Syntactische Annotatie van Hele Grote Corpora in *LASSY*

In het LASSY project beogen we de beschikbaarheid van syntactisch geannoteerde corpora flink uit te breiden. Naast een uitbreiding van handmatig gecorrigeerde syntactische annotaties, gaan we op grote schaal automatisch toegekende syntactische annotaties opleveren. In totaal zal een syntactisch geannoteerd corpus van 500 miljoen woorden ontstaan.

De beschikbaarheid van handmatig gecorrigeerd syntactisch geannoteerd materiaal is cruciaal voor het ontwikkelen, afstemmen en evalueren van een grote verscheidenheid aan toepassingen binnen de natuurlijke-taalverwerking.

Binnen Lassy breiden we het in D-Coi syntactisch geannoteerde corpus uit tot een corpus van één miljoen woorden. Deze syntactische annotaties zijn volledig handmatig gecorrigeerd, en bevatten informatie over de woordsoort van elk woord, het lemma van elk woord, en een dependentiestructuur zoals die binnen het Corpus Gesproken Nederlands is voorgesteld, en verder werd ontwikkeld binnen D-Coi.

Het gebruik van automatisch toegekende syntactische annotaties is tot dusver minder verbreid. Recentelijk is er flinke vooruitgang geboekt op het gebied van de automatisch syntactische analyse van het Nederlands. Binnen LASSY gebruiken we de vrij beschikbare Alpino parser om een corpus van zo'n 500 miljoen woorden syntactisch te annoteren. Wij verwachten hierbij dat het nadeel van de kwaliteit van het materiaal (de kwaliteit loopt natuurlijk terug, omdat de handmatige correctie achterwege blijft) voor een groot aantal toepassingen ruimschoots zal worden gecompenseerd door de veel grotere hoeveelheid materiaal. Dit geldt met name voor toepassingen in informatie extractie, lexicale acquisitie, het afleiden van ontologische informatie en dergelijke. Het is bijvoorbeeld de ervaring van gerelateerde projecten op het gebied van vraag-antwoord systemen, dat het goed mogelijk is betrouwbare feitelijke informatie uit automatisch syntactisch geannoteerde corpora te extraheren (wat betekent welke afkorting, wat is de hoofdstad van welk land, wie bekleedt welke publieke functie, enzovoort). Deze betrouwbaarheid neemt toe in vergelijking tot technieken die de syntactische analyse veronachtzamen en bijvoorbeeld met reguliere expressies direct in de tekst vergelijkbare informatie proberen te achterhalen.

Andere recente onderzoeken laten zien hoe hele grote syntactisch geannoteerde corpora kunnen worden gebruikt om automatisch af te leiden welke woorden semantisch vergelijkbaar zijn, of om beter inzicht in de woordvolgorderegels van het Nederlands te krijgen.

De in het project op te leveren zeer grote hoeveeldheid syntactisch geannoteerd materiaal zal, om echt bruikbaar te zijn, op een goede wijze te doorzoeken en te bewerken moeten zijn. Per zin zal de syntactische analyse als een XML-bestand beschikbaar zijn. Binnen het project zal veel aandacht besteed worden om efficiënt binnen enorme collecties gecomprimeerde XML-bestanden te kunnen zoeken, waarbij minimaal gestreefd wordt naar expressiviteit zoals die binnen XPATH en Xquery beschikbaar is.

Het project zal zeer binnenkort van start gaan. De projectpartners zijn: Universiteit van Groningen (Gertjan van Noord en Gosse Bouma), en de Katholieke Universiteit Leuven (Frank van Eynde en Ineke Schuurman).