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)).

*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.

1.2 *Conditionals as a case study*

(2) a. But **if** we are to believe the environmental movements$_{\text{clause 1}}$, even this is vain hope$_{\text{clause 2}}$.

b. Maar **als** wij de milieubewegingen mogen geloven$_{\text{clause 2}}$, is zelfs dat ijdele hoop$_{\text{clause 2}}$.

- 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 software[1] (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.

[1] See https://github.com/UUDigitalHumanitieslab/timealign for the source code, and https://time-in-translation.hum.uu.nl/timealign/introduction/ for an introduction.

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.

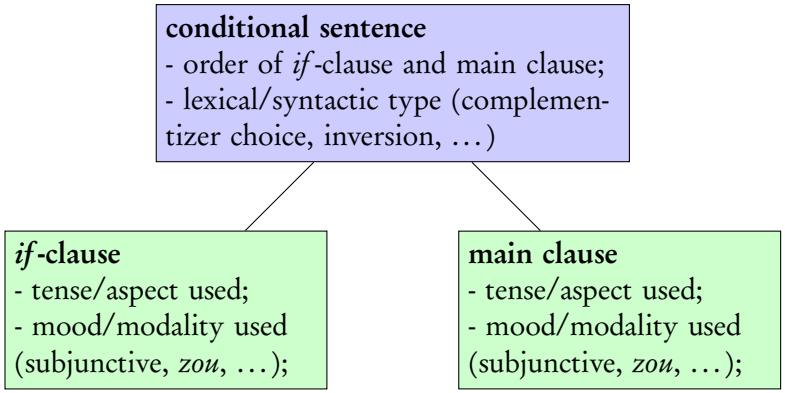

Figure 1. Constructions and subconstructions

[2] https://github.com/UUDigitalHumanitieslab/perfectextractor

- 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.

Step 3: semantic maps and distance function

- When the data and their translations have been annotated, the final step is the creation of semantic maps that visualize variation.
- First, a distance metric is defined as a notion of distance between translations. Then, the statistical technique of multi-dimensional scaling (MDS, see also Wälchli and Cysouw 2012: §5) is used to reduce the multi-dimensional difference data to a two-dimensional representation.

	Simplex	Compositional
example of annotation label:	Perf	[Perf, might, Perf]
example of translation tuple:	⟨Perf, Fut, ott⟩	⟨[Perf,might], [imp, cond], [vtt, zou]⟩
d-function:	$d(s,t)$	$d(S,T)$

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) input: n-tuples of translations s, t
distance: $d(s,t) = \frac{1}{n} \cdot \left|\{1 \leqslant i \leqslant n : s_i \neq t_i\}\right|$
example: $d(\langle \text{Perf}, \text{vtt}, \text{Perfekt}, \text{Fut}\rangle, \langle \text{Perf}, \text{vtt}, \text{Prät}, \text{Imp}\rangle) = 2/4 = .5$
English Dutch German French

- 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) $\mathbf{lift}_i(d)(S,T) = d(\langle S_{1,i}, \ldots, S_{n,i}\rangle, \langle T_{1,i}, \ldots, T_{n,i}\rangle)$
  'only consider the ith coordinate of each annotation label'

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

  (5) $\mathbf{lift}_1(d)(S,T) = d(\langle$Past, ott, imparf $\rangle, \langle$ Past, ott, imparf $\rangle) = 0$;
  $\mathbf{lift}_2(d)(S,T) = d(\langle$Past, vtt, condit $\rangle, \langle$ 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 $f(\mathbf{lift}_1(d), \mathbf{lift}_2(d))$ 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) a. The guests **are to** arrive soon. (Declerck 2010)

  b. If all the costs **were to** be charged on, the question is whether wind energy would not be equally competitive. [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 $N = 4753$).

- As a first investigation, I extracted 100 indicative and 100 subjunctive *be to* conditionals and their Dutch translations. These were then manually annotated for tense.

⟶ Turn page for results.

	were to	*is/are to*
Translated as a conditional?	Cond. 78 Non-cond. 22	Cond. 42 Non-cond. 58
Dutch tense used in *if*-clause:		
zou + infinitive	42	0
ott (Present)	29	41
ovt (Past)	4	1
vvt (Past Perfect)	1	0
mocht-conditional	2	0
Modal verb present in Dutch translation:		
willen 'want'	3	25
moeten / *dienen* 'must'	0	5
kunnen 'can'	3	1
gaan 'go'	1	0
No modal verb	69	11

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

Arregui, A. (2007). When aspect matters: the case of *would*-conditionals. *Natural Language Semantics*, *15*, 221–264.

Bremmers, D., Liu, J., van der Klis, M., & Le Bruyn, B. (2019). Definiteness across languages: from German to Mandarin. In J. J. Schlöder, D. McHugh, & F. Roelofsen (eds.), *Proceedings of the 22nd Amsterdam Colloquium*, pp. 60–70.

Declerck, R. (2010). Future time reference expressed by *be to* in Present-day English. *English Language & Linguistics*, *14*(2), 271–291.

Declerck, R., & Reed, S. (2001). *Conditionals. A Comprehensive Empirical Analysis*. Berlin / New York: Mouton de Gruyter.

Iatridou, S. (2000). The Grammatical Ingredients of Counterfactuality. *Linguistic Inquiry*, *31*(2), 231–270.

Ippolito, M. (2013). *Subjunctive conditionals*. Cambridge, MA: MIT Press.

van der Klis, M., & Bonfil, B. (2020). *Parallel corpus annotation and visualization with TimeAlign*. Poster at CLIN30.

van der Klis, M., Le Bruyn, B., & de Swart, H. (2017). Mapping the PERFECT via Translation Mining. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 497–502).

van der Klis, M., Le Bruyn, B., & de Swart, H. (2019). *A multilingual corpus study of the competition between PAST and PERFECT in narrative discourse*. Ms., Utrecht University.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Machine Translation summit* (Vol. 5, pp. 79–86).

Le Bruyn, B., van der Klis, M., & de Swart, H. (2019). The Perfect in dialogue: evidence from Dutch. *Linguistics in the Netherlands*, *36*, 162–175.

Nieuwint, P. (1984). Werkwoordstijden in Nederlandse counterfactuals [Verb tenses in Dutch counterfactuals]. *De Nieuwe Taalgids*, *77*, 542–555.

Roels, L., Mortelmans, T., & van der Auwera, J. (2007). Dutch equivalents of the German past conjunctive: *zou* + infinitive and the modal preterit. In L. de Saussure, J. Moeschler, & G. Puskás (eds.), *Tense, Mood and Aspect*, pp. 177–196. Brill Rodopi.

Wälchli, B., & Cysouw, M. (2012). Lexical typology through similarity semantics: toward a semantic map of motion verbs. *Journal of Linguistics*, *50*, 671–710.

Appendix: semantic map example

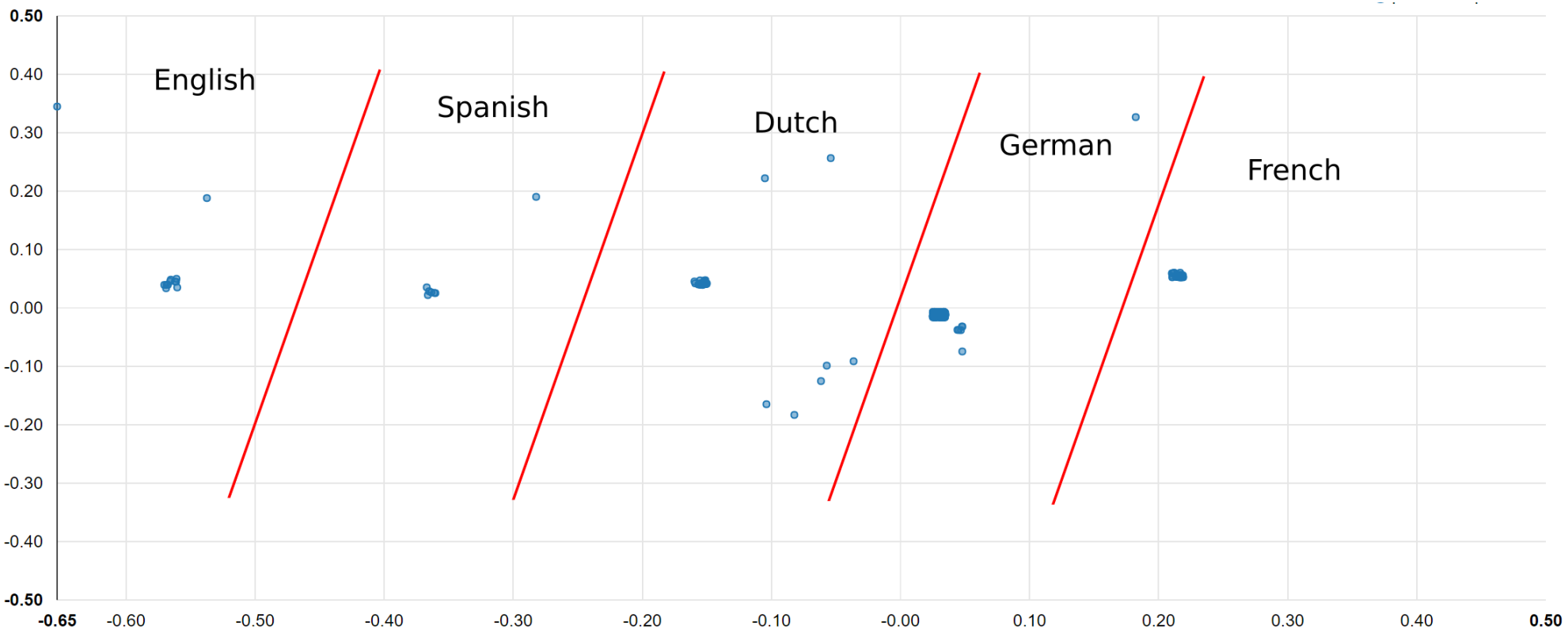

Figure 2. Example of a semantic map obtained from Translation Mining applied to the distribution of Present Perfect forms. The semantic map reveals a richer variation in Present Perfect use than is assumed in theoretical studies. In particular, a subset relation is found distinguishing 'core' uses from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