

Kyoto: multilinguale terminologie op basis van Wordnets

Roxane Segers¹, Wauter Bosma², Piek Vossen²

¹ Faculteit Exacte Wetenschappen, Vrije Universiteit Amsterdam

² Faculteit der Letteren, Vrije Universiteit Amsterdam

Samenvatting

Het Kyotoproject ontwikkelt een platform waarmee domeinexperts uit verschillende taalgebieden hun kennis en terminologie kunnen delen en specificeren. Om de uitwisseling van cultuurspecifieke kennis mogelijk te maken, is er in het project een systeem ontworpen waarmee domeinexperts voor verschillende talen domeinwordnets kunnen maken. Deze domeinwordnets zijn aan elkaar verbonden door de termen in deze wordnets te koppelen aan één gedeelde en taalafhankelijke ontologie. De domeinwordnets en de ontologie worden vervolgens gebruikt om naar specifieke informatie in teksten te zoeken waarbij de gebruiker de keuze heeft om vergelijkbare gegevens te vinden in verschillende talen. Het Kyotoproject richt zich op het mileudomein en ontwikkelt het platform voor het Engels, Nederlands, Spaans, Baskisch, Italiaans, Japans en Chinees.

1. Inleiding

Het doel van het Kyotoproject¹ (2008-2011) is om een online en gebruiksvriendelijk platform te ontwikkelen dat door domeinexperts uit verschillende taalgebieden kan worden gebruikt voor het organiseren en delen van vakkennis, gebaseerd op de termen die binnen het domein een belangrijke rol spelen. Het platform biedt daarnaast toegang tot uitgebreide tekstcollecties die semantisch doorzocht kunnen worden; dat betekent dat er niet alleen op een term gezocht kan worden, maar ook op specifieke relaties tussen termen, bijvoorbeeld tussen *beheersmaatregel*, *grote grazers* en *afschot*. Achter de termorganisatie en zoekfunctionaliteiten van het Kyotoplatform gaat een omvangrijke architectuur schuil waarin twee componenten een centrale rol spelen. De eerste component in de architectuur zijn de domeinwordnets die voor zeven talen worden ontwikkeld. Een wordnet is een semantisch lexicon waarbij woorden die hetzelfde concept uitdrukken een synset vormen. Deze synsets zijn vervolgens aan elkaar verbonden door verschillende semantische relaties waarvan de hiërarchische *is-a* relatie de belangrijkste is. De synset {*vis*} bijvoorbeeld, heeft een *is-a* of hyperoniemrelatie tot {*organisme*} en een heeft-deel-relatie tot {*kieuw*}. In een domeinwordnet wordt volgens dezelfde principes de terminologie van een bepaald vakgebied beschreven. Het domeinwordnet staat niet op zichzelf, maar wordt als een extensie verbonden met het generieke wordnet van de desbetreffende taal. Daardoor profiteert het domeinwordnet van alle concepten die in het generieke wordnet al beschreven zijn.

Een tweede belangrijke component uit de Kyoto-architectuur is de ontologie. Een ontologie is te omschrijven als een formele en taalafhankelijke specificatie van concepten in een bepaald domein (Studer, 1998). Ook een ontologie is hiërarchisch gemodelleerd, maar onderscheidt zich van een wordnet doordat de organisatie van de concepten los staat van of en hoe ze in een taal worden gelexicaliseerd. In een

¹ KYOTO (acroniem van Knowledge Yielding Ontologies for Transition-based Organisation) is een Europees-Aziatisch project gefinancierd door de EU onder projectnummer 211423 in het 7^{de} "Framework in the area of Digital Libraries: FP7-ICT-2007-1, Objective ICT-2007.4.2: Intelligent Content and Semantics". Zie ook: <http://www.kyoto-project.eu> voor meer informatie en de laatste demo's.

(domein)wordnet worden synsets georganiseerd zoals ze in een taal worden geconceptualiseerd; het Engelse of Chinese wordnet heeft daardoor een andere organisatie van de synsets dan het Nederlandse wordnet. Een voorbeeld van een verschillende organisatie van synsets in het Engelse en Nederlandse wordnet is het Engelse *container* waar synsets onder hangen als *spoon*, *envelope* en *refrigerator*. In het Nederlands bestaat geen equivalent van het Engelse *container* waardoor *lepel*, *envelop* en *koelkast* in het Nederlandse wordnet geen semantische groep vormen, maar elk op verschillende plaatsen in de hiërarchie staan. In een ontologie daarentegen kunnen conceptuele onderscheidingen en overeenkomsten expliciet en taalonafhankelijk worden gemodelleerd. Door de synsets uit de verschillende (domein)wordnets niet rechtstreeks aan elkaar maar via een onafhankelijke ontologie met elkaar te verbinden, ontstaat er een netwerk met expliciete en formele relaties waarin precies kan worden uitgedrukt hoe een synset met een concept uit de ontologie en met synsets in andere domeinwordnets verbonden is.

De architectuur van het Kyotoplatform is ontwikkeld door technici en taalkundigen maar wordt in gebruik genomen door domeinspecialisten. Dat betekent dat domeinspecialisten ook zelf hun terminologie beschrijven. In de visie van het Kyotoproject zijn de gebruikers van het platform experts in hun eigen vakterminologie maar ontbreekt het hen aan de middelen om deze zelf te beschrijven en te onderhouden. Het Kyotoplatform biedt de architectuur en de hulpmiddelen waarmee vakexperts hun terminologie zelf kunnen beschrijven. Om dit proces te faciliteren, zijn er diverse ondersteunende componenten in het platform opgenomen zoals een ontologie, geëxtraheerde termlijsten voor het vormen van de domeinwordnets, en een editomgeving waarin de termen kunnen worden georganiseerd en beschreven. De gebruikers van het platform beschrijven hun terminologie met een specifiek doel, namelijk het delen van informatie en het semantisch doorzoeken van vakteksten op relevante informatie. Voor elke term die een vakspecialist toevoegt aan het domeinwordnet, geldt dat er meteen naar specifieke informatie rond deze term gezocht kan worden. In die zin wordt de gebruiker meteen voor zijn inspanning beloond doordat hij direct betere zoekresultaten terugkrijgt.

Dit artikel is verder als volgt georganiseerd: in paragraaf twee beschrijven we aan de hand van voorbeelden uit het milieudomein het belang van een intelligent zoekstelsel dat kan zoeken op betekenis en relaties. In paragraaf drie presenteren we de algemene architectuur van het Kyotoplatform en in paragraaf vier en vijf gaan we in op respectievelijk drie componenten van de architectuur, te weten de termextractie, de domeinwordnets en de centrale ontologie. In paragraaf zes beschrijven we de manier waarop in Kyoto semantisch kan worden gezocht. In paragraaf acht besluiten we met enige algemene opmerkingen over de status van de Kyotoproject en -architectuur.

2. Multilinguale terminologie en semantisch zoeken

De dagelijkse praktijk van experts in het milieudomein bestaat uit het verzamelen van gegevens en het schrijven van beleids- en aanbevelingsrapporten. Informatie komt steeds vaker beschikbaar via gedigitaliseerde tekstdocumenten, maar om bijvoorbeeld complexe vragen rondom de biodiversiteit in een bepaalde regio te beantwoorden, zijn tal van gegevens nodig die uit verschillende documenten moeten worden betrokken. Domeinexperts worden daardoor geconfronteerd met het bekende probleem dat gegevens met de beschikbare zoeksystemen moeilijk te vinden zijn. Een inventaris van de manier waarop de domeinexperts aangeven te willen zoeken, maakt duidelijk waar de meeste systemen niet in kunnen voorzien.

-Betekenisgericht zoeken De meeste beschikbare zoeksystemen zoeken op vorm en niet op de betekenis van de opgegeven zoektermen. Wie bijvoorbeeld binnen een

tekstcollectie of op het internet op zoek gaat naar informatie over de *populatie wilde eenden in Nederland*, krijgt alleen resultaten terug waarin deze zoekwoorden voorkomen. Relevante documenten waarin wordt gesproken over *wintertalingen en Gelderland*, worden alleen gevonden als de opgegeven zoekwoorden in de context staan. Over het algemeen beschikken zoeksystemen niet over de kennis dat een wintertaling een soort eend is, waardoor relevante informatie niet of toevallig wordt gevonden. Een zoekstelsel dat niet zoekt op vorm maar op betekenis, kan deze informatie wél vinden.

-Relationeel zoeken Een ander problematisch aspect is dat de relatie tussen de zoektermen niet kan worden gespecificeerd in de zoekopdracht waardoor de zoekresultaten veel irrelevante gegevens opleveren. Wie informatie zoekt over *ziekten die bedreigend zijn voor vleermuizen*, kan deze specifieke relatie tussen de termen niet opgeven. Het zoekresultaat geeft teksten terug waar deze termen in voorkomen, maar een groot gedeelte zal gaan over bedreigende ziektes die vleermuizen op mensen kunnen overbrengen.

-Volledige indexering Een derde probleem is dat veel zoeksystemen webpagina's en documenten niet volledig indexeren. Een zoekmachine als Google gebruikt slechts een deel van de website en de daar beschikbare documenten voor zijn zoekindex. Wat niet wordt geïndexeerd kan ook niet worden gevonden. Daarbij zijn de zoekresultaten gerangschikt op het aantal in- en uitgaande links van een webpagina. Op die manier kan een document zonder links maar met relevante informatie op een positie in de lijst zoekresultaten belanden waar niemand meer kijkt. Met een platform als Kyoto kunnen gebruikers alle belangrijke en relevante documentatie uploaden en deze wordt ook volledig door het zoekstelsel verwerkt. Zoekresultaten worden vervolgens geordend op inhoudelijke relevantie en niet op populariteit.

-Multilingualiteit Zoeksystemen hebben beperkte mogelijkheden tot multilinguaal zoeken. Juist dit multilinguale aspect is voor milieukundigen van groot belang omdat het domein internationaal georiënteerd is; beschermingsmaatregelen die in Spanje succesvol zijn gebleken, kunnen ook op gelijksoortige situaties in Nederland worden toegepast. Daarnaast houdt een ecosysteem niet op bij de landsgrens, waardoor samenwerking en het uitwisselen van informatie tussen verschillende landen onontbeerlijk zijn voor milieuorganisaties. Om in anderstalige documenten te kunnen zoeken, moet een domeinexpert zijn terminologie wel kunnen vertalen naar de correcte term in de doeltaal. Op dat punt ontstaan al snel problemen omdat een deel van de termen in het milieudomein zeer taal- en cultuurspecifiek is. Zo heeft het Nederlands een uitgebreide terminologie voor waterbeheersing zoals *inlaag*, *wiel* en *kwelwater* die niet of moeilijk vertaalbaar zijn omdat het concept waar de term naar verwijst in andere taal niet bekend is. Daarnaast kan een bepaald concept wel bekend zijn, maar is er in een andere taal sprake van betekenispecialisatie of -generalisatie bij de lexicalisatie van het concept. Zo gebruikt het Engels de term *host* voor zowel dieren als planten die als gastheer kunnen dienen voor een andere soort. In het Nederlands wordt *gastheer* doorgaans exclusief voor dieren gebruikt en bestaat de term *waardplant* specifiek voor planten die als gastheer optreden.

-Domeinkennis Termen zijn de dragers van vakkennis binnen een domein en spelen ook voor milieuorganisaties een belangrijke rol bij het zoeken naar informatie. Een complicerende factor hierbij is dat milieuterminologie veel woorden uit de algemene taal bevat waar een domeinspecifieke rol aan wordt toegekend. Zo worden *snelwegen* gezien als een *migratieobstructie* voor bepaalde diersoorten en fungeren *wolven* onder andere als *biodiversiteitsindicator*. In de bestaande zoeksystemen is het mogelijk om te zoeken naar *biodiversiteitsindicatoren*, maar niet naar wie of wat deze rol kan vervullen. Een zoekstelsel dat op betekenis en relaties kan zoeken, heeft daarom ook domeinspecifieke kennis nodig om deze relatie te kunnen leggen.

Kyotoproject biedt een platform waarmee domeinexperts documenten en terminologie kunnen onderbrengen ten behoeve van een zoekstelsel waarmee multilinguaal en op

domeinspecifieke betekenis kan worden gezocht. Om dat mogelijk te maken, is een uitgebreide architectuur ontworpen die in de volgende paragrafen verder wordt toegelicht.

3. De architectuur van het Kyotoplatform

De architectuur van het Kyotoplatform wordt in deze paragraaf stapsgewijs uitgelegd aan de hand van figuur 1 dat de verschillende componenten van het platform toont in hun onderling verband. Het grootste gedeelte van Kyotoarchitectuur draait overigens op de achtergrond en blijft geheel onzichtbaar voor de eindgebruiker die daardoor zo min mogelijk geconfronteerd wordt met de technische onderdelen van het platform.

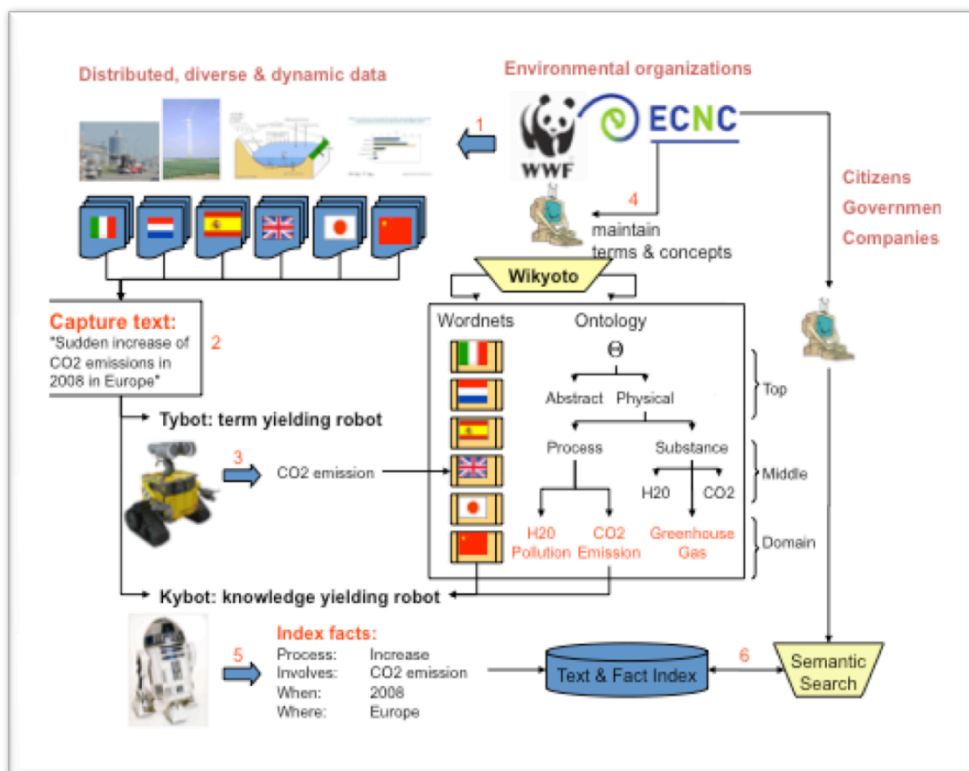


Fig. 1: Architectuur van het Kyotoplatform

1.) De documenten worden door de gebruikers aangeleverd. Voor de invulling van het Kyotoplatform voor het milieudomein hebben de organisaties WWF en ECNC documentcollectie aangelegd in zeven verschillende talen. Deze documentcollectie is niet statisch, maar kan op ieder moment door de gebruiker verder worden aangevuld.

2.) Het Kyotosysteem verwerkt de teksten die zijn aangeleverd in PDF en HTML tot het Kyoto Annotation Format (KAF) (Bosma, 2009), een verdere uitwerking van het LAF formaat (Ide, 2003) dat geschikt is gemaakt voor diepe en taalonafhankelijke syntactische en semantische tekstannotatie. Aan het begin van de cyclus worden de teksten in de verschillende talen syntactisch ontleed; voor de Nederlandse teksten wordt daarvoor de Alpinoparser (Bouma, 2000) gebruikt. Onafhankelijk van de taal en de gebruikte parser worden de woordsoort, de *multiwords* en de dependenties op dezelfde wijze in de teksten gecodeerd. Na het ontleden gaan de teksten door een taalonafhankelijke *Word Sense Disambiguation* (WSD) module (Agirre, 2009; Agirre, 2010), die voor alle verba, substantiva en adjectiva de betekenis voorspelt op basis van synsets uit het generieke wordnet. Bij iedere volgende stap in de Kyotocyclus worden er meer annotatielagen aan de teksten toegevoegd.

3.) De geannoteerde teksten worden gebruikt voor de termextractie. Omdat alle teksten na het ontleden op dezelfde wijze zijn gestructureerd en gecodeerd, is het mogelijk om de termextractie taalonafhankelijk te houden. Het doel van de termextractor is om alle relevante termen voor het domein uit de teksten te halen en deze per taal in kleine hiërarchieën op te slaan. Omdat de teksten nog vóór de extractie door een WSD-module gaan, is het mogelijk om een deel van de termhiërarchieën meteen te koppelen aan een al bestaande synset in het generieke wordnet van de desbetreffende taal. Hierdoor staan de uiteindelijke domeinwordnets die met deze termhiërarchieën worden geconstrueerd, in verbinding met de generieke wordnets en profiteren zo van alle kennis die daar reeds aanwezig is.

4.) De geëxtraheerde termhiërarchieën worden samen met reeds bestaande domeinspecifieke thesauri als de Species2000² aan de eindgebruiker gepresenteerd in een speciale editomgeving, de Wikyoto (Ronzano, 2010). In deze applicatie kunnen de gebruikers de verzamelde termen verder organiseren en onderhouden. In de Wikyoto vindt ook de koppeling plaats van de termen aan de ontologie. De Kyoto ontologie bestaat uit een algemeen en domeinspecifiek gedeelte en is in samenspraak met domeinexperts ontwikkeld. Het koppelen van synsets aan een ontologie is doorgaans werk voor specialisten; in de Wikyoto editomgeving wordt het kiezen van het juiste concept in de ontologie volledig begeleid door een aantal simpele ja/nee-vragen te stellen aan de domeinexperts. Op basis van deze antwoorden kan het systeem automatisch de juiste koppeling en relaties genereren (Segers, 2010).

5.) Alle toegevoegde informatie rond de termen die door de eindgebruikers zijn georganiseerd en gekoppeld aan de ontologie, wordt door het systeem meteen verwerkt in de annotaties van de oorspronkelijke teksten. In de tekstannotatie wordt zondig de betekenis gecorrigeerd van een term; een *wiel* dat eerst door de WSD-module verkeerd was herkend als 'voorwerp dat om een as draait' met het ontologische label *Artefact*, staat nu correct genoteerd als 'klein, diep meertje bij een dijk' en het ontologische label *Body_Of_Water*.

De Knowledge Yielding Robot (Kybot), kan nu op basis van zoekprofielen alle informatie uit de tekst halen die aan de zoekcriteria voldoet. De zoekprofielen bestaan uit conceptuele relaties die worden uitgedrukt in een combinatie van ontologische en morfo-syntactische relaties. Een zoekprofiel gebruikt daarbij de ontologische relaties die voor iedere term in de verwerkte teksten zijn opgeslagen. Hierdoor is het mogelijk om te zoeken op betekenis; een patroon als [Process] + [Bird, *patient*] + [Location] gaat in de teksten op zoek naar:

a.) termen die via het domeinwordnet zijn gelinkt aan het concept Vogel in de ontologie, bijvoorbeeld *vogel, eend, trekvogel* en *broeder*;

b.) processen die volgens de ontologie een relatie hebben met termen die het concept Vogel uitdrukken, waarbij de vogel de patiënt kan zijn van het proces, bijvoorbeeld *uitsterven, predatie, vervuiling*;

c.) alle termen die volgens de ontologie het concept Locatie uitdrukken, bijvoorbeeld *Noordzeekust, wiel, beschermd gebied*.

d.) waarbij alle termen die aan deze voorwaarden voldoen in elkaars nabijheid staan. De Kybot is daarbij niet gebonden aan de voorwaarde dat de termen binnen één zin moeten voorkomen.

De gebruikers hoeven deze zoekprofielen niet zelf te maken maar kunnen een kort fragment uit een document selecteren dat het soort informatie bevat waar ze naar op zoek zijn. Door alle morfo-syntactische en ontologische informatie die in de tekstannotatie is opgeslagen, kan het zoekprofiel dan automatisch gegenereerd worden.

6.) Alle informatie die op basis van een zoekprofiel is gegenereerd, wordt opgeslagen in een Fact Index database.

² zie: www.sp2000.org

7.) Gebruikers kunnen deze database raadplegen om snel informatie te vinden over bepaalde onderwerpen en direct een link volgen naar de originele vindplaats van de informatie.

In de volgende paragrafen worden drie componenten van het Kyotosysteem uitvoeriger besproken: de termextractie, de domeinwordnets en de centrale ontologie.

4. Termextractie

Traditioneel is het doel van termextractie om op basis van een domeincorpus een lijst met termen te vinden die specifiek zijn voor het domein. Vervolgens kan er nog meer informatie over de termen worden verzameld, zoals de relaties tussen de termen. Deze werkwijze impliceert echter dat er termen worden genegeerd omdat ze niet domeinspecifiek zijn, terwijl ze wél een rol spelen in het lexicaliseren van domeinkennis. In Kyoto is ervoor gekozen om de notie domeinspecifiek te vervangen voor domeinrelevant. Zo is een *windturbinepark* niet domeinspecifiek maar wel relevant voor het domein omdat een windturbinepark een verstoring effect kan hebben op de vogeltrek. Door eerst te focussen op domeinrelevante termen kan een zo compleet mogelijk beeld worden gecreëerd van het domein. Daarbij wordt een deel van de betekenis van een term gedefinieerd door de relaties die een term heeft met andere termen. Het vinden van deze relaties is voor de termextractie binnen het Kyotoproject daarom minstens zo belangrijk als het toekennen van domeinrelevantiescores. Zodra duidelijk is wat de relaties zijn tussen de termen, is het ook makkelijker om de termen te voorzien van een relevantiescore. Na het toekennen van de relevantiescore kan de omvang van de termhiërarchieën desgewenst worden gereduceerd door het bepalen van een *threshold* die de termen met de laagste scores er uitfiltert.

In Kyoto worden verschillende en niet-taalspecifieke methoden voor termextractie gebruikt. Het innoverende element hierbij is dat de verschillende methoden serieel worden toegepast waarbij de resultaten van één methode worden doorgegeven aan de volgende extractiemethode. Op deze manier blijkt het mogelijk om termen en relaties te vinden die op basis van één methode alleen niet gevonden hadden kunnen worden. Voor het proces van termextractie begint, zijn de bronteksten eerst getokeniseerd, zijn labels voor woordsoort toegekend, en zijn de teksten ontleed op zinsniveau en voorzien van betekenislabels op woordniveau. De toegekende betekenissen bestaan uit identificatienummers die verwijzen naar een of meerdere mogelijke betekenissen in de algemene wordnets. De WSD-module die hierin voorziet, zoekt ieder lemma in de tekst op in Wordnet en bepaalt aan de hand van lemma's in de directe omgeving de afstand tussen de bijbehorende synsets in het Wordnet. Hoe dichter twee potentiële betekenisandidaten bij elkaar staan, des te waarschijnlijker is het dat dit inderdaad de juiste betekenissen zijn in de tekst. De afstand tussen de synsets *{beer}* (zoogdier) en *{reproductie}* (voortplanting), is bijvoorbeeld korter dan die tussen *{beer}* (werktuig) en *{reproductie}* (kopie). Al deze informatie wordt opgeslagen in Kyoto Annotatieformaat en vormt het startpunt van de taalafhankelijke termextractie.

De termextractie is onderverdeeld in zes opeenvolgende stappen:

1. Extractie van alle kandidaattermen;
2. Morfo-syntactische analyse voor het vinden van hiërarchische relaties;
3. Patroongebaseerde analyse voor het vinden van hiërarchische en deel/heel-relaties;
4. Distributionele statistiek voor het vinden van andere, niet vooraf gespecificeerde relaties
5. Bepalen van de domeinrelevantie door het afwegen van de gevonden relaties tegen de frequentie van een term binnen de documentcollectie;

6. Onderlinge alignment van alle termhiërarchieën in de zeven verschillende talen voor het vinden van nieuwe potentiële termen en relaties.

Deze zes onderdelen van het extractieproces worden in de volgende subparagrafen verder toegelicht.

4.1 Extractie van kandidaat-termen

Twee essentiële kenmerken van potentiële termen zijn dat een term naar een specifiek concept verwijst en dat er syntactische restricties zijn die bepalen of een woordgroep wel of niet een termkandidaat kan vormen. De strategie is om in eerste instantie alle lemma's en woordgroepen te extraheren die voldoen aan de syntactische restrictiecriteria. Zo worden in Kyoto alleen verba, substantiva en adjectiva geëxtraheerd en mogen woordgroepen met een substantief als hoofd bijvoorbeeld niet beginnen met een prepositie of een conjunctie. Het resultaat van deze eerste fase van extractie is een extensieve en platte lijst van termen die niet allemaal domeinrelevant zullen zijn. De hoge recall garandeert echter dat alle termen die belangrijk zouden kunnen zijn voor het domein in de lijst worden opgenomen.

Zowel lemma's als woordgroepen worden geselecteerd als termkandidaat. De lemma's worden per grammaticale categorie toegevoegd aan de lijst kandidaattermen en krijgen een verwijzing naar de vindplaats van deze term in de documentcollectie. Bij de multiword units wordt grofweg dezelfde procedure gevolgd, met dit verschil dat de units nu per syntactische categorie van de groep worden ingedeeld. Daarbij wordt er een extra normalisatie toegepast op het hoofd van de multiword unit zodat *agricultural policies* wordt veranderd in *agricultural policy*, maar dat *migrating species* en *migrated species* wel als twee verschillende termkandidaten behouden blijven.

4.2 Morfo-syntactische analyse

Termen hebben vaak de vorm van multiword units en samenstellingen en deze structuur wordt in Kyoto gebruikt voor het afleiden van hiërarchische relaties tussen de termen. Voor iedere multiword unit en samenstelling wordt gezocht naar de langste eenheid die nog voldoet aan de syntactische criteria om een term te zijn. Als er kandidaattermen zijn die minder elementen hebben dan de omvangrijkste eenheid, dan wordt die gezien als een minder specifieke term; *tropical terrestrial species* wordt opgeslagen als een specifiekere term dan *terrestrial species* en *species*. Dezelfde methode is toegepast voor talen die samenstellingen kennen; *aardwarmte* is dan een specifiekere term dan *warmte*. De termen *warmte* en *species* vormen elk de top van een kleine hiërarchie en zijn gekoppeld aan de synsets {*warmte*} en {*species*} in het algemene Engelse en Nederlandse wordnet.

4.3 Patroonbaseerde analyse

De morfo-syntactische analyse is geschikt voor het vinden van een deel van de hiërarchische relaties; door het toepassen van patroonbaseerde analyse is het mogelijk om extra hiërarchische en meronymierelaties tussen de termen te vinden. Als startpunt voor de patroonbaseerde analyse worden alle teksten uit de documentcollectie gebruikt. Vervolgens wordt voor ieder lemma uit de tekst bekeken of deze is gekoppeld aan het generieke wordnet. Als het daaropvolgende lemma in de tekst ook een koppeling heeft, wordt bekeken of deze twee lemma's een al bestaande meronymie of is-a relatie hebben in wordnet. Als dat zo is, wordt de tekst tussen de lemma's opgeslagen als een potentieel patroon voor die relatie. Op die manier zijn patronen gevonden die kunnen worden gebruikt voor het vinden van relaties tussen nieuwe termen. Een frequent patroon voor een hiërarchische relatie als 'X zoals Y' (*vogels zoals eenden*) kan dan worden gebruikt om een relatie te vinden tussen twee termen die niet in het algemene wordnet staan: *groene daken* zoals *grasdaken*.

4.4 Distributionele statistiek

Een volgende manier om extra relaties tussen de termen te vinden, is het toepassen van distributionele statistiek. De aanname van deze methode is dat termen die in elkaars context staan, op de een of andere wijze aan elkaar zijn gerelateerd. Dit kan een lineaire context zijn (woorden die direct rond de term staan) en een syntactische context; als *marters* en *wilde katten* vaak het subject zijn van *predatie*, dan is dat een indicatie dat deze termen gerelateerd zijn. Vaak zijn deze termen co-hyponiemen van elkaar, dat betekent dat beide termen een is-a relatie hebben naar dezelfde minder specifieke term in de hiërarchie. De statistische methode die voor de distributie van de termen wordt gebruikt is de *Mutual Information score* (Hindle, 1990), waarmee kan worden berekend hoe vaak termen in elkaars context voorkomen in verhouding tot hoe vaak ze los van elkaar voorkomen. Deze score geeft informatie over welke termen sterk aan elkaar zijn gerelateerd en wordt gebruikt om mogelijke synoniemen en co-hyponiemen voor de termen te vinden. Zie (Bosma, 2010) voor de formules die hiervoor worden gebruikt.

3.5 Domeinrelevantie

Na de relatie en termextractie, worden alle termen voorzien van een relevantiescore. Een term die hoog in een termhiërarchie staat en frequent voorkomt binnen de documentcollectie, krijgt daarbij een hoge relevantiescore. Termen die laag in de hiërarchie staan, maar wel een grote frequentie hebben, scoren hoger dan termen met een vergelijkbare positie en een lage frequentie. Termen die niet in hiërarchie staan krijgen een aangepaste score gebaseerd op alleen hun frequentie in de documentcollectie. Zie (Bosma, 2010) voor de formules die zijn gebruikt voor het berekenen van de domeinrelevantie.

3.6 Alignment met termhiërarchieën in andere talen

De termextractie in Kyoto wordt toegepast voor zeven verschillende talen op vergelijkbare documentcollecties. Dat opent perspectieven om de termhiërarchieën voor de verschillende talen aan elkaar te koppelen om zo de termstatus van de vergelijkbare termen te bevestigen en om te detecteren dat bepaalde termstructuren in één van de talen ontbreken. Alle termen die door de termextractor zijn gevonden, hebben een directe of indirecte relatie met een synset uit het algemene wordnet. De wordnets in de verschillende talen zijn weer op synsetniveau gekoppeld aan het Engelse wordnet. De Nederlandse term *invasieve diersoort* en het Spaanse *especie invasora* kunnen bijvoorbeeld aan elkaar worden gerelateerd doordat hun hyperoniem *diersoort* en *especie* uit de generieke wordnets beide een equivalentierelatie hebben met het Engelse *species*. Vervolgens blijkt dat er in de Nederlandse termstructuur ook nog de termen *exotische invasieve diersoort* en *aquatische invasieve diersoort* bevat, en dat ook de termstructuren in andere talen verdere onderverdelingen hebben. Als er termen ontbreken, kan het zijn dat de concepten waar de termen naar verwijzen voor een andere taal niet bestaan, ofwel dat deze termen toevallig niet binnen de tekstcollectie gevonden kunnen worden. In dat geval kan het systeem bij de Spaanse term *especie invasora* aangeven dat er mogelijk specifiekere termen bestaan.

De Kyoto termextractor is geïmplementeerd in verschillende modules die gebruiksvriendelijk op elkaar aansluiten en onderdeel zijn van een grotere keten voor taal- en tekstverwerking (PipeT)³. Binnen dit systeem kunnen (externe) modules en tools die in verschillende programmeertalen zijn geschreven met elkaar samenwerken. Voor meer technische details verwijzen we naar (Bosma, 2010b) en de PipeT website: (<http://pipet.sf.net>).

³ <http://pipet.sf.net/>

4. Ontologie

Een van de belangrijkste componenten van het Kyotoplatfom is de centrale ontologie die de betekenis van de termen in de domeinwordnets verankert. Een ontologie is een formele specificatie van belangrijke concepten in een domein en heeft doorgaans een hiërarchische structuur: de meest abstracte concepten staan in de toplaag en naarmate men de hiërarchie afloopt worden de concepten concreter en specifiek.

Ontologieën bestaan in allerlei vormen en maten en kunnen variëren van vrij klein en domeinspecifiek (PlantOntology⁴, FOAF⁵) tot middelgrote ontologieën die alleen een niet-domeinspecifieke en abstracte toplaag beschrijven (DOLCE⁶) en omvangrijke en algemene ontologieën die zich richten op de top- en middenlaag (SUMO⁷, Cyc⁸). Daarbij kunnen ontologieën meer of minder formeel zijn opgesteld; hoe explicieter de semantiek van de concepten wordt beschreven in de vorm van axioma's en restricties, hoe formeler de ontologie is.

Ontologieën worden veel gebruikt om het uitwisselen en verbinden van heterogene gegevens te vergemakkelijken. Als een museum zijn digitale collectie wil verbinden met die van een ander instituut, kan een domeinontologie als de CIDOC-CRM⁹ worden gebruikt om de metadata van beide collecties op elkaar af te stemmen en correct met elkaar te verbinden. Daarbij is in een ontologie gemodelleerd hoe concepten als *Schilder*, *Kunstwerk* en *Afmeting* met elkaar zijn gerelateerd waardoor duidelijk wordt dat *Afmeting* binnen dit domein alleen kan worden gebruikt voor een *Kunstwerk* en niet voor een *Kunstenaar*. In Kyoto wordt de ontologie gebruikt om de betekenis van domeinsynsets uit verschillende talen te verankeren, en ook in deze toepassing maakt dat het uitwisselen en delen van terminologie en informatie gemakkelijker.

Het is belangrijk om te benadrukken dat een ontologie uit *concepten* bestaat. Doorgaans worden er woorden als *Proces* of *Artefact* gebruikt voor de concepten, maar deze moeten gezien worden als labels waardoor het voor mensen makkelijker is te begrijpen wat er staat. Computers werken op basis van de axioma's en restricties in een ontologie. In die zin kunnen de labels *Proces* en *Artefact* evengoed worden vervangen door cijfers; aan de betekenis van de concepten zal dat niets veranderen. Met de axioma's en restricties kan een computer vervolgens verschillende elegante redeneringen uitvoeren waarbij deductie en het overerven van eigenschappen voor het Kyoto-project het belangrijkste zijn:

Deductie:

Als [Wintertaling] een subklasse (=specifieker concept) is van [Eend], en [Eend] is een subklasse van [Vogel], dan is [Wintertaling] ook een [Vogel].

Top-down overerven van eigenschappen:

Als een [Vogel] de eigenschap 'heeftVeren' heeft, en [Eend] is een subklasse van [Vogel], dan heeft een [Eend] ook de eigenschap 'heeftVeren'.

Maar:

Als [Eend] de eigenschap 'kanVliegen' heeft, en [Vogel] is een superklasse (=minder specifiek concept) van [Eend], dan heeft Vogel *niet* de eigenschap 'kanVliegen'.

⁴ zie: www.plantontology.org

⁵ zie: www.foaf-project.org

⁶ zie: www.loa-cnr.it/Ontologies

⁷ zie: <http://sigma.ontologyportal.org:4010/sigma/Browse.jsp?kb=SUMO> om online in deze ontologie te kunnen zoeken op basis van Engelse synsets die aan deze ontologie gekoppeld zijn.

⁸ zie: <http://cyc.com/>

⁹ zie: www.cidoc-crm.org

Omdat eigenschappen overerven, worden ze toegevoegd aan dat concept in de ontologie dat het meest algemene concept is dat die eigenschap nog kan hebben. Dat heeft als voordeel dat je een eigenschap niet voor iedere klasse opnieuw hoeft te definiëren. Voor redeneersystemen heeft dat als voordeel dat ze minder informatie hoeven te verwerken en daardoor sneller zijn.

In de context van Kyoto is dit redeneren op basis van een ontologie van wezenlijk belang omdat een zoekstelsel nu weet dat de *wintertalingen* uit het voorbeeld in paragraaf 1 een soort *eenden* zijn. Met een ontologie is die kennis formeel vast te leggen en kan allerlei informatie automatisch worden afgeleid, bijvoorbeeld dat een wintertaling veren heeft en kan vliegen.

4.1 De Kyoto ontologie

De Kyoto ontologie is opgebouwd uit drie verschillende lagen en telt in totaal 1133 concepten en 332 formele relaties tussen deze concepten. De top laag is de meest abstracte laag van de ontologie en is gebaseerd op de DOLCE ontologie (Masolo, 2003).

In deze laag bevinden zich concepten als Entiteit, Eigenschap en Kwantiteit.

De middenlaag is gevormd uit de Basis Concepten; dit zijn synsets uit het Engelse wordnet die de meest belangrijke knooppunten en dus concepten in het wordnet representeren (Izquierdo, 2007). De meeste synsets in een wordnet hebben minimaal één relatie naar een andere synset; sommige synsets hebben er echter beduidend meer. Door de boomstructuur van het wordnet van boven naar beneden af te lopen, kan per tak van de boom worden berekend welke synsets zeer veel relaties hebben en daarmee als een belangrijk knooppunt fungeren in de hiërarchie. Deze synsets zijn opgenomen in de niet-domeinspecifieke middenlaag van de ontologie en vervolgens gekoppeld aan de oorspronkelijke synsets in het Engelse wordnet en aan alle equivalenten synsets in de andere algemene wordnets. De belangrijkste knooppunten in de wordnets zijn daarmee al gelijk voorzien van een ontologisch label. De Basis Concepten spelen een belangrijke rol in de koppeling van domeinsynsets aan de centrale ontologie. (zie paragraaf 4.3).

Concepten uit deze niet-domeinspecifieke middenlaag van de ontologie zijn bijvoorbeeld Artefact, Voedsel en Meubelstuk.

De derde laag van de ontologie is domeinspecifiek; hierin staan concepten die door de milieukundigen zijn geselecteerd op domeinrelevantie en bestaat uit concepten als Biodiversiteit, Commerciële_Visserij en Irrigatie.

De Kyoto ontologie is bewust klein gehouden; zo staat er bijvoorbeeld slechts een zeer kleine selectie van de bijna 2 miljoen bekende dier- en plantensoorten in de ontologie. Een goede reden om niet alle soorten op te nemen is vooral van praktische aard: geen enkel redeneersysteem kan op dit moment met zo'n grote ontologie omgaan.

Belangrijker nog is dat het ook niet nodig is om alle soorten en hun eigenschappen in een ontologie op te nemen omdat de vakexperts zijn gespecialiseerd in de wetenschappelijke kenmerken van soorten, en die kenmerken zullen tussen en culturen niet verschillen. Soortenkennis is bovendien ook niet het soort informatie waar zij naar op zoek gaan in hun documentatie; belangrijker is het voor hen om te weten dat er soorten zijn die bepaalde dingen doen of ondergaan en wat dat betekent voor het leefmilieu. Om die vragen te kunnen beantwoorden, volstaat een kleine laag in de ontologie van dier- en plantensoorten waarin vrij algemene concepten staan beschreven als Eend, Kikker en Mos. Vanuit het wordnet kunnen dan desgewenst allerlei specifieke eendensoorten als *wintertaling* aan het concept Eend in de ontologie worden gelinkt. Zo kan dan in de tekst worden teruggevonden dat een *wintertaling* een Eend is, zonder dat het concept *wintertaling* in de ontologie staat.

4.2 Domeinwordnets

Voor alle zeven talen in het Kyotoproject (Engels, Nederlands, Spaans, Baskisch, Italiaans, Japans en Chinees) worden termstructuren gemaakt die het startpunt vormen

voor de uiteindelijke domeinwordnets. Het creëren van de domeinwordnets is een taak voor de domeinexperts aangezien zij de meeste kennis hebben van hun terminologie. De termstructuren bieden daarbij uiteraard veel hulp, aangezien de experts niet vanuit het niets hoeven te beginnen met het organiseren van hun terminologie. De termstructuren zijn daarbij ook niet dwingend; indien een expert liever een andere organisatie had gezien, is dat in de editomgeving makkelijk aan te passen. Voor het ontwikkelen van de domeinwordnets staan ook nog andere hulpmiddelen ter beschikking van de domeinexpert. Zo is de Species2000 database die binnen het domein als een belangrijke bron wordt gezien voor de taxonomische indeling van soorten, geschikt gemaakt om binnen de editomgeving te gebruiken. Experts kunnen desgewenst hele stukken van een taxonomie rechtstreeks in het domeinwordnet plakken. Waar mogelijk worden termen voorzien van definities die automatisch worden geëxtraheerd van Wikipedia; ook deze kunnen door de experts verder worden verfijnd. Naast de editomgeving is er een wiki waar gediscussieerd kan worden over de juiste plaats en betekenis van termen. In die zin voorziet het project ook in een omgeving waar domeinexperts een eigen internationale kennissamenleving kunnen vormen en onderhouden. Het domeinwordnet wordt georganiseerd volgens de belangrijkste relatie uit de generieke wordnets: de is-a of hyperoniemrelatie die zorgt voor een hiërarchische indeling van de termen. De termen die aan het domeinwordnet worden toegevoegd, moeten ook een relatie krijgen met de ontologie. Op deze manier kan de betekenis van een term taalonafhankelijk verankerd worden. De relaties tussen de generieke wordnets, de domeinwordnets, de externe thesauri en de ontologie zijn schematisch weergegeven in figuur 2.

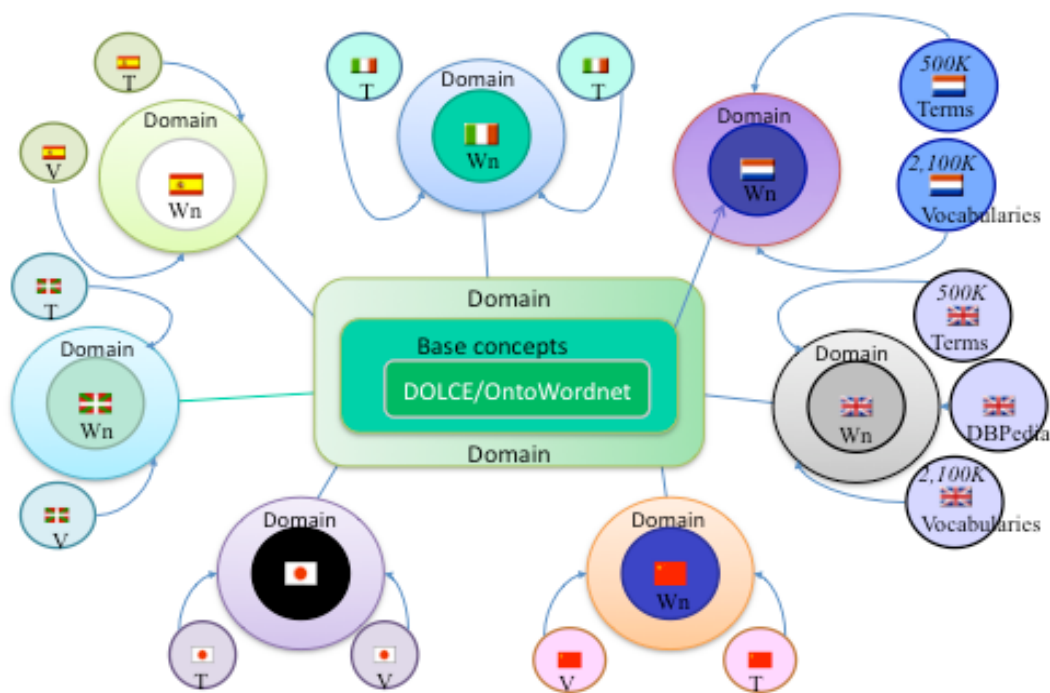


Fig. 2: Relaties tussen de generieke wordnets, de domeinwordnets, de externe thesauri en de ontologie.

Dit figuur toont de situatie voor de zeven verschillende talen in Kyoto, te weten Engels, Nederlands, Spaans, Baskisch, Italiaans, Japans en Chinees. Voor ieder van deze talen bestaat reeds een algemeen wordnet (Wn in het figuur) dat is opgebouwd uit synsets

voor substantiva, verba en adjectiva. De omvang en organisatie van elk wordnet is anders als gevolg van taalintrinsieke kenmerken. Aan deze wordnets worden domeinwordnets toegevoegd (domain in het figuur); binnen Kyoto zijn dat specifiek wordnets voor milieutermologie, maar voor ieder domein kan een dergelijk wordnet worden opgebouwd met behulp van verschillende modules van het Kyotoplatform. Als hulpmiddel voor het creëren van de domeinwordnets zijn er hiërarchische termbanken geëxtraheerd uit domeinspecifieke tekstcollecties (T in het figuur) en staan er externe thesauri (V in het figuur) tot de beschikking van de gebruiker zoals Species2000 en DBpedia, een gestructureerde versie van Wikipedia die wordt gebruikt voor allerlei toepassingen in het semantisch web (Berners-Lee, 2001). De domeintermen worden gekoppeld aan de taalafhankelijke Kyoto-ontologie (afgebeeld in het midden van de figuur). Via de ontologie staan de domeinsynsets uit de verschillende talen met elkaar in verbinding.

4.3 Koppeling van synsets aan de ontologie

De gebruikers van het Kyotoplatform gaan de domeinsynsets zelf koppelen aan de ontologie. Voor deze koppeling zijn er tal van relaties gedefinieerd die tussen een synset en een term uit de ontologie kunnen worden gebruikt. Gebruikers hoeven deze relaties niet te kennen of te begrijpen omdat het systeem door simpele vragen te stellen zelf de juiste relaties kan afleiden.

Voor er relaties kunnen worden gelegd, moet het systeem eerst weten of de term een *type* of een *rol* uitdrukt. Een *huisdier*, *leraar* en een *waardplant* drukken een rol uit die iets of iemand kan dragen, maar die ook weer kan worden afgelegd. 'Huisdier' kan een rol zijn van kat, maar een kat kan ook een *straatkat* worden of een *asielkat*. Een kat kan echter nooit ophouden een kat te zijn. Anders gezegd wordt een kat als kat geboren en gaat als kat dood, maar gedurende zijn leven kan een kat tal van rollen aannemen die allemaal tijdelijk en niet essentieel zijn voor wat het betekent om een kat te zijn. Dit onderscheid tussen een 'altijd' en 'tijdelijk' wordt rigiditeit genoemd. Een kat is dan een rigide concept en de rol huisdier een niet-rigide concept (Guarino, 2002). Om goed te kunnen redeneren met een ontologie is dit onderscheid van wezenlijk belang; in figuur 3 leggen we uit wat er zou gebeuren als dit onderscheid niet wordt gemaakt in een ontologie:

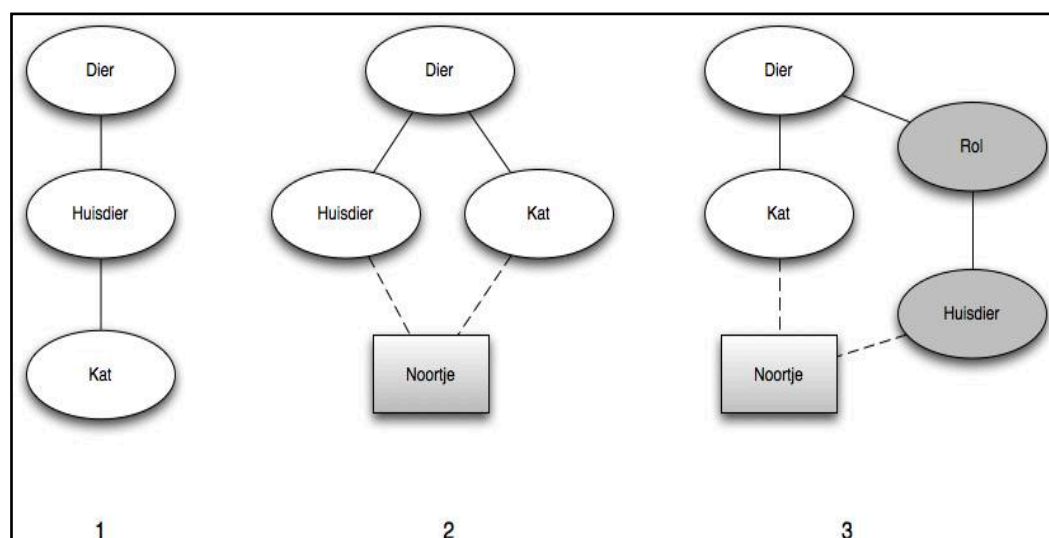


Fig. 3: Het verschil tussen rigide en niet-rigide concepten in een ontologie

In de eerste hiërarchie staat Kat onder Huisdier; dat zou betekenen dat iedere kat altijd een soort huisdier is wat uiteraard niet klopt. Een oplossing zou kunnen zijn om te stellen dat Huisdier en Kat niet in een hiërarchische relatie kunnen staan, maar dat beide een subklasse zijn van Dier, zoals te zien is in de tweede hiërarchie. Een probleem daarbij is dat volgens deze ontologie een Huisdier altijd een soort Dier is. In het dagelijks taalgebruik is het volkomen logisch om dat te zeggen, maar beide concepten hebben toch een wezenlijk ander karakter; een huisdier is immers niet een soort dier zoals kat en hond dat wel zijn, maar een *rol* van bepaalde dieren. Het probleem wordt duidelijk als we een kleine redenatie toepassen met deze ontologie. Stel dat we met deze mini-ontologie zouden willen uitdrukken dat een specifieke kat die Noortje heet, een Huisdier is én een Kat. Op het moment dat Noortje geen instantie van Huisdier meer is, is ze volgens de ene relatie met onze ontologie ook geen Dier meer, maar via de andere relatie wél. Als we de rollen op deze manier beschrijven in de ontologie, leidt dat tot allerlei inconsistenties.

Wat we eigenlijk formeel willen weergeven, is dat Noortje een specifieke Kat is en de rol heeft van Huisdier. In de derde hiërarchie staat de manier waarop dat in de Kyoto-ontologie wordt uitgedrukt. Huisdier is nu expliciet beschreven als een Rol die van toepassing kan zijn op Dier. Welke dieren precies de rol kunnen hebben van Huisdier is vrij cultuurspecifiek en wordt dus niet expliciet in de ontologie beschreven. De kat Noortje kan nu aan de ontologie worden gelinkt als een instantie van Kat en met de rol Huisdier. Op dezelfde manier worden ook domeinsynsets die een rol uitdrukken aan de ontologie gekoppeld.

Voor synsets uit het domeinwordnet die een rol uitdrukken zoals *straatkat* en *waardplant*, zoekt het systeem naar processen, objecten en eigenschappen die aan deze rol gerelateerd zijn. Met deze relaties naar concepten in de ontologie kunnen we de semantiek van een rol als waardplant expliciet maken; bijvoorbeeld dat het om een rol van *planten* gaat, en dat die rol inhoudt dat ze *gegeten* worden door *dieren*.

Nieuwe synsets uit het domeinwordnet die rigide concepten uitdrukken zoals *wiel* en *wintertaling* krijgen een subklasse-relatie met de concepten Meer en Eend in de ontologie. Op die manier wordt uitgedrukt dat een *wiel* een soort Meer is en dus ook alle eigenschappen en relaties heeft die voor dit concept in de ontologie staan beschreven.

De wijze waarop er relaties tussen niet-rigide domeinsynsets en de ontologie worden gemaakt, zal verder worden uitgelegd aan de hand van de term *draadslachtoffer*. (Een draadslachtoffer is een vogel die tegen hoogspanningskabels is gevlogen en daardoor is overleden.)

Omdat de nieuwe domeinwordnets aan de algemene wordnets zijn gekoppeld, heeft iedere nieuwe domeinsynset meteen ook een ontologisch label dat de betekenis van de synsets vrij grof karakteriseert. Zo kan een domeinexpert de term *{draadslachtoffer}* toevoegen aan het domeinwordnet en de synset *{vogel}* kiezen als hyperoniem. De synset *{vogel}* staat al in het algemene wordnet en heeft daar het ontologische label Bird. De precieze semantiek van *{draadslachtoffer}* is dan nog niet helemaal bekend, maar het systeem heeft al wel een startpunt om op zoek te gaan naar hoe de nieuwe synset zich tot de ontologische label Bird verhoudt.

Het eerste wat het systeem nu wil weten, is of draadslachtoffer nu echt een *type* vogel is of een *rol* die een vogel kan aannemen. Een speciale tool die binnen Kyoto is ontwikkeld (Hicks, 2009), kan op basis van specifieke patronen rond een term redelijk accuraat voorspellen of een term een rol is of niet. In dit geval blijkt dat de term een rol is; een vogel is immers niet altijd of essentieel een draadslachtoffer.

Vervolgens bekijkt de gebruiker de definities die voor deze term gevonden zijn en klikt op die processen en eigenschappen die belangrijk zijn voor de term. In dit geval klikte de gebruiker op *hoogspanningskabel* en *overlijden*. Al deze woorden staan als synset in het algemene wordnet. Het systeem gaat via de wordnethiërarchie naar boven en zoekt voor iedere term de eerste synset die een koppeling heeft met de centrale ontologie.

Hoogspanningskabel is gerelateerd aan het concept Device in de ontologie en *overlijden* aan DyingProcess. Bij deze concepten staat in de ontologie al een kort lijstje vragen klaar waarmee het systeem de juiste relatie tussen de synset en de concepten in de ontologie kan leggen. Een vraag bij DyingProcess is bijvoorbeeld: 'ondergaat [draadslachtoffer] een sterfproces? (Ja/Nee)'.

Als de gebruiker de vragen met een simpele 'ja' of 'nee' heeft beantwoord, heeft het systeem automatisch de volgende relaties gelegd:

{ <i>draadslachtoffer</i> }	{ <i>roadkill</i> }
instance: Bird	instance: Animal
patient: DyingProcess	patient: DyingProcess
cause: Device	cause: Vehicle (= subklasse van Device)

Als we deze term en relaties vergelijken met het Engelse *roadkill*, dan zien we dat voor beide termen dezelfde relaties zijn gemaakt naar concepten die qua betekenis dicht bij elkaar staan in de ontologie. *Roadkill* is is een mogelijke rol van allerlei soorten dieren, niet specifiek van vogels, en de veroorzaker is een voertuig. Op basis van deze relaties weet het systeem nu dat een *eend* wel *roadkill* kan zijn (een eend is immers een soort dier), maar een *paard* nooit *draadslachtoffer*. Die restrictie zit besloten in het feit dat de domeinexpert *draadslachtoffer* heeft gerelateerd aan Bird, niet aan Animal. De semantische verschillen en overeenkomsten tussen beide termen staan met deze notatie expliciet en formeel genoteerd. Als een gebruiker nu in de tekstcollectie op zoek wil naar doodsoorzaken van eenden tijdens hun migratie, kan het systeem uit de Nederlandse teksten onder andere informatie genereren rond de term *draadslachtoffer* en desgewenst in de Engelse tekstcollectie informatie geven over *roadkill*.

5. Semantisch zoeken in Kyoto

De zeven domeinwordnets en de ontologie voorzien in een kennisbank die door domeinexperts wordt gebruikt om hun terminologie en kennis te delen. In de eerste plaats staan de termen voor iedere taal duidelijk gespecificeerd door de relaties die zijn gelegd met concepten in de ontologie. Experts kunnen op die manier zoeken naar vergelijkbare termen en hun precieze toepassing in andere talen zoals het Nederlandse *draadslachtoffer* en het Engelse *roadkill* uit de vorige paragraaf.

De kennisbank wordt ook gebruikt voor de semantische zoekmachine. Alle relaties die vanuit de domeinwordnets naar de ontologie zijn gelegd, worden bijgeschreven in de documentcollectie. Deze documentcollectie krijgt daardoor een zeer rijke semantische en taalafhankelijke annotatie op basis waarvan complexe zoekopdrachten kunnen worden uitgevoerd. De zoekopdracht heeft het uiterlijk van een conceptueel profiel dat bestaat uit morfo-syntactische en ontologische relaties. Deze profielen lezen door de tekst heen en halen alle informatie uit de tekst die voldoet aan de criteria van dat zoekpatroon. Een patroon als:

[Object, cause], [Decrease], [Animal, patient], [Place], [Quantity],

gaat op zoek in teksten naar alle woorden die volgens de ontologie kunnen voldoen aan deze labels Object, DecreasingProcess, Bird en Place, en houdt daarbij rekening met de relatie die de term moeten hebben en met de afstand in de tekst tussen de termen.¹⁰ Met dit patroon kan bijvoorbeeld worden gevonden dat in Nederland [Place] jaarlijks honderdduizenden [Quantity] vogels [subklasse van Animal] sterven [gerelateerd aan

¹⁰ Dit zoekprofiel is een vereenvoudigde weergave van de werkelijke en operationele profielen. De volledige codering van de zoekprofielen kan gevonden worden op: www.kyoto-project.eu/

Decrease] doordat ze tegen hoogspanningskabels [Object, cause] vliegen. Het systeem kan echter ook informatie vinden over 'honderdduizenden draadslachtoffers in Nederland' omdat de relatie met een Object, een Vogel en een Decrease in de koppeling van deze term met de ontologie is verwerkt. Op die manier kan een milieukundige ook informatie vinden die in specialistische termen is verpakt en die hij misschien niet kent omdat ze alleen door een kleine groep binnen zijn vakgebied worden gebruikt. Deze zoekprofielen hoeven niet door de milieukundigen te worden geformuleerd; het volstaat om op een stuk tekst te klikken waar informatie in staat die interessant is voor de gebruiker. Op basis van alle semantische annotaties die in de teksten staan, kan het systeem automatisch een zoekpatroon afleiden en door de hele tekstcollectie laten gaan. Het grote voordeel van deze patronen is dat ze kunnen worden toegepast voor alle talen. Omdat de zoekpatronen werken op basis van ontologische labels in de tekst, zijn ze dus niet afhankelijk van hoe de informatie in de tekst wordt geformuleerd. Er zijn bijvoorbeeld in het Nederlands en Engels diverse synsets die een Decrease uitdrukken zoals {*verminderen*} of {*afname*} en {*drop*} of {*downfall*} die allemaal gevonden kunnen worden.

Alle informatie die door de zoekprofielen wordt gevonden, komt terecht in een database die door de gebruikers kan worden geraadpleegd. Deze database geeft een gestructureerd overzicht van alle informatie die voldeed aan de zoekcriteria waarbij de meest relevante informatie bovenaan komt te staan. Omdat de informatie gestructureerd wordt aangeboden, kunnen milieukundigen gegevens makkelijk met elkaar vergelijken. Als een milieukundige bijvoorbeeld op zoek is gegaan naar het aantal korenwolven op een locatie en het bijbehorende jaartal van de telling, is het makkelijk om te bepalen welke aantallen wel of niet bij elkaar kunnen worden opgeteld omdat de locatie overlapt. Daarnaast kan op basis van het jaartal een duidelijk beeld worden verkregen van de toe- of afname van het aantal korenwolven. De database geeft daarbij ook toegang tot de originele documenten waarin de informatie werd gevonden, zodat de expert desgewenst kan nalezen hoe de telling tot stand is gekomen.

6. Conclusie

Het Kyotoplatform biedt een omgeving voor het delen kennis op basis van multilinguale terminologie. De centrale ontologie functioneert daarbij als een taalafhankelijke interlingua die de betekenis en de relaties tussen de termen formeel vastlegt. De domeinwordnets en de ontologie worden gebruikt als kennisbank voor een intelligente zoekmachine die op zoek kan gaan naar informatie in een meertalige tekstcollectie op basis van betekenis.

Het Kyoto platform is ontworpen als een gebruiksvriendelijke en onafhankelijke infrastructuur. Het platform wordt nu toegepast op het milieudomein, maar is na enkele aanpassingen ook geschikt voor nieuwe domeinen. Als het gehele Kyotoplatform voor andere domeinen gebruikt gaat worden, is het met name van belang dat de domeinspecifieke laag in de ontologie die nu is toegespitst op milieutermologie te vervangen voor concepten die belangrijk zijn voor het doeldomein. Indien gewenst, kan er ook worden gewerkt op basis van alleen de top- en middenlaag van de ontologie. In dat geval zullen de zoekpatronen van de zoekmachine wel wat minder domeinspecifieke informatie in een tekstcollectie kunnen terugvinden.

Alle componenten van de Kyotoarchitectuur, zoals bijvoorbeeld de termextractor, de domeinwordnet editor en de onderdelen van de semantische zoekmachine zijn open source, dus vrij beschikbaar voor (her)gebruik. De milieukundige vakteksten en de domeinwordnets die op het platform worden gedeeld, zijn ook als kennisbank beschikbaar voor terminologen en vertalers.

De verschillende componenten van de Kyoto-architectuur kunnen ook los worden gebruikt; de taalafhankelijke termextractor kan bijvoorbeeld worden gebruikt voor

het aanleggen van voorgestructureerde termbanken waarbij de termen zijn gekoppeld aan de vindplaatsen in een tekstcollectie.

Het Kyotoproject wordt medio 2011 afgerond en verkeert daarmee op het moment van schrijven in zijn laatste fase. Diverse componenten van de architectuur zoals de editomgeving, de termextractiemodules en de eerste versie van de semantische zoekmachine zijn gereed en kunnen via de demo's op de website worden geraadpleegd. De module die de gebruiker moet assisteren bij het koppelen van domeinsynsnetten aan de ontologie wordt eind 2010 geïmplementeerd en geëvalueerd. De aanleg van de domeinwordnets is in volle gang en deze zullen ook na de beëindiging van het project verder worden uitgebreid door een gemeenschap van milieukundigen die het platform in gebruik nemen. Op de Kyoto website www.kyoto-project.eu is regelmatig een update te vinden over nieuwe ontwikkelingen en demo's.

Bibliografie

- Agirre E., Lopez de Lacalle O., Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen and Roxane Segers (2010). SemEval-2010 Task 17: All-Words Word Sense Disambiguation on a Specific Domain. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. 75--80. Uppsala, Sweden
- Agirre, E. and Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In: *Proceedings of EACL, 2009*.
- Berners-Lee, Tim; James Hendler and Ora Lassila (mei 2001). "The Semantic Web". In: *Scientific American Magazine*.
- Bosma, W.E., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., and Aliprandi, C. (2009). KAF: a generic semantic annotation format. In: *Proceedings of the GL2009 Workshop on Semantic Annotation*, September 2009.
- Bosma W.E., P. Vossen (2010). Bootstrapping language neutral term extraction. In: *Proceedings of the 7th international conference on Language Resources and Evaluation, (LREC2010)*, Malta, May 17-23, 2010.
- Bosma W.E., P. Vossen. (2010b) *Concept Miners Revised*. Kyoto Deliverable 5.3, VU University Amsterdam.
- Bouma, G., Noord, G.van, and R. Malouf. (2000) Alpino: wide-coverage computational analysis of Dutch. In: *Proceedings of CLIN, 2000*.
- Guarino, N., and C. Welty (2002). Evaluating ontological decisions with ontoclean. In: *Communications of the ACM*, 45(2): 61–65, 2002.
- Hicks, A. and Herold, A. Evaluating ontologies with rudify. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD)*, October 2009.
- Hindle, D. (1990). Noun classification from predicate-argument structures. In: *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, pages 268–275, Morristown, NJ, USA.
- Ide, N., Romary, L. (2003). Outline of the international standard Linguistic Annotation Format. In: *Proceedings of the ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right*.
- Izquierdo, R., Suarez, A., Rigau, G. (2007). Exploring the Automatic Selection of Basic Level Concepts. In: *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'07)*. Borovetz, Bulgaria, September 2007.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A. (2003). *Wonderweb Deliverable D18: Ontology Library*. ISTC-CNR, Trento, Italy.

Ronzano F., Maurizio Tesconi, Salvatore Minutoli, Andrea Marchetti (2010). Collaborative management of KYOTO Multilingual Knowledge Base: the Wikyoto Knowledge Editor. In: *Proceedings of the 5th Global WordNet Conference (GWC2010)*, Mumbai, India. January 31-February 4, 2010.

Segers R., Vossen, P. (2010): Facilitating Non-expert Users of the KYOTO Platform: the TMEKO Editing Protocol for Synset to Ontology Mappings. In: *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC2010)*, Malta, May 17-23, 2010.

Studer, R., Benjamins, V.R., Fensel, D. (1998). Knowledge Engineering: Principles and Methods. In: *Data Knowledge Engineering* 25, p. 161-197.

Vossen, P. (Ed.) (1998) *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.