From parallel corpora to the formal study of compositional variation

Part 1. parallel corpus work connects quantitative and formal approaches
In linguistic typology, a methodology has been established that uses data from massively parallel texts, and employs multi-dimensional scaling (MDS) techniques to visualize cross-linguistic variation (Croft & Poole, 2008; Wälchli & Cysouw, 2012). It has mostly been applied to datasets with a large number of languages to make claims about language classification, but recently, this methodology has been extended to study phenomena in a single language or small set of languages. Phenomena include prepositions (de Swart et al. 2012) and definite determiners (Bremmers et al. 2021). I explain how these studies bridge the gap between quantitative and formal approaches by using parallel corpus data as a starting point for formal analysis (van der Klis & Tellings, 2021).

Part 2. methodological extension to compositional phenomena
I propose to extend this methodology from the lexical/phrasal phenomena listed above, to study sentential phenomena in which multiple parts interact compositionally. The studies cited above base their MDS analyis on a single annotated feature: either the form itself, or a grammatical property of the form. The new methodology annotates multiple features corresponding to different properties of the construction at hand, and MDS allows to investigate the compositional interplay of these features. This way we can uncover cross-linguistic variation with respect to featural combinations. For example, in joint work (2021) with Henriëtte de Swart and Bernhard Wälchli, we find that there is much cross-linguistic variation in how languages express the NPI construction *not … until*, but that the combination of negation with various temporal connectives across languages is stable, as predicted by formal semantic work (de Swart 1996).

Part 3. case study: conditionals
As a case study to illustrate the extended methodology, I consider the domain of conditional sentences. Conditionals have a long tradition of formal analysis (Bhatt and Pancheva 2006), but this body of work has seen little interaction with quantitative or computational approaches.

Based on data in English, Dutch, French, and Spanish from the Europarl corpus (Koehn, 2005), I zoom in on so-called "extraposed" conditionals (Declerck & Reed 2011: §11.11), in English expressed with an expletive pronoun *it* and (typically) an evaluative adjective in the consequent clause ("It would be good if …"). The results show variation with respect to complementizer choice: where English and Dutch use their regular conditional connectives *if* and *als*, French and Spanish show variation between *que* and *si*.

EN It would be splendid **if** the EU institutions were to live up to this.
NL Het zou goed zijn **als** de EU-instellingen zich hieraan hielden.
SP Sería conveniente **que** las instituciones de la UE cumplieran con esto.
FR Il serait bon **que** les institutions de l'UE se montrent à la hauteur de ces principes.

In the subjunctive cases as illustrated above, English *that* is degraded, which I explain by a mismatch between the factivity of the construction (a combination of the predicate and the *that*-CP, see e.g. Schulz 2012) and the counterfactual modality of *would*. The cross-linguistic data reveal variation with respect to the factivity of complementizers across languages.

Finally, I comment on Declerck & Reed's (2011: 396ff.) formal analysis of the English construction, according to which a covert noun phrase gets extraposed, and the *if*-clause represents a conditional meaning. The cross-linguistic data suggest that there is a closer parallel between *if*- and *that*-clauses than can be expected on the basis of their analysis.

References

Bhatt, R. and R. Pancheva (2006). “Conditionals”. In M. Everaert and H. van Riemsdijk (eds.), *The Blackwell Companion to Syntax*, pp. 638–687. Oxford: Blackwell.

Bremmers, D., J. Liu, M. van der Klis, M. and B. Le Bruyn (2021). “Translation Mining: definiteness across languages. A reply to Jenks (2018)”. *Linguistic Inquiry*. To appear.

Croft, William and Keith T. Poole (2008). “Inferring universals from grammatical variation: Multidimensional scaling for typological analysis”. In: *Theoretical Linguistics* 34.1, pp. 1–37.

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

van der Klis, Martijn and Jos Tellings (2021). “Multidimensional scaling and linguistic theory”. Submitted, under review.

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

Schulz, Petra (2012). *Factivity: Its nature and acquisition*. De Gruyter.

de Swart, Henriëtte (1996). Meaning and use of *not… until*. *Journal of Semantics*, 13(3), 221–263.

de Swart, Henriëtte, Jos Tellings and Bernhard Wälchli (2021). “*Not ... until* across languages”. In progress.

de Swart, Peter, Hanne M. Eckhoff, and Olga Thomason (2012). “A Source of Variation: A Corpus-Based Study of the Choice between *apo* and *ek* in the NT Greek Gospels”. In: *Journal of Greek Linguistics* 12.1, pp. 161–187.

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