Mining for variables. Big data from small input.

[redacted]

With the advent of ever larger corpora and ever more powerful computers, big data have become the shiniest new toy in a wide number of scientific disciplines. Many (socio)linguists, however, continue to regard the computational handling of such data as "linguistics with some practically useful but theoretically irrelevant and obfuscating nerdy add-ons" (Spärck Jones 2007: 440). In this talk, we challenge this view with a case study which demonstrates the power of machine solutions to theoretical problems.

The research question addressed is "which elements in the grammar of Dutch are variable – to the extent that there exist 'competing' constructional alternatives for them –, and how can we find out in a theory-neutral way?". This issue has some relevance for sociolinguistics in general – socio-syntax typically focuses on a tiny set of recurrently studied syntactic variables (dative alternation, genitive alternation, particle placement, scrambling, ...) – but it takes on extra significance in the Low Countries, where almost *any* variation issue is steeped in ideology and conflict. Prior to the 2000s, variation in the syntax of Dutch was largely neglected because laymen were for the most part oblivious to it, and analysts did not (want to) believe there was any (Van Haver 1989; Taeldeman 1992). From the 2000s onwards, the increasing interest in syntactic variation revealed differences between Belgian and Netherlandic preferences, but these were acknowledged only in Belgian investigations (De Sutter et al. 2005; Grondelaers et al. 2008; Speelman & Geeraerts 2008), while Netherlandic studies continued to ignore them (Bouma & De Hoop 2008; Van Bergen & De Swart 2010; Vogels & Van Bergen).

It goes without saying that theory-neutral data represent a welcome addition to a research domain which is as sensitive as (syntactic) variation in Dutch. In order to identify all constructional alternations in Dutch, including less conspicuous or emergent loci of syntactic variability, we replicated the bottom-up technology pioneered in Bannard & Callison-Burch (2005), building on a parallel corpus of Dutch translations of the English subtitles to 6700 movies from the Open Subtitle component of Tiedemann's (2012) OPUS resource of freely accessible parallel corpora.

The statistical machine translation software Moses (Koehn et al. 2006) was used to identify plausible mappings between an English n-gram and its aligned Dutch equivalents in order to obtain Dutch *paraphrases*, i.e. stretches of interchangeable text that carry approximately the same meaning. We found 6124 paraphrase pairs which were between 2 and 7 words long. In spite of this limited size, 20.69 % of the pairs represented morpho-syntactic alternations instead of, for instance, idioms or multi-word units. The occurrence in our data of most of the identified syntactic variables in Dutch (*er*-variation, word order alternations, complementizer omission, etc.) validates our bottom-up approach, but we also found evidence for recurrent alternations we had *not* anticipated as interesting variables (notably tense variables and competing subordination strategies).

In light of these findings, we discuss some of the (obvious) pros but also some cons of computational support for sociolinguistics.