The Effect of Annotation Scheme Decisions on Parsing Learner Data

Marwa Ragheb and Markus Dickinson

Department of Linguistics
Indiana University
E-mail: \{mragheb,md7\}@indiana.edu

Abstract

We present a study on the dependency parsing of second language learner data, focusing less on the parsing techniques and more on the effect of the linguistic distinctions made in the data. In particular, we examine syntactic annotation that relies more on morphological form than on meaning. We see the effect of particular linguistic decisions by: 1) converting and transforming a training corpus with a similar annotation scheme, with transformations occurring either before or after parsing; 2) inputting different kinds of part-of-speech (POS) information; and 3) analyzing the output. While we see a general favoritism for parsing with more local dependency relations, this seems to be less the case for parsing the data of lower-level learners.

1 Introduction

An increasingly popular topic in parsing is to parse non-canonical data [11], including the data of second language learners [4, 10, 13, 18]. In this paper, we add to this growing body of work, focusing less on the parsing techniques and more on the effect of the linguistic distinctions made in the data. In particular, we examine syntactic annotation (for English) that makes different assumptions than in previous work, relying more on morphological form than on context or meaning. We will see the effect of particular linguistic decisions by: 1) converting and transforming a training corpus with a similar annotation scheme (section 2), with transformations occurring either before or after parsing (section 3.1); 2) inputting different kinds of part-of-speech (POS) information (section 3.2); and 3) analyzing the output (section 3.3).

We work with a pilot version of the SALLE corpus [7, 19, 21], which has a fairly unique perspective. It is focused on morphologically-driven dependencies and prioritizes syntax—often to the exclusion of semantics—whereas other parsing
work has been more focused on connecting to the semantics of a sentence. Regardless of what one hopes to achieve with the annotation in the end (e.g., acquisition research [10, 12, 20], parsing to assist in error correction or feedback [26], etc.), noting differences in the parsing results is important to improve the parsing more generally, to see the influences of different kinds of information. As one example in our data, there are two POS tags for every position, to reflect different kinds of evidence [see also 5]. Parsing this corpus helps in the process of teasing apart what makes learner language difficult and where parsing can be improved.

2 Data

While we wish to parse one set of data (section 2.1), the training data with the closest available annotation scheme has significant differences (section 2.2). We describe the data sets, followed by how we prepared the training data to be compatible with testing (section 2.3), highlighting differences which can affect parsing.

2.1 Testing Data: Target Annotation

The testing data consists of 25 learner essays (491 sentences, 7381 tokens) gold-annotated with the SALLE scheme [7, 21].1 The essays can be grouped into three different levels—beginner, intermediate and advanced—based on placement scores (1 (low) to 7 (high)) assigned by expert raters for the Intensive English Program at Indiana University. These essays were prompt-based, timed placement exams, and they represent a variety of first languages (L1s) [see 19, for more on the essays].

The annotation scheme [22] annotates lemmas, two POS tags reflecting potentially diverging morphosyntactic and distributional evidence (POS$_{M}$, POS$_{D}$), and dependency relations that are based on the morphosyntactic POS tags. The scheme also encodes subcategorization information. An example annotation is in figure 1, where POS tags are a simplified version of the SUSANNE scheme [25] and the syntactic relations are a modified and expanded set of CHILDES (see section 2.2).

The different layers provide access to innovative learner constructions, through the presence of mismatches between layers. In this example, for instance, the POS$_{M}$ of VV0 (base form verb) for decide conflicts with the POS$_{D}$ of VV (underspecified verbal form), and job subcategorizes for a determiner (<DET>), yet finds two of them. Note that, because both which and my are morphologically valid determiners, both are annotated as DET. That is, decisions are biased towards morphological information [7, 21], a point which affects parsing (see section 3.3).

Another point that may affect the parsing results is the use of underspecified labels. SALLE makes use of underspecification when there is not enough linguistic

---

1Though not huge, the size of the gold-annotated testing data is comparable to other studies directly parsing learner language [e.g., 10] and suits the purpose of investigating the effect of linguistic decisions. We are currently investigating ways to expand the gold annotated corpus by incorporating more semi-automatic steps into the otherwise manual process.
evidence to determine a full label (as is the case with the POSD for decide above) or to select a certain label over another (in either the POS or dependency fields). Underspecified labels (including a complete non-specification, notated as _) are not present in the training data, a point discussed further in section 3.3.

As a side point, since the scheme aims at directly annotating language, without imposing target hypotheses or corrected versions of the learner text [20], it does not encode learner “errors”. For analysis, many of the non-canonical structures in what the learners wrote can nonetheless be arrived at by examining mismatches between layers, for example, VV0 vs. VV for decide and the subcategorization vs. dependency realization for job in figure 1.

2.2 Training Data: Source Annotation

The training data is a subset of the CHILDES corpus of first language acquisition [14], from the US English data, downloaded in March 2012 [19]. This data is appropriate for training because, as a corpus of (L1) acquisitional data, it is likely to include some patterns L2 learners also use, such as overgeneralization of certain inflectional morphemes (e.g., hitted, goed) and similar developmental sequences (e.g., for negation) [cf., e.g., 9]. More importantly, the syntactic annotation scheme [24] formed the basis of SALLE’s scheme [7, 19, 21]. The SALLE labels are based on the CHILDES labels, but there are more labels and some constructions are modified; still, the annotation schemes are relatively compatible.

This data presents a variety of challenges: 1) it is not in a format appropriate for modern trainable parsing systems, and it is occasionally inconsistent in its formatting (section 2.3.1); 2) the schemes have to be harmonized (section 2.3.2); and 3) it is for spoken data, not written data, a fact that partly accounts for lower parsing results (section 3.3). We have automatically removed some of the training data (verbless and anomalous [8] sentences), due to noise; space precludes a full discussion [see 19]. This results in 102,733 sentences and 727,297 tokens.
2.3 Data Preparation

2.3.1 Conversion

The first step to prepare the data is to convert the data from CHAT format [14] to CoNLL-X format [3], to make it appropriate for parsing. The required information for parsing is on three different tiers: the speaker tier (represented in %flo in the top of figure 2), the morphology tier (%mor) and the syntactic tier (%xgra or %gra). We want the representation in CoNLL-X format, as seen in the bottom of figure 2.

%flo: what’s Mamma’s name ?
%mor: pro:wh|what˜aux|be&3S n:prop|Mamma˜poss|s n|name ?
%xgra: 1|2|PRED 2|0|ROOT 3|5|MOD 4|5|MOD 5|2|SUBJ 6|2|PUNCT

Figure 2: CHAT (top) and CoNLL-X (bottom) formats for what’s Mamma’s name?

When the information is properly aligned across all three tiers, the conversion process is straightforward. There are cases with misalignments between two of the three layers, however, requiring additional steps [19]. For example, in figure 2 the tilde is used to mark single %mor tokens which correspond to two %xgra (syntactic) units. In the same essay, a later instance of mama’s contains a tilde in the %mor layer, but only corresponds to only one %xgra unit. Our conversion script covers the majority pattern for special cases (tildes, compound nouns, punctuation), and we corrected the other cases by hand, generally detectable when the CHAT layers have differing numbers of units.

One last step in the conversion process is to change the POS tags from the ones used in CHILDES to the ones used in SALLE. We automatically tag the CHILDES data with TnT [2], using the pre-built SUSANNE model, and then employ a simple mapping scheme to the SALLE labels (a subset of SUSANNE). Although this introduces some noise in training, this does not affect our current focus of determining which transformation model or which input POS results in better performance (see section 3). Additionally, the CHILDES corpus is itself mostly automatically POS-tagged [14, p.147-148], and previous experiments [19, ch. 6] using the CHILDES POS tagset showed similar trends as to what is reported in section 3.
2.3.2 Transformations

There are three main syntactic constructions that CHILDES and SALLE analyze differently: auxiliary verbs, coordination, and possessives. The data thus needs to be transformed to align with SALLE, either before training or after parsing (see section 3). We focus here on the first two areas of difference, as possessives involve a simple swapping of heads, with little effect on surrounding constructions.

The first difference, stemming from SALLE’s prioritization of syntax over semantics, is the *auxiliary-as-head* analysis that SALLE adopts, whereas CHILDES considers the main verb as the head of a verbal unit (e.g., *have run*). The transformation makes the first auxiliary the head of any verbal chain and then, heuristically following SALLE attachments, keeps the following arguments as dependents of the content-ful verb, but preceding arguments as dependents of the auxiliary.

As for the second difference: while CHILDES analyzes the conjunction as the head of the coordination phrase, SALLE adopts a right-branching analysis, since this accounts better for learner innovations [6]. The transformation process likewise switches heads here, but it also has to account for the interaction between coordination and auxiliary transformations [19]. Namely, auxiliary transformations take place before coordinations, so that verbal heads are properly coordinated. Figure 3 shows the difference in coordination analyses between the schemes.

![Figure 3: Coordination in CHILDES (top) and SALLE (bottom)](image)

3 Experiments

3.1 Transformations

Given the transformations from the CHILDES annotation scheme to the SALLE one, we can ask whether it is preferable to train a parser on the CHILDES scheme and then transform the resulting output (*post*) or to first transform the training data, to learn the SALLE model directly (*pre*). As SALLE generally posits more local distinctions (section 2.3.2), and based on preliminary experiments [19] we ex-
pect to see better results with the pre model [cf. 16]. Note that the motivation for SALLE’s scheme was not to make parsing easier, but to be more purely syntactic, which happens to be more local in English.

### 3.2 POS Information

The SALLE data provides two part-of-speech (POS) tags, one more driven by morphology and one by syntactic distribution. The annotation scheme used for parsing itself is more morphologically-based, but syntax by its nature has to rely on contextual information and parsing results can vary based on the POS input [e.g., 15]. Thus, we can ask which POS information works best as input to the parser: morphological ($POS_M$), distributional ($POS_D$), or both ($POS_{Both}$)?

### 3.3 Results

The experiments use MaltParser [17], optimizing the parser settings with help of MaltOptimizer [1]. This chooses the stackproj algorithm based on the nature of the training data. Evaluation is performed using the eval.pl script.\(^2\)

The results for the six models (2 transformations $\times$ 3 POS inputs) are given in table 1. We can immediately draw two conclusions: 1) the $POS_M$ models (top row) are consistently the best (albeit, slightly); and 2) the pre models are consistently better than the post ones. This supports our hypotheses: morphologically-based POS information seems better for parsing with this scheme, and more local (i.e., more adjacent) syntactic relations are preferred over less local ones.

<table>
<thead>
<tr>
<th></th>
<th>pre POS</th>
<th>post POS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LAS</td>
<td>UAS</td>
</tr>
<tr>
<td>$POS_M$</td>
<td>62.8%</td>
<td>74.3%</td>
</tr>
<tr>
<td>$POS_D$</td>
<td>62.7%</td>
<td>74.2%</td>
</tr>
<tr>
<td>$POS_{Both}$</td>
<td>62.7%</td>
<td>74.2%</td>
</tr>
</tbody>
</table>

Table 1: Overall results for the six different models

**Individual results** The results are more complicated when examining individual files, which show great variability, as partly illustrated in table 2. For example, in this table, comparing the pre+$POS_M$ model to the post+$POS_M$ model, the values range from ones where the post+$POS_M$ model has a 4.9% better LAS than the pre+$POS_M$ model (opposite of the overall trend) to ones where the pre+$POS_M$ is 5.8% better than the post+$POS_M$ model. Similarly, though not in the table, the pre+$POS_D$ model varies from having 4.5% worse LAS than the pre+$POS_M$ model to having 2.4% better.

\(^2\)http://ilk.uvt.nl/conll/software/eval.pl
<table>
<thead>
<tr>
<th>Essay</th>
<th>Level</th>
<th>Origin</th>
<th># of Sents</th>
<th># of Words</th>
<th>W/S</th>
<th>Pre LAS</th>
<th>Post LAS</th>
<th>Diff. LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>1</td>
<td>Taj.</td>
<td>2</td>
<td>45</td>
<td>22.5</td>
<td>41.5</td>
<td>46.3</td>
<td>-4.9</td>
</tr>
<tr>
<td>064</td>
<td>3</td>
<td>Kor.</td>
<td>27</td>
<td>261</td>
<td>9.7</td>
<td>60.0</td>
<td>63.7</td>
<td>-3.7</td>
</tr>
<tr>
<td>285</td>
<td>1</td>
<td>Saudi</td>
<td>8</td>
<td>96</td>
<td>12.0</td>
<td>65.5</td>
<td>67.8</td>
<td>-2.3</td>
</tr>
<tr>
<td>017</td>
<td>5</td>
<td>Afg.</td>
<td>34</td>
<td>472</td>
<td>17.8</td>
<td>49.7</td>
<td>51.6</td>
<td>-1.9</td>
</tr>
<tr>
<td>204</td>
<td>4</td>
<td>Thai</td>
<td>10</td>
<td>178</td>
<td>13.9</td>
<td>69.7</td>
<td>71.6</td>
<td>-1.9</td>
</tr>
<tr>
<td>267</td>
<td>3</td>
<td>Saudi</td>
<td>21</td>
<td>220</td>
<td>10.5</td>
<td>65.3</td>
<td>66.8</td>
<td>-1.5</td>
</tr>
<tr>
<td>097</td>
<td>4</td>
<td>Kor.</td>
<td>19</td>
<td>225</td>
<td>11.8</td>
<td>72.8</td>
<td>74.3</td>
<td>-1.5</td>
</tr>
<tr>
<td>021</td>
<td>5</td>
<td>Afg.</td>
<td>12</td>
<td>254</td>
<td>21.2</td>
<td>53.5</td>
<td>54.8</td>
<td>-1.3</td>
</tr>
<tr>
<td>193</td>
<td>7</td>
<td>Thai</td>
<td>15</td>
<td>356</td>
<td>23.7</td>
<td>61.5</td>
<td>62.1</td>
<td>-0.6</td>
</tr>
<tr>
<td>018</td>
<td>6</td>
<td>Afg.</td>
<td>17</td>
<td>293</td>
<td>17.2</td>
<td>62.0</td>
<td>62.0</td>
<td>0.0</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>Saudi</td>
<td>21</td>
<td>261</td>
<td>12.4</td>
<td>59.6</td>
<td>58.6</td>
<td>0.9</td>
</tr>
<tr>
<td>046</td>
<td>5</td>
<td>Jap.</td>
<td>18</td>
<td>329</td>
<td>18.3</td>
<td>53.8</td>
<td>52.7</td>
<td>1.0</td>
</tr>
<tr>
<td>026</td>
<td>4</td>
<td>Chi.</td>
<td>40</td>
<td>462</td>
<td>11.6</td>
<td>61.4</td>
<td>60.1</td>
<td>1.3</td>
</tr>
<tr>
<td>033</td>
<td>7</td>
<td>Jap.</td>
<td>25</td>
<td>308</td>
<td>22.9</td>
<td>60.6</td>
<td>59.2</td>
<td>1.4</td>
</tr>
<tr>
<td>036</td>
<td>6</td>
<td>Jap.</td>
<td>17</td>
<td>390</td>
<td>12.3</td>
<td>65.3</td>
<td>63.4</td>
<td>1.9</td>
</tr>
<tr>
<td>040</td>
<td>5</td>
<td>Jap.</td>
<td>23</td>
<td>317</td>
<td>13.8</td>
<td>67.1</td>
<td>65.0</td>
<td>2.1</td>
</tr>
<tr>
<td>240</td>
<td>6</td>
<td>Hun.</td>
<td>26</td>
<td>377</td>
<td>14.5</td>
<td>67.0</td>
<td>64.6</td>
<td>2.3</td>
</tr>
<tr>
<td>034</td>
<td>6</td>
<td>Jap.</td>
<td>16</td>
<td>327</td>
<td>20.4</td>
<td>63.4</td>
<td>61.0</td>
<td>2.4</td>
</tr>
<tr>
<td>052</td>
<td>5</td>
<td>Kaz.</td>
<td>21</td>
<td>359</td>
<td>17.1</td>
<td>64.6</td>
<td>62.1</td>
<td>2.5</td>
</tr>
<tr>
<td>080</td>
<td>2</td>
<td>Kor.</td>
<td>18</td>
<td>174</td>
<td>9.7</td>
<td>61.9</td>
<td>59.2</td>
<td>2.7</td>
</tr>
<tr>
<td>041</td>
<td>7</td>
<td>Jap.</td>
<td>16</td>
<td>305</td>
<td>19.1</td>
<td>73.6</td>
<td>69.9</td>
<td>3.6</td>
</tr>
<tr>
<td>016</td>
<td>6</td>
<td>Afg.</td>
<td>33</td>
<td>353</td>
<td>26.0</td>
<td>60.2</td>
<td>56.0</td>
<td>4.3</td>
</tr>
<tr>
<td>129</td>
<td>3</td>
<td>Kor.</td>
<td>14</td>
<td>364</td>
<td>10.7</td>
<td>64.0</td>
<td>59.3</td>
<td>4.7</td>
</tr>
<tr>
<td>053</td>
<td>5</td>
<td>Kaz.</td>
<td>18</td>
<td>290</td>
<td>16.1</td>
<td>61.9</td>
<td>56.4</td>
<td>5.6</td>
</tr>
<tr>
<td>020</td>
<td>6</td>
<td>Afg.</td>
<td>20</td>
<td>365</td>
<td>18.3</td>
<td>60.1</td>
<td>54.2</td>
<td>5.9</td>
</tr>
<tr>
<td>-diff</td>
<td>3.7</td>
<td>n/a</td>
<td>16.4</td>
<td>234.1</td>
<td>15.9</td>
<td>59.9</td>
<td>62.1</td>
<td>-2.2</td>
</tr>
<tr>
<td>+diff</td>
<td>5.1</td>
<td>n/a</td>
<td>21.7</td>
<td>332.1</td>
<td>16.2</td>
<td>63.0</td>
<td>60.1</td>
<td>2.8</td>
</tr>
<tr>
<td>overall</td>
<td>4.6</td>
<td>n/a</td>
<td>19.6</td>
<td>295.2</td>
<td>16.1</td>
<td>61.8</td>
<td>60.9</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Table 2: LAS (in %) for POSM models, with Pre and Post transformations, organized by the difference in LAS (Pre − Post), and including number of words, sentences, and average words per sentence; (macro-)averaged values are presented for the different variables, grouped by essays which had a better Post model (-diff, i.e., averaging the first nine rows) or a better Pre model (+diff).

Focusing in on the effect of transformations before training (pre) or after parsing (post), as in table 2, we can observe a major trend, as outlined in the summary statistics at the bottom of the table. Scoring learners on a 1–7 level (see section 2.1), learners who are at a lower level (3.7 on average vs. 5.1) tend to have worse overall parsing performance (e.g., 41.5%/46.3% LAS for learner 201). Correspondingly,
the models which perform transformations after parsing \textit{(post)} tend to work better for such learners. We believe that this may stem from less canonical (and perhaps less local) structures in such data, allowing for noise in the transformation process to have a bigger effect on the results.

While higher average numbers of sentences and words may lead to more difficult parsing in the general case, this seems not to be the case with the learners in this data, at least for the pre+\textit{POS}_M model, where those with a better pre+\textit{POS}_M model had on average more sentences (21.7 vs. 16.4), more words (332.1 vs. 234.1), and much higher parsing accuracy (63.0\% vs. 59.9\% LAS). That is, essays with more words were actually easier to parse.

We can tentatively conclude the following from these results: 1) knowing the level of a learner before parsing may influence the choice of parsing model, in our case differentiating when one performs annotation transformations, i.e, parses with more local decisions or not; and 2) if one does not know the learner level, various features may combine to help select a parsing model (e.g., number of words and sentences). In short, the handling of linguistic decisions for parsing may need to be optimized differently for different kinds of learners.

\textbf{Underspecification} Turning to why the POS\textit{Both} model performs the same as the POS\textit{M} model, one effect we see is of underspecified (distributional) tags in both of these models, tags which the parser has never seen before—i.e., sparsity is a major issue. In figure 4, for instance, we see the effect of different parsing models for (1).

\begin{align*}
(1) & \quad \text{I have experience of living in a country that have see war and experience of living as immigrant and so they all made my characteristics. (Essay 021)}
\end{align*}

In this learner sentence, neither have nor see is correctly inflected, which leads to underspecified POS\textit{D} tags (VH and VV, respectively). Consequently, SALLE treats both of them as potential heads of a modifying clause (CMOD) \cite{7}. Figure 4 shows the difference in dependency labels between the POS\textit{M} and POS\textit{D} parsed trees, as compared to the gold annotation in the top tree. Most clearly, war is correctly an OBJ when see is a known type of verb (VV0) in the POS\textit{M} tree, whereas it becomes a MOD (a label appropriate for nominal modifiers) when the verb label is underspecified and the parser cannot recognize it. The verb see itself is labeled as POBJ (prepositional object) in the POS\textit{D} parsed tree, a label appropriate for a noun but not a verb. With respect to the MCOORD (modificatory coordination) label applied to see in the POS\textit{M}, this most probably is the influence of training data inaccuracies \cite[see 19, for more on that issue]{7}.

\footnote{For more on the treatment of \textit{of living}, see sections 3.3.1 and 4.5 of \cite{7}. The parser makes reasonable decisions in these cases, and the proper handling of \textit{ing} forms is an area of active investigation.}
4 Conclusion and Outlook

We have experimented with parsing learner language directly, without mapping to target hypotheses or correcting errors first [23], using the SALLE scheme, the only scheme we know of which explicitly favors syntactic interpretations over semantic ones and is highly form-based. Using a training corpus with slightly different conventions, we have specifically investigated whether certain linguistic decisions in the annotation affect parsing differently, exploring: 1) different combinations of POS tags, 2) the effect of transforming annotation before or after parsing. We have seen a slight favoritism for morphologically-based POS tags—with the caveat that distributionally-based POS tags suffered from unknown tags—and a clear favoritism for transforming the annotation before training a parser, likely due to differences in locality. Despite this, examining individual learner files shows that parsing the data of lower-level learners may benefit from a different transformation process.

The overall results are lower than in previous parsing experiments for learner data [4, 10], and future work will have to tease apart the effect of the morphologically-driven distinctions (which should in principle be easier) and the training data (e.g., annotation errors) and transformation process. Current experiments indicate that changing the training data (from an L1 acquisition data set to a native English treebank) can yield higher parsing results. Future work will look more closely at the
constructions that are difficult for the parser to analyze correctly. An immediate goal is to increase the size of the SALLE gold-annotated learner data to see if the trend still holds, given more testing data. Developing a parser to pre-process the data should provide a faster route to obtaining more annotated data. One can also further unpack the influence of underspecified information.

Acknowledgements

We would like to thank the three anonymous reviewers for their helpful feedback.

References


