Translation mining in the domain of conditionals: first results

Jos Tellings – j.l.tellings@uu.nl
Utrecht University / UiL-OTS
CLIN 30, 30/01/2020

1 Translation mining

1.1 From the lexical to the clausal level

Translation mining is a methodology of using parallel corpora of translated texts to investigate cross-linguistic semantic variation, that results in a visualization of variation by means of semantic maps (van der Klis et al. 2017). It has been applied to a number of different domains:

(1) a. motion verbs (Wälchli and Cysouw 2012);
   b. definite determiners (Bremmers et al. 2019);
   c. tense (Time in Translation project; Le Bruyn et al. 2019).

An example of a semantic map can be found in Figure 2 at the end of this handout.

The constructions in (1) are all single-word or single-phrase constructions, which I will also refer to as simplex cases. This paper makes the move from translation mining with simplex constructions to translation mining with larger, sentence-size constructions. I intend the concepts developed here to apply to clausal phenomena in general, but I will focus on the area of conditionals (sentences with an if-clause, e.g. (2)).

1.2 Conditionals as a case study

(2) a. But if we are to believe the environmental movements of clause 1, even this is van hopeclause 2.
   b. Maar als wij de milieubewegingen mogen geloverclause 2, is zelfs dat iijdele hoopclause 2.

- Conditionals are a good case study: they are clausal constructions, but tense and aspect play a crucial role. The contributions of tense and aspect in the if-clause and main clause are thought to result in various semantic and pragmatic effects of conditionals (Iatridou 2000, Declerck and Reed 2001: ch. 5, Arregui 2007, Ippolito 2013, and many more).
- Here compositionality comes into play: we are interested in the various building blocks of conditional expressions (in particular tense, aspect, and modal expressions), but also the semantic and pragmatic effects of combining the blocks in various languages.
- The Time in Translation project found a lot of cross-linguistic variation with respect to tense use. What does this variation look like when one restricts the view to the domain of conditionals? Does variation in tense use lead to predictable variation in the meaning of conditionals, given the compositional role formal analyses have assigned to tense and aspect?
- In this talk I will focus on conceptual issues and first results. Alongside, the computational tools required for doing translation mining with clausal data are being developed as an extension of the TimeAlign software1 (van der Klis and Bonfil, this conference).

2 Methodology in steps

The translation mining methodology process consists of three steps: (1) extraction of relevant construction with translations from a parallel corpus; (2) alignment and annotation of relevant properties; (3) creation of semantic maps.

---

*This research is part of the Time in Translation research project (http://time-in-translation.hum.uu.nl), funded by NWO grant 360-80-070, which is gratefully acknowledged. I thank the members of Time in Translation for their valuable input and discussion at various stages of this work.

Step 1: extraction of conditionals from a corpus

- I used the Europarl corpus of proceedings from the European Parliament (Koehn 2005), because its more formal register leads to a high number of (counterfactual) conditionals.

- I wrote a simple script (an adaptation of the PerfectExtractor script from the Time in Translation project)\(^2\) that extracts combinations of ‘if’ and certain tense combinations from the English part of the corpus, together with its translations. This yields a large dataset (for ‘if … would …’: \(N = 21371\)), but contains false positives that need to be manually removed, in particular complement if-clauses.

Step 2: alignment and annotation

- Next, the data need to be annotated for properties of interest for cross-linguistic comparison.

- Conditionals are bi-clausal constructions consisting of an if-clause (antecedent, protasis), and a main clause (consequent, apodosis). Certain properties of interest relate to the construction as a whole, others to the component clauses. See Figure 1.

\(^2\)https://github.com/UUDigitalHumanitieslab/perfectextractor

Step 3: semantic maps and distance function

- Some properties can be annotated automatically (like if-clause/main clause division and ordering), but most will be done by human annotators (following the practice in the tense domain from the Time in Translation research project).

- In order to facilitate human annotators, we want to keep the two clauses separated, yet linked, as illustrated in Figure 1.

<table>
<thead>
<tr>
<th>Simplex</th>
<th>Compositional</th>
</tr>
</thead>
<tbody>
<tr>
<td>example of annotation label:</td>
<td>Perf</td>
</tr>
<tr>
<td>example of translation tuple:</td>
<td>⟨Perf, Fut, ott⟩</td>
</tr>
<tr>
<td>d-function:</td>
<td>(d(s, t))</td>
</tr>
</tbody>
</table>

I’ll write capital \(S, T\) for tuples of sequences of annotation labels (i.e. translation tuples in the compositional case).

- In most earlier work, the d-function used in the simplex case is a version of the Hamming distance:

\[
(3) \quad \text{input: } n\text{-tuples of translations } s, t \\
\text{distance: } d(s, t) = \frac{1}{n} \cdot |\{1 \leq i \leq n : s_i \neq t_i\}| \\
\text{example: } d(⟨\text{Perf, vtt, Perfekt, Fut}⟩, ⟨\text{Perf,vtt,Prät,Imp}⟩) = 2/4 = .5 \\
\]

| English | Dutch | German | French |
---|---|---|---|
| Perf | Fut | vtt | Perfekt |
| Imp | cond | zou |
• In the compositional case, I propose to define a distance function on $S, T$
by lifting any simplex distance function $d(s, t)$ from a function on tuples
of labels, to a function on tuples of sequences of annotations:

\[(4) \quad \text{lift}_i(d)(S, T) = d(\langle S_{1,i}, \ldots, S_{n,i} \rangle, \langle T_{1,i}, \ldots, T_{n,i} \rangle)\]

‘only consider the $i$th coordinate of each annotation label’

• For example, let
  $S = \langle \text{[Past, Past]}, [\text{ott, vtt}], [\text{imparf, conditionnel}] \rangle$, and
  $T = \langle \text{[Past, Perf]}, [\text{ott, vtt}], [\text{imparf, imparf}] \rangle$. Then:

\[(5) \quad \text{lift}_1(d)(S, T) = d(\langle \text{Past, ott, imparf} \rangle, \langle \text{Past, ott, imparf} \rangle) = 0;\]

\[\text{lift}_2(d)(S, T) = d(\langle \text{Past, vtt, conditionnel} \rangle, \langle \text{Perf, vtt, imparf} \rangle) = \frac{2}{3}.\]

• This effectively gives access to the component annotations of the conditionals, so $d_1$ can be used to create a map of the antecedent tense in the conditionals, and compare it to earlier maps.

• On the other hand, some combined distance $f(\text{lift}_1(d), \text{lift}_2(d))$ that weighs the two tense annotations can now be defined. In the compositional setting, this introduces an additional choice for the distance function.

3 Case study: English were to conditionals

• The English *be to* construction combines future time reference with modality, often necessity (Declerck 2010):

\[(6) \quad \text{a. The guests are to arrive soon.} \quad \text{(Declerck 2010)}\]
\[\text{b. If all the costs were to be charged on, the question is whether} \]
\[\text{wind energy would not be equally competitive. \text{[Europarl]}}\]

• Its use in conditionals has been noted (see Declerck and Reed 2001: §6.4.1; Declerck 2010: §6). Indeed, in my dataset the construction is quite frequent (subjunctive $N = 2506$; indicative $N = 4753$).
Translation mining in the domain of conditionals

<table>
<thead>
<tr>
<th>Translated as a conditional?</th>
<th>Cond. 78</th>
<th>Cond. 42</th>
<th>Non-cond. 22</th>
<th>Non-cond. 58</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch tense used in if-clause:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zou + infinitive</td>
<td>42</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ott (Present)</td>
<td>29</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ovt (Past)</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vvt (Past Perfect)</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mocht-conditional</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modal verb present in Dutch translation:</td>
<td>3</td>
<td>25</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>willen ‘want’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>moeten / dienen ‘must’</td>
<td>0</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kunnen ‘can’</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gaan ‘go’</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No modal verb</td>
<td>69</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Tense/modal properties for 100 subjunctive and 100 indicative be to conditionals

Three striking results from Table 1:

1. A high number of non-conditional translations in Dutch. In many cases, Dutch uses an infinitival clause:

   (7) a. We must be more ambitious and more decisive if we are to resolve the causes of inequality, violence and poverty.
   
   b. Wij moeten meer wilskracht en daadkracht tonen om de ongelijkheid, het geweld en de armoede uit de wereld te helpen.

   This raises the question if in English, too, an infinitival clause can represent the meaning of an if-clause.

2. Many translations have the present tense, also in the subjunctive case.

   (8) a. If this were to be recognised in the EU, she would then be able to bring her entire family to join her.

   b. Als dit in de EU wordt erkend (Pres) dan kan zij vervolgens ook haar gezin hier naartoe laten komen.

   Not all instances of were to can be translated by a Dutch present tense, so this leads into questions about different readings of were to (see Declerck 2010), as well as the futurate use of Dutch present tense.

3. Many modal verbs are present in the indicative case.

   (9) a. If we were to propose that, we would of course consult the Parliament.

   b. Als wij een dergelijke uitbreiding willen voorstellen (Pres), zullen wij het Parlement natuurlijk raadplegen.

   Note that the use of modal auxiliary zou is frequent in the subjunctive case. This is known to play a role as a subjunctive marker (see Nieuwint 1984; Roels et al. 2007), but may also add modal flavor.

   (10) a. If we were to adopt the proposal [...], the entire computer system would need to be changed.

   b. Als wij het voorstel [...] zouden volgen (zou + inf) [...] dan moet je je hele computersysteem veranderen.

   • Declerck (2010: 286) writes: “Be to can be used in if-clauses where the hypothesis that is made concerns a goal which the referent of the subject may wish to attain”.

   (11) If we are to make any progress at all, we must set about finding a solution to this problem. (ibid., p. 286)

   This use seems to be reflected in the frequent occurrence of modal willen ‘want’.
4 Conclusion

- Even though this is a small and preliminary case study, it illustrates the potential of applying Translation Mining to investigate variation in a clausal domain:
  - variation at a higher level (various ways of how conditionality is expressed across languages);
  - variation at a lower level (differences in the use of tense and modality inside conditional constructions).
- The methodology is not restricted to conditional sentences, but can be used for other compositional and clausal phenomena.
- The computational interface for annotation is under development (van der Klis and Bonfil, this conference). When this is in place, we will be able to look at larger datasets, and create semantic maps.

References


Appendix: semantic map example

Figure 2. Example of a semantic map obtained from Translation Mining applied to the distribution of Present Perfect forms. The semantic map reveals a richer variation in Present Perfect use than is assumed in theoretical studies. In particular, a subset relation is found distinguishing ‘core’ uses from typologically more marked uses (see van der Klis et al. 2019 for more details).