DPC, een nieuw vertaalcorpus

Hans Paulussen

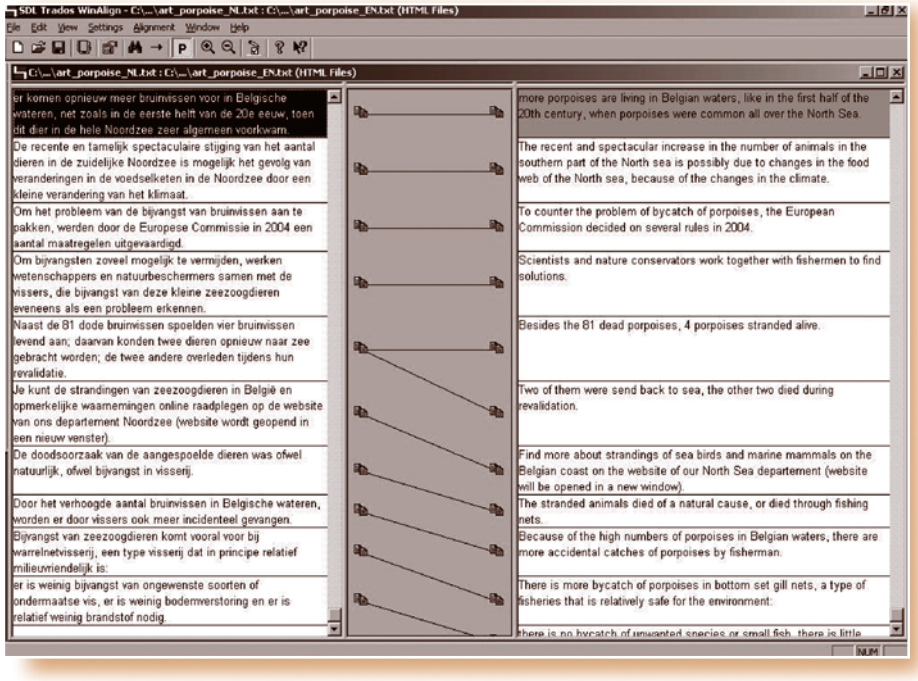
Onlangs werd een project opgestart rond een nieuw vertaalcorpus, dat de naam DPC (Dutch Parallel Corpus) meekreeg. Het project, dat onder leiding staat van professor Piet Desmet, is toevertrouwd aan de goede zorgen van een onderzoeksteam van de K.U.Leuven Campus Kortrijk en de Hogeschool Gent Departement Vertaalkunde. Ook onderzoekspartners van universiteiten en hogescholen uit Nederland en Vlaanderen zijn erbij betrokken, net als de gebruikersgroep, die bestaat uit experts uit de academische en industriële wereld.

Maar voor we inzoomen op DPC, lichten we eerst toe wat een corpus is. Een corpus is een verzameling van elektronisch opgeslagen teksten die zodanig gestructureerd zijn dat je er op een eenvoudige manier allerlei zaken in kunt opsporen. Dat kan gaan van taalkundige aspecten (woorden, woordsoorten, zinsconstructies) tot stilistische kenmerken, die dan eerder bij literatuur en forensisch onderzoek thuishoren. Het volstaat dus niet gewoonweg een elektronische versie van een tekst te hebben. Die elektronische versie moet ook *ontsluitbaar* zijn, met andere woorden gemakkelijk toegankelijk. Heel wat tekstformaten zijn namelijk gesloten *proprietaire* formaten, alleen maar toegankelijk via het programma (gewoonlijk een tekstverwerker) waarmee die tekst werd aangemaakt. Het formaat moet *open* zijn, zodat je dat met eender welk programma kunt lezen. Het is ook noodzakelijk dat de verzameling teksten op een of andere manier gemeenschappelijke kenmerken hebben, al kan dat op verschillende manieren gedefinieerd worden. Bij voorkeur gelijkaardige teksten, omdat je moeilijk appels met peren kunt vergelijken. Zo bestaan er onder andere al heel wat corpora met krantenmateriaal. Maar vanuit taalkundig standpunt komen alle soorten teksten in aanmerking. Eigenlijk zou een corpus een representatief staal van de taal moeten zijn. In werkelijkheid ligt het allemaal wel een beetje moeilijker.

Hoe dan ook, tekstcorpora zijn in de taalkunde tegenwoordig een alledaags begrip. Zij vormen de toetssteen voor heel wat lexicologische en stilistische theorieën rond taal, soms op het gevaar af dat men het corpus voor heilig verklaart. Maar ook het corpus moeten we kunnen relativeren. Corpora vormen het bronnenmateriaal om bepaalde taalkundige fenomenen te staven met wat J. Sinclair "real language usage" noemde. Toen er nog geen corpora waren, bestonden er voor taalkundigen slechts twee mogelijkheden om aan voorbeelden te geraken: ofwel vonden ze zelf voorbeelden uit op basis van hun intuïtieve talenkennis, ofwel wonnen ze taalkundige informatie in bij proefpersonen binnen een nauwkeurig opgezette testomgeving. Beide methodes blijven een belangrijke vorm om aan taalvoorbeelden te geraken, maar zijn beperkt, vooral wat de natuurlijkheid van zulke introspectieve methode betreft. Corpusvoorbeelden daarentegen geven duidelijk aan wat er effectief gebruikt wordt. Maar veel hangt af van het representatief karakter van het corpus. Zo zullen bepaalde woorden alleen maar in een bepaald type tekst voorkomen. Daarom blijven introspectieve methodes en de corpustaalkunde complementaire methodes.

De eerste corpora waren alleen beschikbaar voor het Engels, maar het jongste decennium, en onder invloed van de verbeterde informaticatechnologie en het internet, is daar veel verandering in gekomen. Nu vind je voor zo goed als iedere taal een corpus. (We hebben het hier dan wel over geschreven corpora. Er zijn ook corpora voor gesproken taal, maar die laten we hier even buiten beschouwing.) Eerst waren er corpora voor talen, daarna voor taalvariëteiten (van Brits en Amerikaans Engels, tot de Engelse varianten van Australië tot Singapore), en uiteindelijk kwamen er ook meertalige corpora, en zo belanden we bij de parallelle corpora.

Eigenlijk zijn er twee soorten meertalige corpora. Ofwel gebruik je een verzameling teksten over een bepaald onderwerp in een bepaalde taal, en je verbindt het met een gelijkaardige verzameling teksten in een andere taal. Het gaat dus niet om vertaalde teksten, maar om authentieke teksten in twee (of meerdere) talen. Zo'n corpus noemt men in het Engels een *comparable* corpus: een *vergelijkbaar* corpus. Daarnaast hebben we ook vertaalcorpora of parallelle corpora, waarvan DPC een voorbeeld is. Hier gaat het om een verzameling vertaalde teksten in twee of meerdere talen die met elkaar gealigneerd zijn op paragraaf-, zins- of woordniveau. DPC beoogt een alignering op zinsniveau. Dat betekent dat je de corresponderende zinnen in twee talen automatisch verbindt, zodat je bij het opvragen van een Nederlandse zin automatisch de equivalente Franstalige zin meekrijgt. Dit soort alignering biedt heel wat toepassingen voor taalkunde, vertaalkunde en taaltechnologie.



De motivatie om een vertaalcorpus te ontwikkelen heeft te maken met de vraag naar dergelijke corpora, en de tot nog toe beperkte beschikbaarheid ervan. Gealigneerde parallelle corpora vormen noodzakelijk bronmateriaal voor een groot aantal multitalige toepassingen, zoals machinevertaling (in het bijzonder corpusgebaseerde machinevertaling zoals statische en

example-based machine translation), computerondersteunde vertaaltools, informatie-extractie, multilinguale terminologie-extractie, en computerondersteund talenonderwijs.

De financiering van het DPC-project gebeurt door de Nederlandse Taalunie in het kader van het STEVIN-programma (Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands), een meerjarig onderzoeks- en stimuleringsprogramma voor Nederlandstalige taal- en spraaktechnologie. Op dit ogenblik zijn er slechts weinig kwaliteitsvolle parallelle corpora met Nederlands als centrale taal beschikbaar, en hun beschikbaarheid voor de onderzoeksgemeenschap wordt bemoeilijkt door auteursrechterlijke restricties. Daarom was de aanmaak van een parallel corpus een van de prioriteiten van het STEVIN-programma.

Voor het DPC-project zal een kwaliteitsvol zinsgealigneerd parallel corpus van 10 miljoen woorden aangemaakt worden voor de talenparen Nederlands-Engels en Nederlands-Frans. Het corpus zal bidirectioneel zijn (Nederlands als brontaal en doeltaal), zodat het ook kan gebruikt worden als een *comparable* corpus (waarbij oorspronkelijk in het Nederlands geschreven teksten kunnen vergeleken worden met teksten vertaald naar het Nederlands). Een gedeelte van het corpus zal drietalig zijn, waarbij Nederlandse teksten vertalingen hebben naar het Engels én het Frans. Het corpus wordt verrijkt met taalkundige annotaties, meer bepaald de aanduiding van lemmata en woordsoorten.

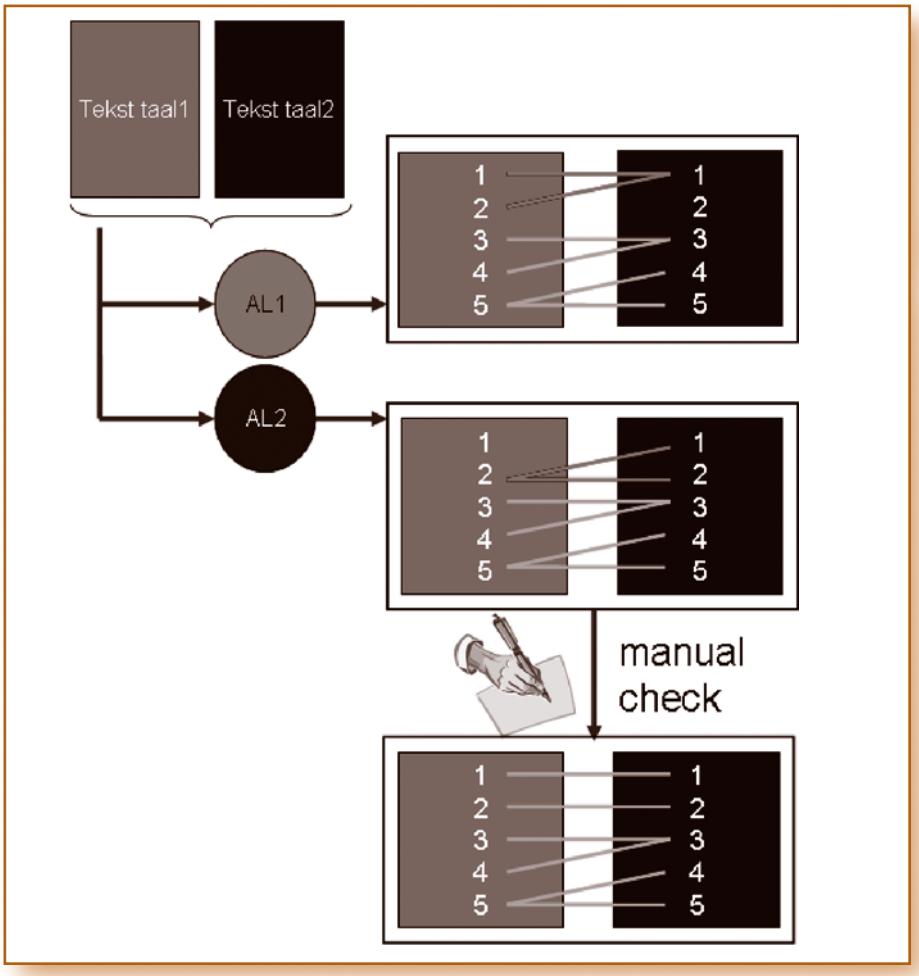
Om de kwaliteit van het corpus te waarborgen, evenals de multilinguale beschikbaarheid voor de gehele onderzoeksgemeenschap, zal iedere stap in de aanmaak, het structureren en het annoteren gevalideerd worden door een gebruikersgroep van specialisten in de taalkunde en taaltechnologie. Aangezien het Nederlands de *scharniertaal* is van het corpus, wordt ook nauw samengewerkt met de onderzoekers van het D-COI project, die een pilootcorpus aanmaakten van 50 miljoen woorden van hedendaags geschreven Nederlands.

De bedoeling is om een zo evenwichtig mogelijk vertaalcorpus samen te stellen. Daarom zijn verschillende tekstsoorten zo belangrijk. Er zullen zowel teksten uit fictie als non-fictie in voorkomen. De tweede soort wordt verder ingedeeld in journalistieke, zakelijke, technische en administratieve teksten, zodat een ruime spreiding van tekstsoorten wordt bekomen.

Type tekst	Sector
journalistiek	kranten & tijdschriften
essayistisch & literair	literaire uitgevers
zakelijk	bank & verzekeringen
technisch	softwarebedrijven, medische sector
administratief	overheid

Het aligneren van teksten op zinsniveau gebeurt door een aligneringsprogramma. De kwaliteit van de alignering hangt sterk af van de kwaliteit van het programma en de soorten vertaalde teksten, want in dat laatste bestaat er ook heel wat variatie. Om de kwaliteit van de alignering te controleren, wordt een bepaald gedeelte van het gealigneerde corpus met de hand gecontroleerd. Daarnaast wordt ook een semi-automatische controle uitgevoerd. Hierbij wordt het corpus een tweede keer gealigneerd met een tweede aligneringsprogramma dat een volledig andere algoritme gebruikt. De veronderstelling is dan dat alleen nog moet gecon-

troleerd worden waar de output van de twee aligneringsprogamma's met elkaar verschillen, wat de controle van de alignering aanzienlijk verlicht. Eens het corpus is afgewerkt, zal het beschikbaar worden gesteld voor de hele onderzoeksgemeenschap, en dit via de TST-centrale, een afdeling van het INL die instaat voor de distributie van dergelijke corpora.



De K.U.Leuven heeft al heel wat ervaring opgebouwd op het gebied van corpuscompilatie en corpusonderzoek. Zo werd onder meer meegewerkt aan de ontwikkeling van het CGN (Corpus Gesproken Nederlands) en werden er corpora ontwikkeld voor het Frans (bv. ELICOP, een corpus gesproken Frans) en andere talen. Het allereerste woordenboek Arabisch-Nederlands gebaseerd op een elektronisch corpus, werd ook al ontwikkeld aan de K.U.Leuven. Het DPC-project hoopt bij te dragen aan verdere corpustoepassingen die gebaseerd zijn op parallelle corpora.

Meer informatie op de volgende (uiteraard drietalige) website:
http://www.kuleuven-kortrijk.be/DPC