Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank

Tak-sum Wong*, Kim Gerdes*, Herman Leung*, John Lee*

*Department of Linguistics and Translation
City University of Hong Kong

*Sorbonne Nouvelle, LPP (CNRS)
Paris, France
Introduction

• Cantonese, a Sinitic language, spoken by 55M people mostly in Canton, Hong Kong, Macao. “Cantonese is the most widely known and influential variety of Chinese other than Mandarin” (Matthews & Yip 1994)
• The special status of Hong Kong and Macao and the economic and educational importance of the region has made Cantonese a relatively well-studied and well-resourced language.
• A number of POS-tagged corpora exist but no syntactic treebank has been published.
• We are presenting the first parallel dependency treebank for Cantonese and Mandarin and analyze the statistical differences.
Treebank Construction

- Annotation scheme was adapted from existing UD guidelines for standard Chinese (Leung et al., 2016)
- Source Material: Hong Kong television programmes, with Mandarin subtitles
- Size: 569 parallel sentences
- Sentence-aligned
- Semi-planned spoken text
- Cantonese transcription was done independently of Mandarin subtitles
- Subtitles are always condensed, and simplified dialogues
- Treebank is not as strictly parallel

<table>
<thead>
<tr>
<th>Language</th>
<th>#tokens</th>
<th>avg sent length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandarin</td>
<td>4149</td>
<td>7.29</td>
</tr>
<tr>
<td>Cantonese</td>
<td>5428</td>
<td>9.54</td>
</tr>
</tbody>
</table>
# Statistical Measures

## Categorical differences

<table>
<thead>
<tr>
<th>Type</th>
<th>Specificity</th>
<th>Cantonese</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUNCT</td>
<td>31</td>
<td>999</td>
<td>1344</td>
</tr>
<tr>
<td>INTJ</td>
<td>23</td>
<td>97</td>
<td>97</td>
</tr>
<tr>
<td>PART</td>
<td>10</td>
<td>619</td>
<td>898</td>
</tr>
<tr>
<td>AUX</td>
<td>0</td>
<td>246</td>
<td>428</td>
</tr>
<tr>
<td>CCONJ</td>
<td>0</td>
<td>18</td>
<td>33</td>
</tr>
<tr>
<td>SCONJ</td>
<td>0</td>
<td>23</td>
<td>41</td>
</tr>
<tr>
<td>ADJ</td>
<td>-1</td>
<td>97</td>
<td>186</td>
</tr>
<tr>
<td>NOUN</td>
<td>-1</td>
<td>801</td>
<td>1449</td>
</tr>
<tr>
<td>NUM</td>
<td>-1</td>
<td>54</td>
<td>104</td>
</tr>
<tr>
<td>PROPN</td>
<td>-1</td>
<td>84</td>
<td>155</td>
</tr>
<tr>
<td>DET</td>
<td>-4</td>
<td>60</td>
<td>144</td>
</tr>
<tr>
<td>VERB</td>
<td>-4</td>
<td>347</td>
<td>688</td>
</tr>
<tr>
<td>PRON</td>
<td>-5</td>
<td>462</td>
<td>915</td>
</tr>
<tr>
<td>ADP</td>
<td>-8</td>
<td>93</td>
<td>239</td>
</tr>
<tr>
<td>ADV</td>
<td>-11</td>
<td>511</td>
<td>1080</td>
</tr>
</tbody>
</table>

## Functional measures

<table>
<thead>
<tr>
<th>Type</th>
<th>Spec</th>
<th>Cantonese</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>punct</td>
<td>31</td>
<td>1002</td>
<td>1345</td>
</tr>
<tr>
<td>discourse</td>
<td>26</td>
<td>204</td>
<td>226</td>
</tr>
<tr>
<td>discourse:sp</td>
<td>11</td>
<td>443</td>
<td>619</td>
</tr>
<tr>
<td>advc:ecoverb</td>
<td>9</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>det</td>
<td>3</td>
<td>193</td>
<td>286</td>
</tr>
<tr>
<td>advc</td>
<td>-2</td>
<td>91</td>
<td>184</td>
</tr>
<tr>
<td>nmod</td>
<td>-2</td>
<td>99</td>
<td>204</td>
</tr>
<tr>
<td>obj</td>
<td>-2</td>
<td>393</td>
<td>726</td>
</tr>
<tr>
<td>mark:rel</td>
<td>-3</td>
<td>20</td>
<td>56</td>
</tr>
<tr>
<td>nsubj</td>
<td>-3</td>
<td>362</td>
<td>707</td>
</tr>
<tr>
<td>xcomp</td>
<td>-3</td>
<td>64</td>
<td>140</td>
</tr>
<tr>
<td>dislocated</td>
<td>-4</td>
<td>62</td>
<td>148</td>
</tr>
<tr>
<td>obl</td>
<td>-5</td>
<td>58</td>
<td>147</td>
</tr>
<tr>
<td>ccomp</td>
<td>-6</td>
<td>56</td>
<td>145</td>
</tr>
<tr>
<td>advmod</td>
<td>-7</td>
<td>541</td>
<td>1087</td>
</tr>
<tr>
<td>obl:obj</td>
<td>-7</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>case</td>
<td>-14</td>
<td>80</td>
<td>245</td>
</tr>
</tbody>
</table>
## Statistical Measures

### Mixed measures

<table>
<thead>
<tr>
<th>Type</th>
<th>Spec</th>
<th>Cantonese</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>VERB-punct→PUNCT</td>
<td>24</td>
<td>595</td>
<td>781</td>
</tr>
<tr>
<td>INTJ-punct→PUNCT</td>
<td>22</td>
<td>93</td>
<td>93</td>
</tr>
<tr>
<td>NOUN-det→NOUN</td>
<td>19</td>
<td>126</td>
<td>135</td>
</tr>
<tr>
<td>VERB-discourse→INTJ</td>
<td>15</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>VERB-discourse→PART</td>
<td>12</td>
<td>369</td>
<td>503</td>
</tr>
</tbody>
</table>

### Directional measures

<table>
<thead>
<tr>
<th>name</th>
<th>advmod</th>
<th>aux</th>
<th>obj</th>
<th>obl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cantonese</td>
<td>13,74</td>
<td>48,82</td>
<td>100</td>
<td>28,08</td>
</tr>
<tr>
<td>Mandarin</td>
<td>3,81</td>
<td>35,16</td>
<td>100</td>
<td>19,67</td>
</tr>
</tbody>
</table>

VERB-advmod→ADV          | -10   | 332  | 729  |
AUX-ccomp→VERB           | -14   | 0    | 38   |
Artefacts vs. typology

• Parallel corpus, but:
  – Artefacts:
    • Different conventions
      → *punct* much more frequent in Cantonese
    • Translationese (genre)
      → *INTJ* much more frequent in Cantonese
  – Typology:
    • All points without explanation as artefact
      – Some conscious annotation choices
      – Some discoveries post-annotation
Preposition and (co)verb

- Cantonese coverb is tagged as VERB+advcl:coverb
- Mandarin coverb is tagged as ADP (preposition) +case

Cantonese

I am talking with her

Mandarin
Noun(classifier) and determiner

- “Bare classifier” construction in Cantonese: [classifier + noun] as definite NP
- Aligned to a Mandarin demonstrative
Sentence particle and adverb

- Some Cantonese sentence particles correspond to Mandarin adverbs

| Cantonese | | | | | | Mandarin |
|---|---|---|---|---|---|
| 食 | 吃 | PRF | cold | thing | 先 / ADV |
| 冻 | cold | NOM |
| 先 | first |

Eat the cold [things] first
Conclusions

• A method of empirical comparative syntax using statistical measures on a sentence-aligned parallel dependency treebank.

• Significant observations can be explained by actual differences in the language structure.

• Subtle genre differences on the two sides of our treebank: transcription vs subtitle is still visible.
On-going Work

• Development of word alignment between Mandarin and Cantonese
• Transcribe materials distributed on Youtube for free language resource
• Analysing other constructions showing asymmetric difference between these two languages
• Application: for teaching Cantonese as a foreign language
Fisher Test and Specificity

Specificity = \begin{cases} 
-\log_{10}(p) \\
\log_{10}(1-p) 
\end{cases}

- Cantonese: lower frequency of adverbs
- prominence of Cantonese post-verbal particles
- Mandarin: uses adverb more often
- Mandarin: zhèngzài + V
- Cantonese: V-gán
Some Interesting Constructions

**Double objects**

For a ditransitive verb, in Cantonese we have the following word order:

verb + direct object + indirect object.

畀 一枝花 我
Pei yatji fa ngoh
give a flower 1sg
‘Give me a flower.’

In Mandarin it is

verb + indirect object + direct object.

給 我 一枝花
Gei wo yizhi huar
give 1sg a flower
‘Give me a flower.’

**Object marker**

門咗度門啦！
Saan jo douh muhn la!
close PERF CLF door SFP
‘Close the door!’

PERF = perfective particle
CLF = classifier
SFP = sentence final particle

vs.

將度門門咗（佢）啦！
Jewng douh muhn sanan jo (keuih) la!
OM CLF door close PERF (3SG) SFP
‘the Door, close (it)’

These two alternative constructions recall the English dative shift alternation.
Some Interesting Constructions

Post-verbal modifiers

Cantonese:

```
Wa! Jáu saai làh?
Wow go all SFP
‘Wow! All of them have gone already’ / ‘They have all gone?’ / ‘They have all been released from duty?’
```

Mandarin:

```
Dōu xiàbān le ma
all off-duty ASP SFP
```

Coverb constructions

Cantonese:

```
Ngóh pūih léihdeih jahpheui á
1sg accompany 2pl go.inside SFP
‘Let me enter / go into the shop with you!’
```

Mandarin (0_28):

```
Wò péi nīmen jìnqù ba
1sg accompany/with 2pl go.inside SFP
```
Some Interesting Constructions

Expletives

大家 飲勝 佢！
Daaihgā jāmsing kēuih
everyone cheers KEUHI
‘Everyone! Cheers (to it)!’

我 不如 死 咬 佢 好過 啦!
Ngóh bātyūh séi jó kēuih hōugwo lā
1SG had.better die PERF KEUHI better SFP
‘It would be better for me to die.’