Eenwoordsconstituenten in GrETEL – Jan Odijk

(Universiteit Utrecht)

Beste Frank,

In de special issue van het taalkundig tijdschrift *Lingua* over taalkundig onderzoek in de CLARIN infrastructuur onderzoek je samen met enkele collega's de congruentie in getal tussen het onderwerp en het naamwoordelijk deel van het gezegde, waarbij je de treebank Lassy-Klein als datamateriaal gebruikt en GrETEL als toepassing om deze treebank te bevragen (Van Eynde 2016).

Je extraheert uit de treebank alle zinnen waarin zowel een onderwerp gemarkeerd voor getal als een naamwoordelijk deel van het gezegde gemarkeerd voor getal voorkomen in combinatie met een werkwoord dat als hoofd van de zin fungeert. Je merkt op dat in de LASSY-treebank constituenten die uit één woord bestaan alleen een knoop bevatten voor dat woord als woord maar niet voor dat woord als constituent. Dit is een gevolg van de afspraken die er gemaakt zijn omtrent de precieze vorm van treebanks voor het Nederlands, oorspronkelijk voor de treebank van het Corpus Gesproken Nederlands en later uitgebreid en aangepast voor de Lassy-treebank.

Omdat zowel het onderwerp als het naamwoordelijk deel van het gezegde een constituent kan zijn die een of meerdere woorden bevat, heb je daarom vier query's moeten schrijven om de relevante data te extraheren.

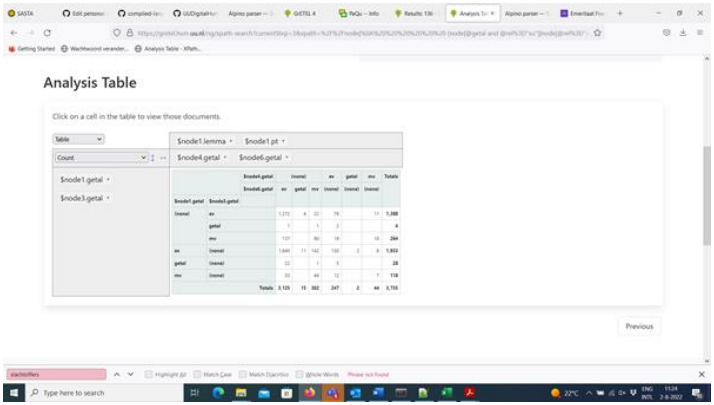
In werkoverleg over de verdere ontwikkeling van GrETEL heb je daarom ook gesuggereerd dat de treebanks beter aangepast kunnen worden zodat er ook voor eenwoordsconstituenten een knoop is voor dat woord als constituent. Dan volstaat één enkele query. Ik ben zelf dit probleem voor andere verschijnselen ook meerdere keren tegengekomen, en was (en ben) het met je eens dat aanpassing van de treebanks nuttig zou zijn. Helaas paste werk daaraan niet in de door Utrecht voorgenomen plannen, en kreeg het dus lage prioriteit, en er is dan ook nog niet van gekomen.

Maar er is een alternatief. De queries die jij gebruikte kunnen wel degelijk in een enkele query geformuleerd worden, en wel als volgt. We zoeken voor het onderwerp naar een knoop die gemarkeerd is voor getal en relatie *su* heeft, of naar een knoop die gemarkeerd is voor getal maar het hoofd is van een knoop die de relatie *su* heeft. Dat kan als volgt geformuleerd worden in Xpath:

(1) node[@getal and
@rel="su"]|node[@rel="su"]/node[@getal and
@rel="hd"]

De delen voor en na de verticale bar leveren nodesets op die verenigd worden in een nodeset door de |-operator. We kunnen dit analoog doen voor het naamwoordelijk deel van het gezegde door *su* overal te vervangen door *predc*. Gecombineerd krijgen we dan de query in (2):

(2) //node[(node[@getal and @rel="su"]|node[@rel="su"]/node[@getal and @rel="hd"]) and (node[@getal and @rel="predc"]|node[@rel="predc"]/node[@getal and @rel="hd"]) and node[@rel="hd" and @pt="ww"]]



Figuur 1. Screenshot van de draaitabel van de GrETEL analysecomponent voor het attribuut getal voor de knopen node1 (subjectwoord), node3 (hoofd van de subjectconstituent), node4 (predc-woord) en node6 (hoofd van de predc-constituent).

Wanneer we deze query uitvoeren op de Lassy-Klein Treebank (via deze link), krijgen we precies de 3735 resultaten die jouw 4 query's ook vonden.[1]

In de analysecomponent van GrETEL die door het werk in Utrecht aan GrETEL is toegevoegd (Odijk et al. 2018) kunnen

[1] Hiervan worden er maar 500 in GrETEL 4 getoond, maar de telling geeft aan dat er 3.735 hits zijn

we nu het getalsattribuut van iedere knoop selecteren en in een draaitabel zetten, waardoor we een compact overzicht krijgen van de situatie.

Een screenshot hiervan is weergegeven in Figuur 1, en een tekstuele weergave (geëxtraheerd uit GrETEL) is gegeven in Tabel 1, waarin *ev*=enkelvoud, *mv*=meervoud, en *getal*=ongespecificeerd voor getal. De cijfers die we hier vinden corresponderen precies met de cijfers die in Tabel 3 van jouw publicatie in Lingua staan (Van Eynde 2016: 107), zij het dat jij bepaalde cijfers daar geaggregeerd hebt.

Natuurlijk is het niet gewenst de complexe query om een woord zowel als eenwoordsconstituent als als hoofd van een meerwoordige constituent te vangen iedere keer weer opnieuw moeten schrijven. Het zou wenselijk zijn het macro-mechanisme oorspronkelijk geïntroduceerd in DACT en overgenomen in PaQu en GrETEL uit te breiden met een parametermechanisme om een compacte query hiervoor te creëren.

En het ligt dan ook voor de hand om de query-by-example interface van GrETEL uit te breiden met een optie die een enkel woord in de voorbeeldzin omzet in een query die zowel naar een enkel woord als naar een constituent met dit woord als hoofd zoekt, en andersom, dat een constituent in de voorbeeldzin ook leidt tot het zoeken naar een enkel woord. En misschien moet dat zelfs wel de defaultoptie worden, omdat dit veel vaker gewenst is dan apart zoeken voor eenwoordsconstituenten en meerwoordsconstituenten.

$node1.getal	$node3.getal	$node4.getal	$node6.getal	Count
(none)	Ev	(none)	ev	1272
(none)	Ev	(none)	getal	4
(none)	Ev	(none)	mv	22
(none)	Ev	ev	(none)	79
(none)	Ev	mv	(none)	11
(none)	Getal	(none)	ev	1
(none)	Getal	(none)	mv	1
(none)	Getal	ev	(none)	2
(none)	Mv	(none)	ev	137
(none)	Mv	(none)	mv	90
(none)	Mv	ev	(none)	19
(none)	Mv	mv	(none)	18
ev	(none)	(none)	ev	1640
ev	(none)	(none)	getal	11
ev	(none)	(none)	mv	142
ev	(none)	ev	(none)	130
ev	(none)	getal	(none)	2

ev	(none)	mv	(none)	8
getal	(none)	(none)	Ev	22
getal	(none)	(none)	Mv	1
getal	(none)	ev	(none)	5
mv	(none)	(none)	Ev	53
mv	(none)	(none)	Mv	46
mv	(none)	ev	(none)	12
mv	*(none)*	*mv*	*(none)*	*7*

Tabel 2. Extractie uit GrETEL van de data waarop de draaitabel in Figuur 1 is gebaseerd

Hoewel dit alternatief het mogelijk maakt jouw vier aparte queries in een query samen te vatten en de resultaten ervan in de GrETEL 4 analysecomponent gezamenlijk te analyseren, zijn er nog andere configuraties, die jij buiten beschouwing hebt gelaten, maar waar we eigenlijk wel rekening mee moeten houden.

In de eerste plaats betreft dat zogenaamde indexknopen, knopen zonder *word* of *cat* attribuut maar met een *index* attribuut en een antecedentknoop met de volledige eigenschappen. Hiervoor is in PaQu al een oplossing geïmplementeerd, namelijk door als optie aan te bieden dat gezocht wordt in een treebank waarin de indexknopen

geëxpandeerd zijn tot hun antecedent en diens kinderen.[2] Het is wenselijk deze oplossing ook in GrETEL te implementeren.

In de tweede plaats betreft dat coördinatie. Jij laat in jouw artikel gecoördineerde structuren buiten beschouwing omdat die geen hoofd hebben (Van Eynde 2016:107, voetnoot 5). Ik denk dat het zinvol is hier ook iets speciaals te voorzien, en daarbij de strategie van PaQu te volgen, die eigenlijk aanneemt dat ieder van de kinderen (zowel de conjuncten als de voegwoorden) van een gecoördineerde structuur een hoofd is [Odijk et al. 2017:287]. Maar dat is iets voor de toekomst.

Zoals je ziet, er is sinds jouw artikel uit 2016 flinke vooruitgang geboekt om het gebruik van treebanks bij taalkundig onderzoek nog gebruikersvriendelijker te maken, maar tegelijkertijd is er ook nog veel werk aan de winkel. Maar niet voor jou, want jij gaat met een welverdiend emeritaat!

Referenties

[Odijk et al. 2017] Jan Odijk, Gertjan van Noord, Peter Kleiweg & Erik Tjong Kim Sang (2017). The Parse and Query (PaQu) Application. In: Jan Odijk, Arjan van Hessen (eds*.), CLARIN in the low countries*. London: Ubiquity Press.

[Odijk et al. 2018] Jan Odijk, Martijn van der Klis & Sheean Spoel (2018). Extensions to the GrETEL Treebank Query Application. In E. Bejcek (Ed.), *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)* (]pp. 46-55). Prague, Czech Republic: Charles University.

[2] Zie https://paqu.let.rug.nl:8068/info.html#expanded

[Van Eynde et al. 2016] Frank Van Eynde, Liesbeth Augustinus & Vincent Vandeghinste (2016). Number agreement in copular constructions: A treebank-based investigation, in 'Linguistic Research using CLARIN', *Lingua Special Issue on CLARIN* Vol 178, pp. 1-4. [DOI][URL]

Ken je dat verhaaltje… – GPT-2

Ken je dat verhaaltje van die ene prof computerlinguïstiek? Mijn woordengebruik is de regel dat ik de tekstkeuze verleg. Als het maar foutloos is. In spoilerlip: Nee dat heeft geen invloed op de kwaliteit. Ik spreek altijd mijn begrip uit dat ik denk, ik heb een fout. En die ene vinger op de toets en de twijfel is weg. Waarom zijn we daar niet altijd bewust mee bezig? haha ik ben niet op zoek naar wat jullie doen. de logica, het principe... ja alles wat een beetje ingewikkeld of vaag is, wordt enorm gecompliceerd en onbelangrijk. LOL, ik vindt dat als iemand praat, dat de spreker is. Maar wanneer jij zegt dat de tekst ook mooi is en mooi in elkaar zit, dan lijkt het alsof de spreker simpel gezegd uit een onverwachte hoek komt. Wait?! Ik krijg echt de kriebels van alle praatwoorden in een kleine tien minuten. Ik zou zeggen: laat iedereen maar praten, maar dat ga ik nooit zo doen. Ik vind zulke taal ook wel leuk, maar als je daar niet heel veel mee doet, valt het best wel mee en lijkt het dus alsof je maar aan het lopen bent. Hmm, je kan wel proberen met een verwijzing naar het beknopte genre synoniemen en die hebben het eigenlijk ook voor mij best leuk. Je verzint een titel of een dan. En stuurt de "kop" van de gelezen tekst door naar een vertrouwd generalismediagram of niet?