Dirk Speelman, Stefan Grondelaers & Benedict Szmrecsanyi

Manueel en automatisch, top-down en bottom-up: Is de Nederlandse syntaxis “lexicaler” dan de Vlaamse?

De standaardmethode voor de corpusgebaseerde studie van syntactische alternanties zoals *De koningin gaf Sneeuwwitje de appel* vs. *De koningin gaf de appel aan Sneeuwwitje* is logistische regressie: om de voorkeur voor één van die alternatieven te modelleren, wordt doorgaans gebruik gemaakt van een mengeling van automatisch en manueel gecodeerde predictoren. Daarbij vereisen in het bijzonder “high-level” predictoren van semantische of pragmatische aard (zoals, bijvoorbeeld, de “given-new” status van de nominale elementen) tijdsintensieve en vaak lastig objectiveerbare manuele codering, waardoor ze een bottleneck vormen voor de schaalbaarheid van de methoden.

De vooruitgang in Natural Language Processing en Artificial Intelligence heeft niet alleen het ontstaan van nieuwe analysetechnieken (zoals memory-based learning – Daelemans & Van den Bosch, 2005) gestimuleerd, maar ook van nieuwe methodes om semantische en pragmatische factoren computationeel te modelleren. Voorbeelden van die laatste evolutie zijn de automatische semantische classificatie van cruciale woorden op basis van semantische vectoren (distributionele modellen – zie Levshina & Heylen 2014) en de automatische bepaling van de contextuele voorspelbaarheid van cruciale woorden op basis van language models (Van den Bosch & Berck, 2009). Bovendien hebben sommige van de nieuwe technieken (memory-based learning) geen higher-level factoren nodig om syntactische variatie te modelleren, maar baseren ze zich louter op low-level informatie (ongeclassificeerde lexemen). Kortom, de nieuwe methodes hebben het potentieel om de studie van syntactische variatie substantieel op te schalen.

Met name voor de studie van de *syntactische* dimensie van het verschil tussen Belgisch- en Nederlands-Nederlands is de mogelijkheid om meer predictoren in de studie van een veel groter aantal syntactische kandidaat-variabelen te betrekken erg nuttig. Maar er zijn ook belangrijke theoretische voordelen. Omdat we nu de mogelijkheid hebben om syntactische variatie zowel op basis van high-level als low-level predictoren te modelleren, kunnen we ons een beter beeld vormen van de relatieve impact van die twee predictortypes in Belgisch- en Nederlands-Nederlands, en van hun granulariteitsniveau.

In deze presentatie bespreken we op basis van Nederlandse en Vlaamse corpusdata verschillende analyses van de aan- of afwezigheid van *er* in bepalingsinitiële zinnen zoals *In de asbak lag (er) een peuk*. Methodologisch gezien gebruiken we een manueel geannoteerde dataset als referentiepunt om te bepalen hoe succesvol we de alternantie (semi-)automatisch kunnen modelleren. Inhoudelijk bekijken we hoe de modelleerbaarheid van de Nederlandse en de Vlaamse data verschilt. De Nederlandse gegevens laten zich weliswaar beter modelleren op basis van louter “low-level” factoren, maar een higher-level factor zoals de voorspelbaarheid van het onderwerp blijkt in de Nederlandse data relatief meer bij te dragen tot de verklarende kracht van de modellen dan in de Vlaamse data. Omgekeerd blijkt een higher-level factor zoals de semantische klasse van het hoofd van het onderwerp een krachtiger predictor voor de Vlaamse dan voor de Nederlandse data.

Hoewel onze bevindingen voorlopig op slechts één variabele gebaseerd zijn, bewijzen ze dat Belgisch en Nederlands Nederlands anders dan vermoed een nogal verschillende syntactische motor hebben, en dat computationele verrijking een onontbeerlijk uitbreiding is voor syntactische-variatieonderzoek.

Referenties

Broekhuis, H., Corver, N., & Vos, R. 2015. Syntax of Duch: Verbs and verb phrases, volume 1. Amsterdam: Amsterdam University Press.

Daelemans, Walter en Van den Bosch, Antal. 2005. Memory-Based Language Processing. Cambridge, UK: Cambridge University Press.

Levshina, Natalia en Kris Heylen. 2014. A Radically Data-driven Construction Grammar: Experiments with Dutch causative constructions. In: Boogaart, Ronny, Timothy Colleman and Gijsbert Rutten (Eds.), Constructions in Germanic – Extending the scope, 17-–46. Berlin: Mouton de Gruyter. doi:10.1515/9783110366273.17

Van den Bosch, Antal en Berck, Peter. 2009. Memory-Based Machine Translation and Language Modeling. The Prague Bulletin of Mathematical Linguistics 91, 17–26.