two tools, and ii) answer questionnaires. The tool for each session is randomly selected, whereas the annotated file is always the same for each evaluation session. We name each file-tool combination as an experiment setting. Such that we get two experiment settings, A and B. Where each experiment setting considers 6 participants and the participants use both tools in turn for the annotation task. For instance, a participant in setting A uses SLATE in the first evaluation session and Parlay in the second. In Table 2 we present the two experiment settings in detail.

Validation Criteria. Parlay and SLATE are evaluated in terms of usability, task load, annotation time and performance. The usability is measured by the System Usability Scale (SUS) (Brooke, 1996). The task load is measured using NASA/TLX (Hart and Staveland, 1988). The annotation time is determined from the annotation-logs. Lastly, the performance or quality of the annotation is measured by calculating the Cohen’s Kappa (Cohen, 1960) for agreement between each participant’s annotations and the expert adjudicated annotations (gold standard) presented in Kummerfeld et al. (2019) work. We say that the higher the Kappa score (κ) the better the performance.
Hypotheses. To analyze this study results, we performed some statistical tests to verify the following alternative hypotheses:

- **H1**: The participants’ performance is affected by the annotation tool.
- **H2**: The participants’ performance is affected by the experience gained in the annotation sessions.

We evaluated the significance of the performance *kappa* using the statistical t-test (De Winter, 2013) for small sample size (De Winter, 2013). We reject the null hypotheses if $p$-value $< 0.05$.

### 4 Results

In this section we present the results of our methodology of evaluation for i) the task load, ii) usability, iii) performance and iv) annotation time.

**Task load.** Table 3 shows the averaged results of participants’ answers for the NASA/TLX dimensions. On average mental demand, physical demand, effort and frustration are lower in Parlay, whereas SLATE shows less temporal demand, although the difference is slight. Lastly, the users report that they performed better using Parlay. Overall Parlay reports less task load with the exception of being slightly more temporal demanding.

**Usability** Table 4 shows the mean and standard deviation values of SUS scores with the two annotation tools. If we pay attention to questions regarding positive (i.e., Q1, Q3, Q5, Q7 and Q9) and negative (i.e., Q2, Q4, Q6, Q8 and Q10) aspects of usability we find that: i) SLATE has similar average scores in positive and negative; and ii) Parlay rates higher to questions regarding the positive and lower in negative aspects of usability. A closer look at the ten component SUS scores we can see that Parlay is perceived more usable by the participants. This suggests that Parlay achieved better usability compared to SLATE.

<table>
<thead>
<tr>
<th>Mental</th>
<th>SLATE</th>
<th>Mean</th>
<th>SD</th>
<th>Parlay</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>6.167</td>
<td>2.167</td>
<td>5.833</td>
<td>1.992</td>
<td>2.216</td>
</tr>
<tr>
<td>Physical</td>
<td></td>
<td>3.417</td>
<td>1.514</td>
<td>3.417</td>
<td>2.429</td>
<td>2.216</td>
</tr>
<tr>
<td>Temporal</td>
<td></td>
<td>4.417</td>
<td>2.429</td>
<td>4.417</td>
<td>2.429</td>
<td>2.216</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td>5.5</td>
<td>2.908</td>
<td>5.5</td>
<td>2.908</td>
<td>2.429</td>
</tr>
<tr>
<td>Effort</td>
<td></td>
<td>5.5</td>
<td>2.908</td>
<td>5.5</td>
<td>2.908</td>
<td>2.216</td>
</tr>
<tr>
<td>Frustration</td>
<td></td>
<td>5.667</td>
<td>1.969</td>
<td>5.667</td>
<td>1.969</td>
<td>2.216</td>
</tr>
</tbody>
</table>

Table 3: Comparison of average and standard deviation of NASA/TLX scores between Parlay and SLATE.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Sample</th>
<th>Time (min)</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parlay</td>
<td>12</td>
<td>48.5</td>
<td>21.211</td>
</tr>
<tr>
<td>SLATE</td>
<td>12</td>
<td>41.083</td>
<td>16.714</td>
</tr>
</tbody>
</table>

Table 4: Comparison of average and standard deviation of SUS scores between Parlay and SLATE.

**Annotation time.** SLATE reported less annotation time in minutes in our participants (Mean=41.083, SD=16.714) compared to Parlay (Mean=48.5, SD=21.211). Lastly the second evaluation session reported less annotation time in our participants (Mean=39.417, SD=21.249) compared to Parlay (Mean=50.167, SD=15.643).

**Performance.** We assess our hypotheses by calculating the performance ($\kappa$) by two criteria: i) the tools used in the annotation task (H1) and ii) the session in which the participants annotate (H2). For H1 there was a non-significant difference in the scores for Parlay and SLATE with $p$-value=0.1583. For H2 there was a significant difference in the scores for the first evaluation session and second evaluation session with $p$-value=0.007636. Therefore, the quality of annotations increases as annotators gain experience. Finally, the annotation tool used does not influence the quality of annotations.

### 5 Conclusion

In this paper we evaluate the influence of the annotation tool for non-expert users in terms of data quality and user experience in the chat-untangling task. To achieve our purpose we introduce a new annotation tool named Parlay and establish SLATE as our baseline. Subsequently, we conducted a controlled experiment with 12 non-expert annotators in which each participant annotated 100 messages...
on each tool in turn. Each tool was evaluated under i) usability, ii) task load, iii) annotation time and iv) annotation performance. Lastly, we establish that the quality of annotations is measured by calculating the agreement ($\kappa$) between participants’ annotations and the gold annotated data (Kummerfeld et al., 2019).

The results indicate that the tools did not show significant differences in the annotators’ performance outcome on the chat-untangling task. On the other hand, participants showed better performance in the second session, presumably due to a gain in experience. The annotation time is also lower in SLATE. Complementary to these results, we also report that Parlay scored better in usability and task load. Where participants highlighted that i) link representation between annotated messages and ii) mouse interaction where the main characteristics that made Parlay. In conclusion, Parlay offers a better user experience while achieving comparable annotation performance.

The study of annotation quality is an important issue for future development of NLP models. Tools remain as an important factor in the development of high-quality training datasets for NLP tasks (Grosman et al., 2020). Despite the results presented in this study, further work is required to get a thorough understanding of how the annotation tool affects the quality of chat-untangling data.

This study contributes to the area by introducing Parlay, a new tool for the chat-untangling task. We believe that an important contribution of our work is the methodology we propose, which allows us to compare annotation tools according to both data quality and user experience. For future work we aim to study the inter-annotator agreement of each tool. As well as, the discrepancies between expert and non-expert annotations.

Acknowledgements

We thank Renato Cerro for his review and editing and we thank 3 “anonymous” reviewers for their so-called insights. Felipe Bravo-Marquez was funded by ANID FONDECYT grant 11200290, U-Inicia VID Project U1-004/20 and ANID - Millennium Science Initiative Program - Code ICN17_002. Alexandre Bergel thanks Lam Research and the ANID FONDECYT Regular 1200067 for partially sponsoring the work presented in this paper.

References


Torsten Holmer. 2008. Discourse structure analysis of chat communication. Language@ Internet, 5(10).


Kristina Yordanova, Adeline Paiement, Max Schröder, Emma Tonkin, Przemyslaw Woźnowski, Carl Magnus Olsson, Joseph Rafferty, and Timo Sztyler. 2018. Challenges in annotation of user data for ubiquitous systems: Results from the 1st arduous workshop.