From parallel corpora to the formal study of compositional variation

Jos Tellings

Utrecht University
j.tellings@gmail.com

28 May 2021

Plan:

1. Introduce the *Translation Mining* methodology, and how it connects functional and formal approaches to cross-linguistic variation.
2. Extend the methodology to sentential constructions in order to investigate the compositional interaction of multiple features
3. Case study: data from conditionals

Part 1: Translation Mining connects functional and formal approaches

- Translation mining (term due to van der Klis et al. 2017):
  - the use of parallel corpus data to analyze cross-linguistic variation;
  - the use of multidimensional scaling (MDS) to visualize variation.

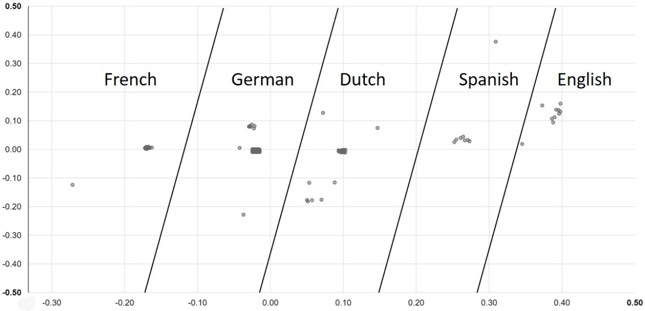
  The visualization is the starting point for further (formal) linguistic analysis.
- Bible corpora, translated novels, Europarl (proceedings of European parliament), film subtitles, . . .
- Used in typology, for language classification (example: Dahl and Wälchli (2016), 1107 languages), but more recently also in formal studies of a single language.

- The statistical technique of multidimensional scaling (MDS) is used to reduce high-dimensional variation data so that it can be visualized.
- Items that are similar to each other are displayed close to each other in the visualization.

  ⇒ example similarity measure between contexts: how many languages use the same form in both contexts?
- The dots in the scatterplot correspond to contexts in the corpus, and provide the stepping stone towards formal analysis.
- Map interpretation proceeds by identifying clusters and interpreting dimensions.

  [web interface example at time-in-translation.hum.uu.nl]
- van der Klis and Tellings (2021): overview paper with explanation of MDS, and how it connects to formal linguistic analysis.

- Translation mining is also used in small/single language studies, in combination with formal approaches (see van der Klis and Tellings 2021).
- Examples:
  - P. de Swart et al. (2012) on Greek prepositions;
  - Bremmers et al. (2021) on definite determiners;
  - Perfect tense in European languages (van der Klis et al. 2022)



Part 2: extend methodology

- Extend the methodology to study larger units than just lexical items.
- In sentence-size constructions, there will typically be multiple features that vary across languages, although their interaction may not vary.
  ⇒ negation, modality, clause order, . . .
- Annotate multiple features per construction, and study the interaction of these features.
  [web interface example]
- A MDS analysis can be run that is based on a similarity measure that takes various features into account.

- Example: joint work with Henriëtte de Swart and Bernhard Wälchli (2021)
  ⇒ Europarl study with 7 languages on temporal (NPI) connectives *not . . . until*.
- Connects a formal approach (H. de Swart 1996) with a functional (typological) approach (Wälchli 2018).
- Preliminary findings: there is a lot of lexical variation, but the combination of negation and connective is stable, as predicted by formal approaches.
- Some German connectives we encountered:

(1) *nicht . . . solange . . . nicht*, *erst . . . wenn*, *bis*, *nicht . . . bevor*, . . .

Part 3: conditionals

- I extracted conditional sentences from Europarl in English, and translations in Dutch, French, and Spanish.
- I zoom in today on a construction that Declerck and Reed (2001) call "extraposed" semi-nominal conditionals:

(2) a. It would be splendid **if** the EU institutions were to live up to this.

b. Het zou goed zijn **als** de EU-instellingen zich hieraan hielden.

c. Sería conveniente **que** las instituciones de la UE cumplieran con esto.

d. Il serait bon **que** les institutions de l'UE se montrent à la hauteur de ces principes.

- Cf. standard CP extraposition:

(3) It is good [$_{\text{CP}}$ that Mary is here].

- Main pattern: in subjunctive conditionals, English uses *if*, whereas Spanish and French use both *que* and *si* (first noted in work by Angelica Hill (2020)).

(4) a. It would be better still **if** inspections were to take place biannually [. . .]

b. Het is nòg beter **als** de controle tweejaarlijks wordt [. . .]

c. Sería todavía mejor **que** el control se hiciera cada dos años [. . .]

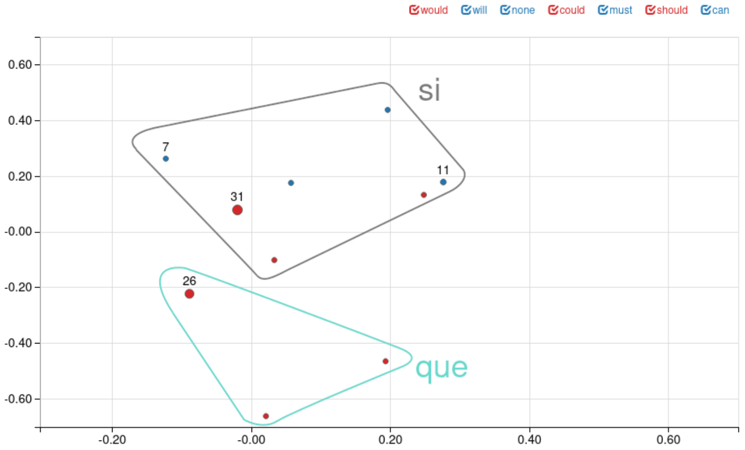
d. Ce serait encore mieux **si** le contrôle devenait biennal [. . .]

- The distribution of both complementizers is not the same in French and Spanish, and there appears to be no meaning difference between *si* and *que*.

- Interaction between conditional mood and complementizer choice. In English, complementizer choice is linked to the subjunctive/indicative contrast:

(5) a. It would be good {if, *that} you were here.
    b. It is good {that, #if} you are here.

- Annotated features include both modal structure of the conditional, and complementizer choice (among others).
- In the Romance languages, the choice between *si* vs. *que* carves up the space differently.


☑would
☑will
☑none
☑could
☑must
☑should
☑can
0.60
0.40
0.20
-0.00
-0.20
-0.40
-0.60
-0.20
-0.00
0.20
0.40
0.60
si
que
7
11
31
26

- In Dutch, there is limited variation between *als* ‘if’ and *dat* ‘that’ in seminominal constructions. A search in the SoNaR corpus finds the following:

	NL	B	Mixed	Total
het zou ADJ zijn dat	14	207	18	239
het zou ADJ zijn als	596	681	151	1428
	610	888	169	1667

- The use of *dat* in this construction seems to be mostly a Belgian Dutch phenomenon.
- A similar search for the English construction ‘it would be ADJ if/that’ in the iWeb corpus finds various adjectives for *if* (most commonly *nice* 15864, and *great* 12953). The construction with *that* only appears 821 times in total, with six adjectives, mostly modal ones (*unlikely*, *clear*, *obvious*, *possible*, *likely*, *important*).

- Factivity literature: *that*-clauses are not inherently factive, it depends on the predicate (Herriman 2000; Schulz 2012).

(6) a. Sue remembered that Mary went to London. (Schulz 2012, p. 15) [factive]

b. Sue thought that Mary went to London. [non-factive]

(7) a. It's good that Mary is coming tomorrow. [factive]

b. It's likely that Mary is coming tomorrow. [non-factive]

- However, there is a clash between the counterfactual inference of *would* and the *that*-phrase:

(8) *It would be nice that you could come tomorrow.

- *That*-phrase are possible, however, in counterfactual contexts:

(9) [Why did you take that expensive medical training??]
Well, if there were an emergency now, you would be happy that I took that training.

- The modal *would* carries a counterfactual presupposition (Condoravdi 2003).

(10) A holiday to Spain would be nice.
$\rightsquigarrow$ no holiday to Spain is currently planned

- *Would* often appears in (implicit) conditionals:

(11) John would be a great teacher.
= If John were a teacher, he would be a great teacher.
$\rightsquigarrow$ John is not a teacher

- This explains the badness of *that*-clauses in subjunctive constructions (unless the clause is true in the counterfactual worlds).

(12) *It would be nice that you could come tomorrow.

- There is cross-linguistic variation w.r.t to the factivity of complementizers.
  ⇒ Romance *que* does not clash with non-factive environments
- Declerck and Reed's (2001) derivation: not the *if*-clause gets extraposed, but an underlying *that*-clause.

(13) a. It would be nice if everybody minded their own business.
b. If everybody minded their own business, it would be nice that everybody minded their own business.

- They claim that the *if*-clause represents a true conditional meaning. This seems at odds with the corpus findings that many of these constructions get translated by non-conditional constructions.

- Use of infinitival clauses:

(14) a. It would have been better **if** she had obtained channels in all languages.

b. Il serait préférable de pouvoir capter des chaînes dans toutes les langues.

(15) a. It would be preferable if products made from PVC were to be replaced ...

b. Het zou beter zijn **om** PVC-producten te vervangen ...

- The extended *Translation Mining* methodology is suitable for comparing constructional meaning across languages.

Conclusion

- The methodology of *Translation Mining* connects functional and formal approaches to cross-linguistic variation.
- Extension of the methodology has been applied to data on temporal connectives, and conditional sentences.
- Conditional data show how this methodology can shed light on the cross-linguistic variation of factivity of complementizers.

Bremmers, D. et al. (2021). "Translation Mining: definiteness across languages. A reply to Jenks (2018)". Submitted.

Condoravdi, C. (2003). "Moods and Modalities for *Will* and *Would*". Handout of talk presented at Amsterdam Colloquium, December 2003.

Dahl, Ö. and B. Wälchli (2016). "Perfects and iamitives: two gram types in one grammatical space". In: *Letras de Hoje* 51.3, pp. 325–348.

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

De Swart, H. (1996). "Meaning and use of *not . . . until*". In: *Journal of Semantics* 13.3, pp. 221–263.

De Swart, H., J. Tellings, and B. Wälchli (2021). "*Not . . . until* across languages". In progress.

De Swart, P., H. M. Eckhoff, and O. Thomason (2012). "A Source of Variation: A Corpus-Based Study of the Choice between ἀπό and ἐκ in the NT Greek Gospels". In: *Journal of Greek Linguistics* 12.1, pp. 161–187.

Herriman, J. (2000). "Extraposition in English: A study of the interaction between the matrix predicate and the type of extraposed clause". In: *English Studies* 81.6, pp. 582–599.

Hill, A. (2020). "If Not *If* then *Que*: A Comparison of Spanish and English Counterfactuals". Manuscript.

Schulz, P. (2012). *Factivity: Its nature and acquisition*. Walter de Gruyter.

van der Klis, M., B. Le Bruyn, and H. de Swart (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., B. Le Bruyn, and H. de Swart (2022). "A multilingual corpus study of the competition between PAST and PERFECT in narrative discourse". To appear in *Journal of Linguistics* 2022.

van der Klis, M. and J. Tellings (2021). "Multidimensional scaling and linguistic theory". Submitted.

Wälchli, B. (2018). "'As long as', 'until' and 'before' clauses: Zooming in on linguistic diversity". In: *Baltic Linguistics* 9, pp. 141–236.

Wälchli, B. and M. Cysouw (2012). "Lexical typology through similarity semantics: Toward a semantic map of motion verbs". In: *Linguistics* 50.3, pp. 671–710.