Reliability of automatic linguistic annotation: native vs non-native texts

Elena Volodina\textsuperscript{1}, David Alfter\textsuperscript{1}, Therese Lindström Tiedemann*, Maisa Lauriala*, Daniela Piipponen*

* = University of Helsinki, \textsuperscript{1} = University of Gothenburg
“L2 profiles”

Development of Lexical and Grammatical Competences in Immigrant Swedish

Project financed by
Riksbankens jubileumsfond (2018–2020 (2021))
P17-0716:1

Collaboration:
Swedish language, University of Gothenburg
“Nordica”, Dept of Finnish, Finnougrian and Scandinavian Studies, University of Helsinki

PI: Elena Volodina, Gothenburg

Homepage:
https://spraakbanken.gu.se/en/projects/l2profiles
Swedish (L2) Profile
– lexical, grammatical, morphological.

https://spraakbanken.gu.se/larkalabb/svlp

Language Acquisition Reusing Korp

Lexical profile
Grammatical profile
Morphological profile
Swedish (L2) Profile
– **lexical, grammatical, morphological.**

<table>
<thead>
<tr>
<th>Exercise Generator</th>
<th>Tools</th>
<th>Experimental prototypes</th>
<th>Swedish Profile</th>
</tr>
</thead>
</table>

Language Acquisition Reusing Korp

Lexical profile

Grammatical profile

Morphological profile

[https://spraakbanken.gu.se/larkalabb/svlp](https://spraakbanken.gu.se/larkalabb/svlp)

Reliable?
Manual checking

**Sparv pipeline** (Borin et al. 2016)

- Lemmatization
- Part-of-speech tagging
- Multi-word expression detection
- Word sense disambiguation
- Dependency parsing
Hypotheses

1. Pipelines trained on a standard language (L1) do not perform as well on non-standard language varieties such as learner language (L2)

2. Normalization of non-standard language, e.g. through error correction, improves tool performance
Hypotheses

1. Pipelines trained on a standard language (L1) do not perform as well on non-standard language varieties such as learner language (L2)

2. Normalization of non-standard language, e.g. through error correction, improves tool performance

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Lemma</th>
<th>PoS</th>
<th>DepRel</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Coctail</td>
<td>0.93±0.0</td>
<td>0.98±0.0</td>
<td>74.89</td>
</tr>
<tr>
<td>L2 orig</td>
<td>0.90±0.02</td>
<td>0.95±0.0</td>
<td>63.01</td>
</tr>
<tr>
<td>L2 norm</td>
<td>0.93±0.02</td>
<td>0.97±0.0</td>
<td>69.02</td>
</tr>
</tbody>
</table>

Table 2: Lemmatization and PoS tagging: accuracy and standard deviation; Dependency: LAS
Hypotheses

1. Pipelines trained on a standard language (L1) do not perform as well on non-standard language varieties such as learner language (L2).

2. Normalization of non-standard language, e.g. through error correction, improves tool performance.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Lemma</th>
<th>PoS</th>
<th>DepRel</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Coctaill</td>
<td>0.93±0.0</td>
<td>0.98±0.0</td>
<td>74.89</td>
</tr>
<tr>
<td>L2 orig</td>
<td>0.90±0.02</td>
<td>0.95±0.0</td>
<td>63.01</td>
</tr>
<tr>
<td>L2 norm</td>
<td>0.93±0.02</td>
<td>0.97±0.0</td>
<td>69.02</td>
</tr>
</tbody>
</table>

Table 2: Lemmatization and PoS tagging: accuracy and standard deviation; Dependency: LAS

<table>
<thead>
<tr>
<th></th>
<th>#tokens excl punct</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>WSD Accuracy (corr/tot)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Coctaill</td>
<td>1900</td>
<td>0.80</td>
<td>0.71</td>
<td>0.75</td>
<td>0.84±0.03</td>
</tr>
<tr>
<td>L2 orig</td>
<td>3635</td>
<td>0.90</td>
<td>0.72</td>
<td>0.80</td>
<td>0.82±0.07</td>
</tr>
<tr>
<td>L2 norm</td>
<td>3565</td>
<td>0.85</td>
<td>0.78</td>
<td>0.81</td>
<td>0.83±0.04</td>
</tr>
</tbody>
</table>

Table 3: Number of correctly identified MWEs including precision, recall and F1 score and automatic word sense disambiguation (WSD)
Conclusions

Reliable?
- yes, for lemmatization, POS tagging, WSD & MWE detection
- not really for DepRel

Hypotheses?
1. yes, in general, but depend on linguistic features
2. yes, even if sometimes marginally
Thank you!

Questions? Comments? Suggestions?