Naar een semi-automatische analyse van tekststructuren in corpora

Markus Egg & Gisela Redeker [1]

In deze bijdrage presenteren wij eerste verkennende bevindingen van een project voor de semi-automatische analyse van tekststructuren in syntactisch geannoteerde corpora. We gebruiken de syntactische informatie om een in eerste instantie zeer ondergespecificeerde structuur van locale relaties op te bouwen. Vervolgens worden mede op basis van corpusanalyse constraints geformuleerd om deze structuur nader te specificeren. Om de gemodelleerde structuren te toetsen, zal de corpusannotatie verrijkt worden met een handmatige annotatie. In een pilot study bleek de syntactische informatie gemiddeld ca. 40 % van de retorische neven-/onderschikkingsrelaties te kunnen voorspellen.

1 Inleiding

Een van de centrale vragen in onderzoek naar tekststructuur betreft de mate waarin en manier waarop linguïstische informatie (zonder beroep op niet-linguïstische kennis) deze structuur bepaalt. Om deze vraag te kunnen beantwoorden is een representatie van de beschikbare linguïstische informatie nodig en een goed uitgewerkt analysemodel voor tekststructuren, zoals *Rhetorical Structure Theory (RST)* (Mann & Thompson, 1988). Sanders en van Wijk (1996) bijvoorbeeld hanteren in hun tekstanalysesysteem *PISA* een vereenvoudigde variant van RST verrijkt met genrespecifieke informatie op lexicaal, syntactisch en tekstniveau.

Idealiter zou het voor het modelleren van tekststructuren van een complete syntactische en semantische analyse op zinsniveau uitgegaan moeten worden. De vorderingen in deze richting in de computationele taalkunde zijn bemoedigend en er komen steeds omvangrijkere corpora beschikbaar die part-of-speech tagging en deels ook syntactische annotaties bieden, zoals het Corpus Gesproken Nederlands (www.tst.inl.nl.cgn.htm), PAROLE corpus (parole.inl.nl) en de Alpino Treebank (van der Beek, Bouma, & van Noord, 2002). Maar wat betreft de semantiek zijn de mogelijkheden voor automatische analyse nog beperkt. Een complete modellering van discourse-structuren, zoals in Asher en Lascarides (2003) voorgesteld, vereist redenaties met wereldkennis en pragmatische kennis die ver buiten de huidige mogelijkheden van automatische analyse liggen.

Wij beperken ons daarom in eerste instantie tot zuiver uit syntactische analyse herleidbare informatie en onderzoeken in hoeverre alleen op basis van deze informatie hiërarchische discourse-structuren bepaald kunnen worden. Om deze structuren te beschrijven, maken wij gebruik van een methode voor de representatie van ondergespecificeerde semantische structuren (Egg, Koller, & Niehren, 2001). De bijdrage van syntactische informatie, zoals inbedding, onderschikking of parallellie, wordt geformuleerd als een 'constraint', dwz. als een beperking van de set van toelaatbare discourse-structuren. Als leidraad voor de formulering van constraints dient een op RST gebaseerde manuele annotatie van de tekststructuren.

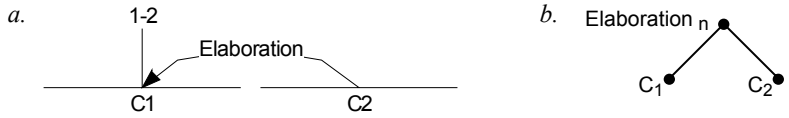
In deze bijdrage zullen wij aan de hand van voorbeelden en resultaten van een beperkt corpusonderzoek demonstreren hoe de exploitatie van syntactische informatie kan bijdragen tot de bepaling van hiërarchische discourse-structuren. We zullen

mogelijkheden bespreken om tot verdere specificatie te komen met behulp van beperkte semantische informatie, bijvoorbeeld op basis van eenvoudige ontologieën.

2 Hiërarchische discourse-structuren

Voor de beschrijving en modellering van hiërarchische discourse-structuren worden in de literatuur verschillende systemen van coherentierelaties voorgesteld (Marcu, 1997, 2000; Redeker, 2000; Carlson, Marcu, & Okurowski, 2003; Soricut & Marcu, 2003, Asher & Lascarides, 2003). Het meest gedetailleerd uitgewerkte analysesysteem is ontwikkeld door Marcu (1997, Carlson & Marcu, 2001) op basis van *Rhetorical Structure Theory* (RST) (Mann & Thompson, 1988). In RST worden coherentierelaties gedefinieerd in termen van discourse-functies. De analist wordt geacht de contextueel meest plausibele interpretatie te kiezen. RST is empirisch zeer succesvol gebleken voor uiteenlopende tekstgenres en talen (Mann, Matthiessen, & Thompson, 1992; Abelen, Redeker, & Thompson, 1993; Reitter & Stede, 2003; Stede 2004).

In deze bijdrage hanteren wij, waar niet anders vermeld, de 'klassieke' RST-relatiedefinities. Voor de representatie van de hiërarchische structuur vertalen wij de RST-diagrammen naar *binaire boomstructuren.* In RST wordt een ELABORATION-relatie waarbij C_2 nadere informatie toevoegt aan (een element in) C_1 gerepresenteerd door de zgn. *satelliet* C_2 met een pijl te verbinden aan de *nucleus* C_1 (zie figuur 1a). In de binaire boomrepresentatie introduceert the ELABORATION-relatie een knoop boven C_1 en C_2, waarbij C_1 en C_2 de *argumenten* van de *functor* ELABORATION zijn (zie figuur 1b). Het onderscheid tussen nucleus en satelliet wordt gemarkeerd door een index (n of s) die de status van de linker dochterknoop (hier: C_1) van de relatie aangeeft.



Figuur 1: ELABORATION-relatie (a) in RST en (b) als binaire boomstructuur

Relaties tussen gelijkwaardige segmenten (de zgn. *multinucleaire relaties* in RST, bijv. LIST of CONJUNCTION) worden, als zij meer dan twee segmenten (*nuclei*) bevatten, opgesplitst in binaire deelrelaties (zie figuur 2 in paragraaf 2.1).

2.1 De tekstuele realisatie van discourse-structuren

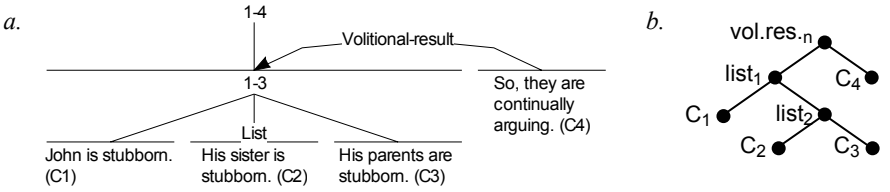
Discourse-relaties kunnen expliciet gemarkeerd zijn door conjuncties zoals *dus, maar* of *hoewel* (of Engels *so* in voorbeeld (1)), maar kunnen ook impliciet zijn, zoals de relaties tussen C_1, C_2 en C_3 in voorbeeld (1):

(1) John is stubborn (C_1). His sister is stubborn (C_2). His parents are stubborn (C_3). So, they are continually arguing (C_4).

Dit veelgeciteerde voorbeeld (uit Webber 2004) illustreert drie belangrijke eigenschappen van discourse-structuren, die in deze paragraaf aan de orde zullen komen:

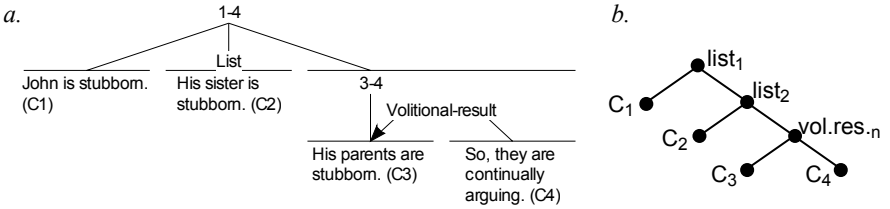
- interactie van discourse-structuur en referentie
- ambiguïteit m.b.t. het bereik van discourse-connectieven
- de Right Frontier Constraint (RFC).

In de meest plausibele interpretatie van dit fragment is C_4 het resultaat van C_1-C_3 (zie figuur 2), waarbij *they* in C_4 naar alle vier personen verwijst en niet bijvoorbeeld alleen naar *his parents*.



Figuur 2: RST-analyse (a) en binaire boomstructuur (b) voor voorbeeld (1)

In de minder geprefereerde lezing waar *they* alleen naar *his parents* verwijst, zou C_4 alleen het resultaat van C_3 (en niet ook van C_1 en C_2) zijn. De relatie C_3-C_4 is dan ingebed in een LIST-relatie met de nuclei C_1, C_2, en C_3-C_4 (zie figuur 3).

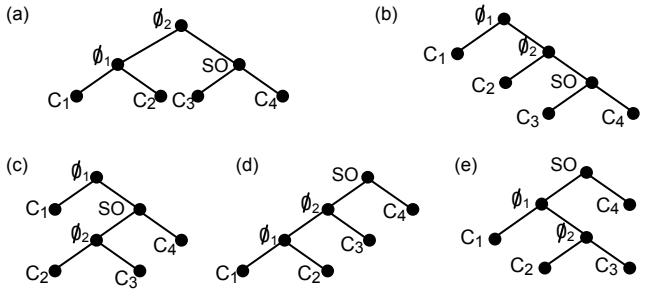


Figuur 3: Alternatieve RST-analyse (a) en binaire boomstructuur (b) voor voorbeeld (1)

Voorbeeld (1) heeft geen lezing waar C_4 het resultaat van C_2-C_3 is, maar dat is juist de geprefereerde lezing voor het iets veranderde voorbeeld (2):

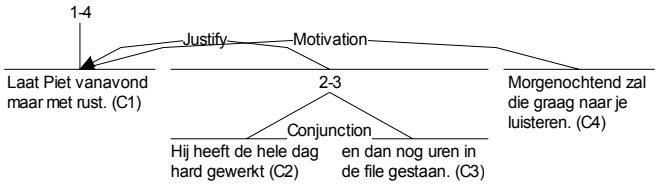
(2) John is spending a few days with his family (C_1). His sister is stubborn (C_2). His parents are stubborn (C_3). So, they are continually arguing (C_4).

Zonder naar de inhoud van de discourse-segmenten te kijken, is het dus niet duidelijk of het connectief *so* (*dus*) in voorbeelden als (1) en (2) C_4 aan C_3 koppelt, aan C_2-C_3, of aan C_1-C_3. Doordat ook de onderlinge relatie van C_1, C_2 en C_3 nog kan variëren, zijn in feite vijf structuren mogelijk (zie figuur (4); '∅' staat voor impliciete relaties zonder connectief).



Figuur 4: Mogelijke discourse-structuren van voorbeelden als (1) en (2)

Zuiver combinatorisch zouden met vier segmenten nog meer structuren mogelijk zijn, bijvoorbeeld structuren waar C_4 alleen aan C_1 of aan C_1-C_2 gerelateerd is, maar niet aan C_3, of waar het segment C_3-C_4 aan C_1 maar niet aan C_2 gerelateerd is. In RST zijn dergelijke structuren mogelijk in zgn. *multi-satellite constructions* (MSC's) (zie figuur 5), die echter niet zonder meer als binaire bomen gerepresenteerd kunnen worden.



Figuur 5: Voorbeeld van een multi-satellite construction in RST

Carlson et al. (2003) en Stede (2004) hanteren voor hun omvangrijke corpora van krantenteksten varianten van RST zonder multi-satelliet-constructies. Inderdaad kan voor gevallen zoals in figuur 5 een binaire representatie, waarbij C_4 niet op C_1, maar op C_1-C_3 zou slaan, vaak een redelijk plausibel alternatief bieden. Abelen et al. (1993) echter troffen in fondswervingsbrieven tweezijdige MSC's aan, waarbij de satellieten aan weerszijden van de nucleus staan (bijv. een of meer JUSTIFY of MOTIVATION satellieten vóór het centrale verzoek, en MOTIVATION en ENABLEMENT satellieten erna). Het sterke retorische effect van dergelijke constructies zou bij opsplitsen in successieve binaire relaties niet adequaat gerepresenteerd worden.

Om een idee te krijgen van het voorkomen van enkelzijdige (in principe binair representeerbare) en de meer problematische tweezijdige MSC's, hebben wij de RST-analyses geraadpleegd die op de RST-website gepubliceerd zijn. Er blijken in de 15 gerapporteerde analyses (met in totaal 224 atomaire tekstsegmenten) 12 multi-satelliet-constructies voor te komen, waarvan drie tweezijdige; dat is 1,3 gevallen per 100 segmenten. Voor de vier persuasieve teksten ligt het voorkomen hoger (twee bij 56 segmenten oftewel 3,6 per 100 segmenten), bij niet-persuasieve teksten lager (één tweezijdige MSC op 168 segmenten oftewel 0,6 per 100 segmenten). Overigens is bij zeven van de acht enkelzijdige MSC's telkens maar één type relatie betrokken (bijv. twee EVIDENCE-satellieten), zodat de MSC omgezet kan worden in een JOINT of LIJST.

De problematische tweezijdige multi-satelliet-constructies blijken dus zeer zeldzaam te zijn en praktisch alleen in persuasieve teksten af en toe voor te komen. De door niet-binaire representatievormen (zoals klassieke RST) geboden mogelijkheid om deze gevallen te accommoderen, zou dan ook o.i. niet opwegen tegen het grote voordeel van binaire bomen, die een veel makkelijkere behandeling van constraints bieden.

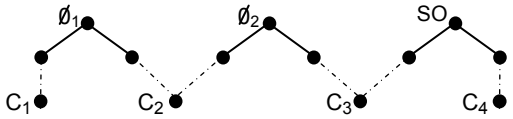
Wat zowel bij RST als bij binaire bomen uitgesloten is, zijn structuren waar een latere uiting over een nucleus of knoop heen aansluit bij een ingebedde eerdere uiting en niet bij de tussenliggende segmenten waarbinnen die eerdere uiting ingebed is. Dit is de zgn. *Right Frontier Constraint (RFC)*: nieuwe segmenten kunnen alleen bij de meest rechts in de representatie staande, 'open' segmenten aansluiten (Polanyi & Scha, 1983; Grosz & Sidner, 1986). Bij binaire boomstructuren zijn dat de laatste knoop (= de laatste uiting) en alle hem dominerende knopen (die relaties tussen eerdere uitingen representeren). Bij RST komt daar voor nucleus-satelliet-constructies nog de bovengenoemde mogelijkheid bij om over een satelliet heen bij een eerdere nucleus aan te sluiten (echter nooit over een nucleus heen naar een eerdere nucleus of satelliet).

3 Partiële representatie van discourse-structuren

Zoals in de inleiding vermeld, zijn er nog steeds geen goede, breed toepasbare computationele modellen van wereldkennis en pragmatische kennis ontwikkeld. Voor een formele representatie van discourse-structuren beschikken wij dus niet over de betekenis van de discourse-segmenten. Wel kunnen op basis van syntactisch geanalyseerde teksten discourse-segmenten, connectieven, anaforen, parallele structuren enz. worden geïdentificeerd. Met deze onvolledige informatie kunnen wij de discourse-structuur partieel beschrijven met behulp van zgn. ondergespecificeerde representaties.

We representeren (partiële) informatie over de discourse-structuur met uitdrukkingen van de *Constraint Language for Lambda Structures* (CLLS; Egg et al. 2001). Deze uitdrukkingen ('constraints') beschrijven alle over een discourse-boom beschikbare informatie. Meestal zijn de constraints compatibel met meer dan één discourse-structuur. In tegenstelling tot andere modelleringen hoeven deze alternatieven (de bij een constraint behorende zgn. 'oplossingen') in deze representatie niet geënummereerd te worden. Vergelijkbare formalismen voor partiële discourse-structuurinformatie zijn o.m. *Underspecified Discourse Representation Theory* (UDRT; Reyle 1993) en *Minimal Recursion Semantics* (MRS; Copestake, Flickinger, Pollard, & Sag, 2005).

De grafische representatie van een discourse-constraint is een discourse-boom waarbij de (nog) niet volledig gespecificeerde dominantierelaties met stippellijntjes weergegeven worden. De doorgetrokken lijnen in de constraint geven aan welke segmenten onder welke relaties vallen (door een relatie-knoop gedomineerd worden). Figuur (6) geeft de initiële discourse-constraint voor voorbeeld (1). 'SO' staat voor een functor die het connectief *so* representeert; impliciete relaties worden met '∅' gelabeled (waarbij de tokens door indices geïdentificeerd worden). Deze constraint incorporeert de *right frontier constraint* en is compatibel met (heeft als oplossingen) alle in figuur (4) beschreven discourse-bomen.



Figuur 6: De initiële ondergespecificeerde discourse-constraint voor voorbeeld (1)

Het linker deel van figuur (6) modelleert bijvoorbeeld dat het impliciete connectief $\emptyset_1$ een verbinding legt tussen C_1 en een ander discourse-segment, dat tenminste C_2 en misschien ook nog verdere segmenten omvat.

Voor verdere specificering van deze constraint wordt gebruik gemaakt van syntactische en semantische informatie. De regels die daarbij toegepast worden, hebben de status van empirische hypothesen die met corpusonderzoek gestaafd dienen te worden. Vaak zal het daarbij niet om wetmatige samenhangen gaan, maar om heuristische *defaults,* die in de meeste gevallen tot de juiste oplossing zullen komen. Of de default in een bepaald geval van toepassing is, zal soms door convergerende evidentie gesteund zijn en anders door manuele controle van de annotatie vastgesteld moeten worden (vandaar dat wij beogen het analyse-instrument *semi*-automatisch te maken).

Voor voorbeeld (1) zou de formulering van aanscherpende constraints als volgt kunnen gaan. Rekening houdend met de parallellie van C_1, C_2 en C_3 [X BE stubborn] zou een semi-automatische analyse kunnen voorstellen om deze drie segmenten als LIST te analyseren, corresponderend aan de structuurvarianten (d) en (e) uit figuur (4). Wij stellen voor om de binaire representatie van LIST-relaties

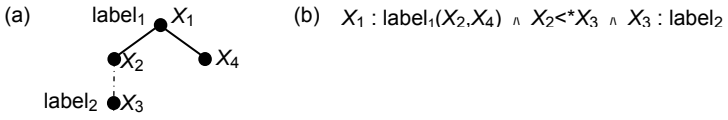
rechtsvertakkend te construeren en kiezen dus voor variant (e), zoals ook in figuur (2) aangegeven is (voor andere multinucleaire relaties, bijv. CONJUNCTION, waarbinnen sprake zou kunnen zijn van een climax, zou eventueel linksvertakking als default gesteld kunnen worden).

Verder zou het connectief *so* op grond van zijn lexicale betekenis gerelateerd kunnen worden aan de causale RST-relaties: VOLITIONAL / NON-VOLITIONAL CAUSE / RESULT. De semantiek van [BE arguing] geeft aan dat het om een handeling gaat en leidt tot de keuze van VOLITIONAL. Voor de keuze tussen CAUSE (met C_4 als nucleus) en RESULT (met C_1-C_3 als nucleus) zou corpusanalytische informatie over het connectief *so* gebruikt kunnen worden: volgens Schiffrin (1987: 193-226) functioneert *so* als een 'superordinate marker' die een conclusie of een terugkeer naar de hoofdlijn signaleert.

3.1 Formele grondslagen van discourse-representaties in CLLS

In deze paragraaf geven wij een beknopte (en enigszins vereenvoudigde) schets van de *Constraint Language for Lambda Structures* (CLLS) en de benodigde uitbreidingen van het formalisme ten behoeve van de representatie van discourse-structuren.

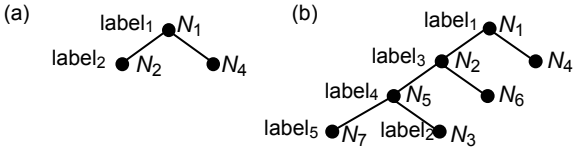
De in deze bijdrage tot nu toe alleen grafisch gerepresenteerde constraints worden in CLLS uitgedrukt als conjuncties van zgn. atomaire constraints die labels aan knopen toewijzen of dominantierelaties tussen knopen aangeven. Figuur 7 illustreert dit (<* staat voor 'domineert'). De formule in (7b) bestaat uit drie atomaire constraints.



Figuur 7: Grafische en formele representatie van een eenvoudige constraint in CLLS

De formulering in termen van atomaire constraints maakt een partiële ordening van constraints mogelijk: C_1 is minstens zo sterk als C_2 precies dan als C_1 minstens alle in C_2 aanwezige atomaire constraints bevat.

De hiërarchische structuur van discourse-relaties in een tekst kan nu als volgt gespecificeerd worden: een boomstructuur is compatibel met een constraint, als er een variabelentoewijzing (van de knoopvariabelen in de constraint naar knopen in de boom) mogelijk is zodat elke atomaire constraint binnen de constraint in kwestie vervuld is. Zo zijn bijvoorbeeld de boomstructuren in figuur (8a) en (8b) beide compatibel met de constraint in figuur 7. In (8a) zijn X_2 en X_3 allebei aan de knoop N_2 toegewezen. Dit is compatibel met de atomaire constraint X_2<*X_3, omdat dominantie ook als identiteit gerealiseerd kan zijn. Uiteraard is er een eenvoudigere constraint waarmee (8a) beschreven kan worden: X_1 : label$_1$(X_2,X_4) ∧ X_2 : label$_2$.



Figuur 8: Twee boomstructuren die compatibel zijn met de constraint in figuur 7

Deze voorbeelden illustreren dat een constraint zoals in figuur 7 in feite een onbeperkt aantal constraints beschrijft, omdat niet nader bepaald is wat er door de knoop X_2 gedomineerd wordt, met als enige voorwaarde dat het de aan X_3 toegewezen knoop moet omvatten. Voor het modelleren van een gegeven tekst zullen overigens alleen

mappings gebruikt worden waarbij elke knoop in de oplossing met een knoopvariabele in de constraint correspondeert.

Voor de specificatie van discourse-relaties moet CLLS uitgebreid worden. Hiervoor postuleren wij een *join-semilattice* structuur $< L, \leq >$ voor de set van labels L, waarvan het minst specifieke element $\emptyset$ is, het label dat door impliciete relaties geïntroduceerd wordt. De elementen van L modelleren de (vaak partiële) informatie die connectieven over discourse-relaties geven. De partiële ordening van de elementen geeft een hiërarchie aan van (klassen van) discourse-relaties, waarbij bijv. CAUSE 'groter' (breder, minder gespecificeerd) is dan VOLITIONAL CAUSE. Hiermee kan aansluiting gezocht worden bij een taxonomische benadering van discourse-connectieven en discourse-relaties (Knott & Sanders, 1998).

Met de partiële ordening van relatielabels in L wordt het mogelijk om onder atomaire constraints met concurrerende labeltoewijzingen de meest specifieke te kiezen. Op deze manier ontstaat een partiële ordening van de oplossingen, waarbij ook de relatielabels een rol kunnen spelen. Voor een meer technische beschrijving van het formalisme zie Egg en Redeker (te verschijnen), waar onze benadering ook vergeleken wordt met die van Schilder (2002) en met D-LTAG (Lexicalized Tree Adjoining Grammar for Discourse, Webber 2004).

Volledig gespecificeerde discourse-structuren worden dus gemodelleerd als oplossingen van ondergespecificeerde constraints. Het oplossen van constraints geschiedt door toevoegen van dominantierelaties tussen knoopvariabelen en specificaties van relatielabels. Zo is in de modellering van voorbeeld (1) de cruciale stap het toevoegen van een dominantierelatie tussen de linker dochter van de SO-knoopvariabele en de $\emptyset_2$-knoopvariabele. Voor de twee LIST-relaties die aan $\emptyset_1$ en aan $\emptyset_2$ toegewezen worden, blijven beide ordeningen mogelijk (althans als er geen andere constraints toegevoegd worden, bijvoorbeeld op basis van een regel dat LIST-relaties als linksvertakkend gemodelleerd worden, zoals boven besproken).

Meer in het algemeen kan voor veel soorten bijzinnen gepostuleerd worden dat zij als satellieten van de bijbehorende hoofdzin geanalyseerd moeten worden. Dit geldt met name voor beknopte bijzinnen (waarin het werkwoord in de infinitief of als deelwoord voorkomt). Als (in afwijking van klassieke RST) ook onderwerp-, voorwerp- en complementzinnen en beperkende betrekkelijke bijzinnen als aparte segmenten gezien worden, zijn ook zij altijd satellieten van hun inbeddende hoofdzin. Voor uitbreidende betrekkelijke bijzinnen en adverbiale bijzinnen ligt het iets ingewikkelder. Hier kunnen de plaats ten opzichte van de hoofdzin, aspectuele kenmerken van de bijzin en de semantiek van de voegwoorden vaak uitkomst bieden. Zo zal een mediaal geplaatste (parenthetische) bijzin altijd een satelliet zijn. Ook bij vooropgeplaatste adverbiale bijzinnen is dit een sterke tendentie: zij fungeren meestal als oriëntatie of verankering voor de hoofdzin (Chafe 1984: 'guideposts', zie ook Matthiessen & Thompson, 1988). Bijzinnen die op de hoofdzin volgen hebben vaak een cataforische en een anaforische functie (Givón 1987), waardoor zij in de RST-structuur soms als nuclei kunnen optreden. Hier zal voor een voorspelling van de satellietstatus andere informatie nodig zijn, bijvoorbeeld uit aspectuele kenmerken, bijwoordelijke bepalingen en de semantiek van het voegwoord. Zo markeren causale onderschikkende voegwoorden (*omdat, doordat, ondanks,* enz.) doorgaans satellieten, terwijl bijzinnen met een temporeel voegwoord (*toen, wanneer, terwijl,* enz.) vaker nucleusstatus kunnen hebben. Het betreft dan bijvoorbeeld contrastief gebruik van *terwijl* (zoals in de vorige zin hierboven) of, zoals Thompson (1987) aangetoond heeft, narratieve zinnen als "We naderden de top, toen er ineens ...", waarbij de hoofdzin de context aangeeft en de gebeurtenis in de bijzin staat. Overigens vinden Carlson et al. (2003) in het Marcu-corpus dat bijzinnen met *since* en *as* in 84% van de 282 voorkomens satellietstatus

hebben (zij maken echter in hun rapport geen onderscheid tussen temporeel en causaal gebruik van deze connectieven).

Waar met behulp van al deze middelen een voorkeur aangegeven kan worden, kan het systeem een suggestie doen. Is een voorspelling ook in deze voorzichtige vorm niet mogelijk, dan kan de onderschikking/nevenschikking ongespecificeerd blijven, zonder dat de representatie daardoor incoherent zou worden.

4 Empirische validatie

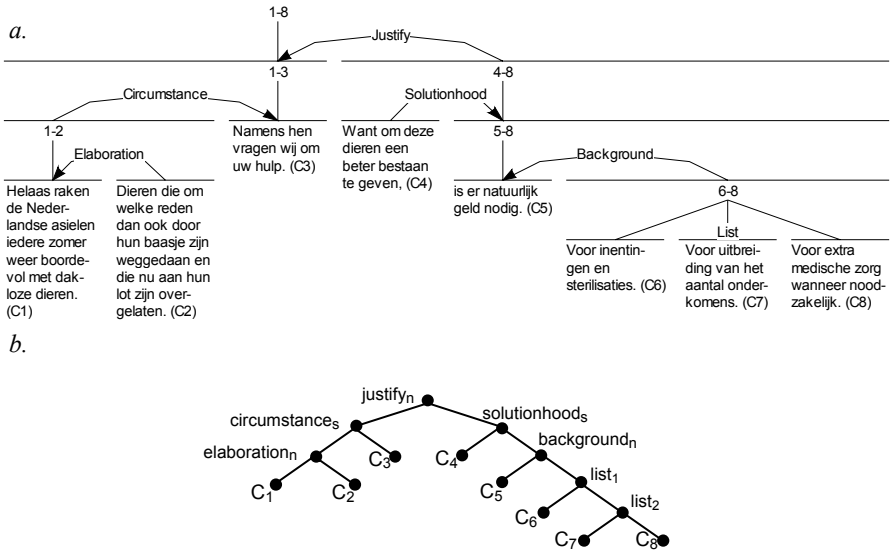
In deze paragraaf zal de herleiding van *constraints* uit de in een tekst aanwezige semantische en syntactische informatie gedemonstreerd worden met een voorbeeld. Daarna worden de resultaten van een pilot study gepresenteerd waarmee getoetst werd in welke mate deze informatie volstaat voor de correcte (plausibele) bepaling van neven- en onderschikking van discourse-segmenten.

4.1 Een uitgewerkt voorbeeld

Aan de hand van een fragment uit een fondswervingsbrief (voorbeeld 3) zal nu geïllustreerd worden hoe de ondergespecificeerde structuren verrijkt zouden kunnen worden op basis van in de tekst aanwezige semantische en syntactische informatie.

(3) Helaas raken de Nederlandse asielen iedere zomer weer boordevol met dakloze dieren. (C_1) Dieren die om welke reden dan ook door hun baasje zijn weggedaan en die nu aan hun lot zijn overgelaten. (C_2) Namens hen vragen wij om uw hulp. (C_3) Want om deze dieren een beter bestaan te geven, (C_4) is er natuurlijk geld nodig. (C_5) Voor inentingen en sterilisaties. (C_6) Voor uitbreiding van het aantal onderkomens. (C_7) Voor extra medische zorg wanneer noodzakelijk. (C_8)

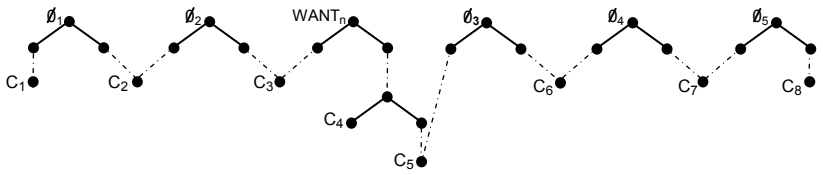
De volledige analyse van dit fragment is weergegeven in figuur (9) in de vorm van een RST-diagram (a) en als binaire boomstructuur (b).



Figuur 9: RST-analyse (a) en binaire boomstructuur (b voor voorbeeld (3)

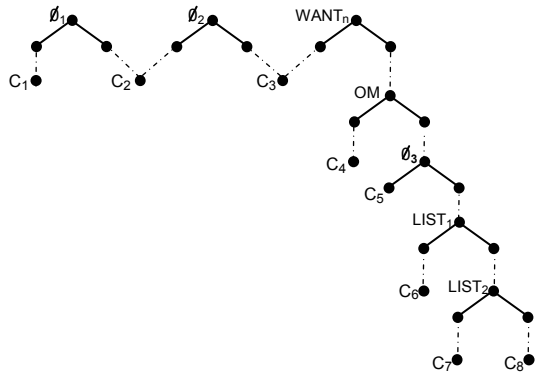
Het fragment bevat twee voegwoorden, *want* en *om*, beide aan het begin van segment C_4. Syntactisch gezien hoort *om* bij de voorop geplaatste beknopte bijzin en *want* bij de volzin als geheel. Uit de semantiek van het voegwoord *want* kan afgeleid worden dat het voorafgaande segment een nucleus en het met *want* geïntroduceerde segment een satelliet zal zijn. De omvang van nucleus en satelliet is echter nog niet gespecificeerd en kan meer dan de onmiddellijk voorafgaande en volgende uitingen bevatten (zeker als *want,* zoals hier, aan het begin van een aparte zin staat).

Voor de syntactisch ondergeschikte beknopte bijzin met *om .. te* is duidelijk dat hij ook in de discourse-structuur ondergeschikt zal zijn aan het volgende segment, bestaande uit C_5 en eventueel daarnaast ook C_6 en volgende. Vandaar dat in de initiële constraint-representatie (figuur 9) de relatie tussen C_4 en C_5 wat C_4 betreft gesloten kan worden (geen stippellijn).



Figuur 10: Initiële representatie van de discourse-structuur van voorbeeld (3)

Nadere inspectie van de syntactische kenmerken van de segmenten C_6 t/m C_8 leidt tot een versterking van de constraint op twee punten. Ten eerste vertonen C_6, C_7 en C_8 syntactische parallellie (drie voorzetselgroepen met *voor*) en vormen daarmee een LIST-relatie. Daarnaast is uit hun elliptische vorm af te leiden dat deze lijst zeer waarschijnlijk een satelliet is bij het voorafgaande segment (C_5). De resulterende versterkte constraint is in figuur 11 afgebeeld.



Figuur 11:De middels syntactische informatie versterkte constraint voor voorbeeld (3)

Om naast deze constraints op de hiërarchische relaties ook voorspellingen te herleiden over de labels van de relaties, kan met name gebruik gemaakt worden van de semantiek van de connectieven. Anders dan *so* in voorbeeld (1) laten de conjuncties *want* en *om* echter een vrij groot aantal RST-relaties toe. Voor onderzoek waarbij de semantiek van connectieven gecombineerd wordt met syntactische kenmerken van de segmenten zie bijvoorbeeld Webber (2004).

4.2 Pilot study: corpusonderzoek

In de twee tot nu toe besproken voorbeelden leveren de syntaxis en de semantiek van connectieven aanzienlijke winst in de specificatie van constraints op de discourse-structuur. Om te toetsen of dit meer in het algemeen voor teksten uit verschillende genres geldt, is een beperkte pilot study uitgevoerd met teksten waarvan wij RST-analyses beschikbaar hadden. Het gaat om 15 teksten, waaronder zes fondswervings-brieven, zes advertenties en drie filmrecensies. In totaal bevatten deze teksten 274 elementaire segmenten ('atomen') en 249 relaties tussen elementaire en/of complexe segmenten. We hebben hierbij de segmentatieregels van de klassieke RST gehanteerd, wat onze schattingen conservatief maakt (zie onze discussie van onderschikking en satelliet-status aan het eind van paragraaf 3).

Voor elke relatie is onderzocht of de relatieve hiërarchische positie van de twee gerelateerde segmenten voorspelbaar is op grond van syntaxis en voegwoorden. In tabel 1 is te zien dat de percentages correcte voorspelling van onder- en neven-schikking varieerden tussen 31,9 % voor de fondswervingsbrieven en 51,8 % voor de filmrecensies, met een gemiddelde van 39,5 %.

Tabel 1 Syntactische disambiguering van discourse-hiërarchie in een pilot corpus

Genre	Atomen	Relaties	Onder- of neven-schikking correct	Percentages
Fondswervingsbrieven	130	119	38	31,9 %
Advertenties	86	74	31	41,9 %
Filmrecensies	58	56	29	51,8 %
Totaal	274	249	98	39,5 %

Voor meer dan de helft van de relaties (60,5%) was geen voorspelling mogelijk. Daar staat tegenover dat foute voorspellingen van onder- of nevenschikking in slechts twee gevallen (0,8% van de relaties) voorkwamen. Het gaat om een MEANS-relatie waarin een *zodat*-bijzin de nucleus is, terwijl hij syntactisch ondergeschikt is (onderstreept in (4)) en een PURPOSE-relatie tussen twee met *en* gecoördineerde zinnen, die op syntactische gronden als nevenschikkend (multinucleair) geclassificeerd zou worden, maar in de RST-analyse als satelliet optreedt (5). Beide voorbeelden komen uit een fondswervingsbrief van de CliniClowns.

(4) Door de clowneske afleiding van de CliniClowns vergeten de kinderen even alles om zich heen. De clowns laten de kinderen zelf het spel bepalen, zodat ze weer even onbezorgd kind kunnen zijn.
(5) Vul vandaag nog de acceptgiro in en steun het werk van CliniClowns.

5 Conclusie

In deze bijdrage hebben wij een benadering van discourse-structuren geschetst die de basis kan vormen voor een semi-automatische discourse-analyse van syntactisch geannoteerde tekstcorpora. Inspectie van voorbeelden en een beperkte, handmatig uitgevoerde pilot study hebben aangetoond dat de hiërarchische relaties van discourse-segmenten voor een aanzienlijk deel voorspeld kunnen worden op grond van in de tekst aanwezige syntactische informatie. Deze voorlopige bevindingen dienen uiteraard met substantiëler corpusonderzoek op basis van syntactisch geannoteerde teksten onderbouwd te worden.

Op basis van de zover mogelijk verrijkte (nog steeds sterk ondergespecificeerde) representaties kan met behulp van geannoteerde corpora getoetst worden in hoeverre andere informatiebronnen tot aanvullende constraints kunnen leiden. Te denken valt aan andere dan de hier besproken syntactische informatie, bijv. over retorische dominantie in gekloofde zinnen (Delin en Oberlander 1995, Oberlander en Delin 1996). Verder zal lexicale informatie zeker een rol moeten spelen. Hier zullen wij ons echter beperken tot relatief oppervlakkige (makkelijk af te leiden) semantische informatie, bijv. conceptuele ontologieën.

Noot

1. Wij danken John Hoeks en twee anonieme reviewers voor hun waardevolle commentaren bij eerdere versies van deze bijdrage.

Literatuur

Abelen, E., Redeker, G. & Thompson, S.A. (1993). The rhetorical structure of US-American and Dutch fund-raising letters. *Text, 13,* 323-350.

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge: Cambridge University Press.

Beek, L. van der, Bouma, G., & Noord, G. van (2002). Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde, 7*, 353-374.

Carlson, L., & Marcu, D. (2001). *Discourse tagging reference manual.* Available from http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf.

Carlson, L., Marcu, D., & Okurowski, M.E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt & R. Smith (Eds.), *Current Directions in Discourse and Dialogue* (pp. 85-112). Dordrecht: Kluwer.

Chafe, W.L. (1984). How people use adverbial clauses. In C. Brugman & M. Macaulay (Eds.), *Proceedings of the Tenth Annual Meeting of the Berkeley Linguistics Society* (pp. 437-449). Berkeley, CA: Berkeley Linguistics Society.

Copestake, A., Flickinger, D., Pollard, C., & Sag, I. (2005). Minimal Recursion Semantics. An introduction. *Research on Language and Computation, 3,* 281-332.

Delin, J., & Oberlander, J. (1995). Syntactic constraints on discourse structure: the case of it-clefts. *Linguistics, 33,* 465-500.

Egg, M., Koller, A., & Niehren, J. (2001). The Constraint Language for Lambda-Structures. *Journal of Logic, Language, and Information, 10*, 457-485.

Egg, M., & Redeker, G. (te verschijnen). Underspecified discourse representation. In A. Benz & P. Kühnlein (Eds.), *Constraints in discourse*. Amsterdam: Benjamins.

Givón, T. (1987). Beyond foreground and background. In R.S. Tomlin (Ed.), *Coherence and grounding in discourse* (pp. 175-188). Amsterdam: Benjamins.

Grosz, B., & C. Sidner (1986). Attention, intentions and the structure of discourse. *Computational Linguistics, 12(3),* 175-204.

Knott, A., & Sanders, T. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics, 30,* 135-175.

Mann, W., & Thompson, S.A. (1988). Rhetorical Structure Theory: Towards a functional theory of text organization. *Text, 8*, 243-281.

Mann, W.C., Matthiessen, C.M.I.M., & Thompson, S.A. (1992). Rhetorical Structure Theory and Text Analysis. In W.C. Mann & S.A. Thompson (Eds.), *Discourse*

Description: Diverse linguistic analyses of a fund-raising text (pp. 39-78). Amsterdam, John Benjamins.

Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.* Ph.D. thesis, Department of Computer Science, University of Toronto.

Matthiessen, C., & Thompson, S.A. (1988). The structure of discourse and 'subordination'. In J. Haiman & S.A. Thompson (Eds.), *Clause combining in grammar and discourse* (pp. 275-329). Amsterdam: Benjamins.

Oberlander, J., & Delin, J. (1996). The function and interpretation of reverse wh-clefts in spoken discourse. *Language and Speech*, *39*, 185-227.

Polanyi, L., & Scha, R. (1983). The syntax of discourse. *Text, 3*, 261-270.

Redeker, G. (2000). Coherence and structure in text and discourse. In W. Black & H. Bunt (Eds.), *Abduction, Belief and Context in Dialogue* (pp. 233-263). Amsterdam: Benjamins.

Reitter, D., & Stede, M. (2003). Step by step: underspecified markup in incremental rhetorical analysis. In *Proceedings 4th International Workshop on Linguistically Interpreted Corpora (LINC-03),* Budapest.

Reyle, U. (1993). Dealing with ambiguities by underspecification: construction, representation, and deduction. *Journal of Semantics, 10*, 123-179.

Sanders, T., & Wijk, C. van (1996). PISA – A procedure for analyzing the structure of explanatory texts. *Text, 16*, 91-132.

Schiffrin, D. (1987). *Discourse Markers.* Cambridge: Cambridge University Press.

Schilder, F. (2002). Robust discourse parsing via discourse markers, topicality and position. *Natural Language Engineering, 8*, 235-255.

Soricut, R., & Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of NAACL 2003*.

Stede, M. (2004). The Potsdam Commentary Corpus. In B. Webber & D. Byron (Eds.), *ACL 2004 Workshop on Discourse Annotation*, Barcelona, Spain (pp. 96-102). Association for Computational Linguistics.

Thompson, S.A. (1987). "Subordination" and narrative event structure. In R.S. Tomlin (Ed.), *Coherence and grounding in discourse* (pp. 435-454). Amsterdam: Benjamins.

Webber, B. (2004). D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science, 28*, 751-779.