Automatische extractie van hyponiemrelaties uit grote tekstcorpora

Afstudeerscriptie Informatiekunde

Leonie IJzereef

Scriptiebegeleider en eerste lezer:
dr. G. Bouma
Tweede lezer:
dr. G. van Noord

1 augustus 2004

Inhoudsopgave

Hoofdstuk 1

Inleiding

1.1 Introductie

Met de explosieve groei van de op het internet beschikbare informatie groeit ook de vraag naar een goede zoekmethode. De huidige zoekmachines geven aan de hand van ingegeven *keywords* de meest relevante documenten als resultaat. Vaak levert zo'n zoekactie vele duizenden weinig of ongestructureerde documenten op. Het is dan aan de gebruiker om uit deze documenten de gewenste informatie te halen.

Gebruikers hebben echter behoefte aan een zoekmachine die geen heel document, maar alleen relevante informatie als zoekresultaat geeft. Een techniek op het gebied van Information Retrieval die hierop inspringt en waarnaar op dit moment veel onderzoek wordt gedaan is Question Answering. Toepassing van deze techniek moet een zoekmachine opleveren waarbij de gebruiker een vraag ingeeft en vervolgens een of meer relevante antwoorden als resultaat krijgt.

Bij Question Answering (QA) wordt gebruikt gemaakt van allerlei technieken uit de computationele taalkunde. Zo worden syntactische en semantische kenmerken van de vraag en de potentiële antwoorden gebruikt om de meest waarschijnlijke antwoorden te vinden. Hiervoor is syntactische en semantische kennis nodig. Een voorbeeld van zo'n bron van semantische kennis is een ontologie; een grote database met woorden en de semantische relaties tussen woorden, zoals bijvoorbeeld hyponiem- en synoniemrelaties.

Voor het Nederlands is op dit moment één zo'n grote, algemene ontologie beschikbaar: EuroWordNet (Vossen 1998). EuroWordNet is handmatig

gebouwd en bestaat uit algemene ontologieën voor zeven verschillende Europese talen, waaronder Nederlands. Hoewel het Nederlandse deel van EuroWordNet 70.366 woorden[1] bevat, is dit erg weinig om goed bij te kunnen dragen aan de selectie van relevante antwoorden door een QA-systeem. In het algemene lexicon van EuroWordNet zijn veel van de benodigde specifieke relaties niet aanwezig, en daarom is uitbreiding gewenst. Uitbreiding van een ontologie kan handmatig gebeuren, maar dit is erg tijdrovend. Een automatische uitbreidingsmethode die behalve snel ook accuraat is, zou uitkomst kunnen bieden.

In deze scriptie wordt een onderzoek naar een automatische uitbreidingsmethode voor ontologieën beschreven. Centraal staat een methode waarmee hyponiemrelaties tussen nomina m.b.v. lexicaal-syntactische patronen uit grote tekstcorpora kunnen worden geëxtraheerd. Deze hyponiemrelaties kunnen vervolgens (in later onderzoek) gebruikt worden voor de uitbreiding van een ontologie zoals EuroWordNet.
In het vervolg van deze inleiding wordt geschetst wat de bijdrage van ontologieën kan zijn aan QA-algoritmes. Daarnaast wordt kort de uitbreiding van ontologieën besproken en wordt aangegeven welk deel van dit proces in deze scriptie aan bod zal komen. Tenslotte wordt de opzet van deze scriptie uitgelegd.

1.2 Bijdrage van ontologieën aan Question Answering

Een ontologie bestaat uit een groot lexicon met woorden (van verschillende woordsoorten, zoals nomina, adjectieven en werkwoorden) en uit allerlei relaties tussen woorden, zoals hyponiemrelaties en synoniemrelaties.

Een hyponiemrelatie `hyponiem(X,Y)` tussen twee woorden of begrippen is een relatie waarvoor geldt dat:

> X is een (soort van) Y
>
> hyponiem("hond","dier")
> hyponiem("Nederland","Europees land")

[1] 54.439 nomina, 14.278 werkwoorden, 1.597 adjectieven en 52 adverbia

Een synoniemrelatie is een relatie waarvoor geldt dat:

semantische betekenis (X) = semantische betekenis (Y)

synoniem("auteur","schrijver")
synoniem("godsdienst","religie")

Hoewel het bij de zojuist gegeven voorbeelden om relaties tussen nomina gaat, kunnen deze relaties zich ook voordoen bij adjectieven en werkwoorden.

De semantische kennis van een ontologie kan het QA-algoritme helpen bij het selecteren van relevante antwoorden. Zo kan de kennis die een ontologie heeft over synoniemen gebruikt worden om het bereik van de zoekactie te vergroten.
Neem de volgende vraag:

Vraag: Wie is de schrijver van het boek De Tweeling?

Als het QA-algoritme vervolgens voor alle woorden in de vraag in een ontologie synoniemen zoekt, zou de volgende relatie gevonden kunnen worden:

synoniem("auteur","schrijver")

Om de kans op een correct antwoord te vergroten, kan zowel naar *schrijver* als naar het synoniem *auteur* worden gezocht:

Vraag: Wie is de [schrijver|auteur] van het boek De Tweeling?

Toevoeging van het synoniem *auteur* kan het volgende zoekresultaat opleveren, waaruit het antwoord op de vraag kan worden afgeleid:

Relevante informatie: Tessa de Loo is de auteur van het boek De Tweeling.
Antwoord: Tessa de Loo

Ook de kennis die een ontologie heeft over hyponiemrelaties kan van pas komen bij het selecteren van relevante antwoorden. Stel dat de volgende vraag wordt ingegeven in de QA-zoekmachine:

Vraag: Welk Europees land heeft gebieden met eeuwige sneeuw?

Deze vraag zou de volgende relevante informatie kunnen opleveren, waaruit het juiste antwoord wordt geselecteerd:

> Relevante informatie: In Oostenrijk is boven de 2800 meter eeuwige sneeuw te vinden.
> Antwoord 1: Oostenrijk
>
> Relevante informatie: Peru kent vlak bij de evenaar enorme bergen met eeuwige sneeuw.
> Antwoord 2: Peru

In beide antwoorden is immers een landennaam en het begrip *eeuwige sneeuw* terug te vinden.

Door deze antwoorden te combineren met de volgende ontologische kennis, kan het meest waarschijnlijke antwoord worden gekozen:

> hyponiem("Oostenrijk","Europees land")
> hyponiem("Peru","Zuid-Amerikaans land")

Door kennis van de bovenstaande hyponiemrelaties kan *Oostenrijk* als meest waarschijnlijke antwoord worden geselecteerd.

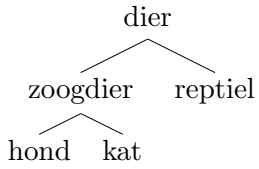
Dit zijn twee voorbeelden die de relevantie van ontologieën voor Question Answering laten zien. Ontologieën zijn waardevoller voor QA-algoritmes naarmate zij meer (gedetailleerde) informatie bevatten, omdat zij dan met meer kennis kunnen bijdragen aan de selectie van het juiste antwoord.

1.3 Uitbreiding van een bestaande ontologie

Ontologieën kennen, door de combinatie van hyponiemrelaties, een hiërarchische structuur. De volgende hyponiemrelaties

> hyponiem("reptiel","dier")
> hyponiem("zoogdier","dier")
> hyponiem("hond","zoogdier")
> hyponiem("kat","zoogdier")

kunnen, aangenomen dat deze woorden niet ambigu zijn, in de volgende hiërarchische boomstructuur worden vastgelegd:



Voor het uitbreiden van een bestaande ontologie zijn nieuwe woorden en relaties tussen woorden nodig, die op de juiste plaats in de hiërarchie geplaatst moeten worden.
De identificatie van deze nieuwe woorden en relaties kan ruwweg op twee manieren. De eerste methode is gebaseerd op de selectievoorkeur van werkwoorden en werkt met domein-specifieke teksten (bijv. Hastings 1994, Hahn & Schnattinger 1998). Door in deze teksten te kijken van welk werkwoord een bepaald concept het subject of object is, kan worden bepaald waar het concept in de ontologie geplaatst moet worden:

Selectie:	het object van *vermoorden* is een levend wezen vermoorden(X,levend wezen).
Zin in corpus:	Jan heeft onze hond vermoord. vermoorden(Jan,hond).
Conclusie:	hyponiem("hond","levend wezen") *hond* wordt in de ontologie geplaatst als hyponiem van *levend wezen*.

De tweede methode maakt gebruik van lexicaal-syntactische patronen om hyponiemrelaties te identificeren in grote corpora met vrije tekst[2] (bijv. Hearst 1992, Rydin 2002).

Lexicaal-syntactisch patroon:	NP_0 *zoals* $\{NP_1, NP_2\ \ldots,\ (en\|of)\}$ NP_n *voor alle* $NP_i,\ 1 \leq i \leq n,$ *hyponiem* (NP_i, NP_0)
Zin in corpus:	Huisdieren, zoals honden en katten, zijn niet toegestaan.
Conclusie:	hyponiem("honden","huisdieren") hyponiem("katten","huisdieren")

Deze lexicaal-syntactische patronen worden niet alleen gebruikt voor de identificatie van nieuwe hyponiemparen, maar ook voor het uitbreiden van het

[2]'Vrije tekst' is tekst uit bijvoorbeeld boeken en kranten, die niet speciaal gestructureerd is voor computationele doelen. Deze tekst bevat mogelijk typefouten, spelfouten of ongrammaticale zinnen.

aantal hyponiemen onder elk hyperniem (bijv. m.b.v. coördinatiepatronen, zie Widdows & Dorow 2002, Roark & Charniak 1998, Riloff & Jones 1999).

De hierboven beschreven methode voor automatische extractie van hyponiemparen, gebaseerd op lexicaal-syntactische patronen, neemt een centrale plaats in in verschillende methodes voor de opbouw of uitbreiding van ontologieën (Alfonseca & Manandhar 2002, Caraballo 1999, Rydin 2002).

In deze scriptie staat de ontwikkeling van twee methodes voor automatische extractie van hyponiemrelaties uit Nederlandse corpora, m.b.v. lexicaal-syntactische patronen, centraal. Het gaat hierbij om extractie van hyponiemrelaties uit vrije tekst en extractie uit een gestructureerde encyclopedie. De ontwikkelde methodes kunnen vervolgens gebruikt worden bij de automatische uitbreiding van bestaande Nederlandse ontologieën zoals EuroWordNet.

Hoewel deze scriptie zich zal beperken tot onderzoek naar een automatische extractiemethode voor hyponiemrelaties, zullen bestaande methodes voor de uitbreiding van ontologieën wel worden besproken en zullen enkele aanbevelingen worden gedaan voor de te ontwikkelen Nederlandse methode.

1.4 Opbouw van de scriptie

In hoofdstuk 2 zal relevant onderzoek op het gebied van uitbreiding van ontologieën worden besproken. Zowel methodes voor de extractie van hyponiemrelaties uit vrije tekst als het plaatsen van deze hyponiemrelaties in een bestaande ontologie zullen besproken worden.

Het eigen onderzoek komt aan bod in hoofdstuk 3. Aan de hand van goede en slechte punten van de besproken onderzoeken wordt een methode voor hyponiemextractie voor het Nederlands opgesteld. Vervolgens wordt een korte beschrijving gegeven van alle stappen van het onderzoek.

In het vierde hoofdstuk wordt de Nederlandse methode voor extractie van hyponiemparen m.b.v. lexicaal-syntactische patronen ontwikkeld. Hier komt zowel de extractie uit vrije tekst als de extractie uit de meer gestructureerde encyclopedie aan de orde. Daarnaast wordt besproken hoe de hyponiemparen m.b.v. stemming uniform kunnen worden opgeslagen.

De resultaten van de extractiemethodes komen in hoofdstuk 5 aan de orde, waarin door middel van vergelijking met EuroWordNet en een steekproef beoordeeld door menselijke beoordelaars de methodes geëvalueerd worden.

In hoofdstuk 6, de conclusie en discussie, wordt kritisch bekeken of de voor het Nederlands ontwikkelde methode voor de extractie van hyponiemparen bij kan dragen aan het uitbreiden van de bestaande ontologieën zoals EuroWordNet. Daarnaast worden enkele aanbevelingen gedaan voor de daadwerkelijke uitbreiding van de ontologie, waarbij gebruik gemaakt wordt van de automatische extractie van hyponiemrelaties.

Hoofdstuk 2

Eerder onderzoek

Hoewel de vraag naar het automatisch kunnen genereren of uitbreiden van ontologieën pas de laatste jaren duidelijk is opgekomen, publiceerde Hearst al in 1992 haar onderzoek naar de automatische acquisitie van hyponiemrelaties. Tot op heden wordt de door Hearst beschreven methode, het extraheren van hyponiemrelaties uit vrije tekst m.b.v. lexicaal-syntactische patronen, nog vaak toegepast in onderzoek naar het automatisch genereren of uitbreiden van ontologieën.

In dit hoofdstuk wordt relevant onderzoek op het gebied van de uitbreiding van ontologieën besproken. Hierbij zullen de verschillende aspecten van het bouwen van een ontologie aan bod komen: het verzamelen van hyponiemparen m.b.v. lexicaal-syntactische patronen, het gebruik van algoritmes voor het identificeren van coördinaties van nomina, en het bouwen of uitbreiden van de ontologie. Hoewel deze laatste twee aspecten niet aan bod zullen komen in het eigen onderzoek wordt wel een kritische blik geworpen op de bestaande methodes.

2.1 Literatuur

2.1.1 Extractie van hyponiemrelaties uit vrije tekstcorpora

Lexicaal-syntactische patronen

Hearst (1992,1998) beschrijft een methode om met behulp van lexicaal-syntactische patronen hyponiemrelaties in domein-onafhankelijke teksten te

identificeren.
Beschouw de volgende zin (Hearst 1992):

> The *bow lute*, such as the *Bambara ndang*, is plucked and has an individual curved neck for each string.

Hearst (1992) geeft aan dat vloeiende lezers van het Engels die de term *Bambara ndang* nog nooit eerder hebben gehoord, toch zonder veel moeite zullen concluderen dat een *Bambara ndang* een soort van *bow lute* is.

Dit komt doordat semantiek van de lexicaal-syntactische constructie:

$$NP_0 \; such \; as \; \{NP_1, NP_2 \ldots, \; (and|or)\} \; NP_n$$

veronderstelt dat:

$$voor \; alle \; NP_i, 1 \leq i \leq n, \; hyponiem \; (NP_i, NP_0)$$

En hieruit kan worden geconcludeerd dat

hyponiem("Bambara ndang", "bow lute").

Hearst veronderstelt dat er meer van dit soort lexicaal-syntactische constructies zijn die hyponiemrelaties aanduiden. Om deze patronen, en dus ook nieuwe hyponiemrelaties, automatisch te identificeren, heeft zij het volgende algoritme bedacht:

1. Verzamel een lijst van hyperniem-hyponiemparen; handmatig, uit een bestaande ontologie of automatisch door dit algoritme.
2. Doorzoek een corpus naar zinnen waarin deze paren syntactisch dicht bij elkaar voorkomen, en sla de omgeving van het paar op.
3. Zoek gemeenschappelijke kenmerken in de (syntactische) omgeving, en identificeer deze als nieuwe lexicaal-syntactische patronen.
4. Gebruik deze patronen om nieuwe hyperniem-hyperniemparen te vinden, en ga verder bij stap 2.

Hearst (1992) geeft aan dat, op basis van handmatige evaluatie van relaties gevonden door het hierboven besproken *such as*-patroon, geconcludeerd kan worden dat de kwaliteit van de gevonden relaties erg hoog is. Hearst (1998) laat zien dat op basis van het *or other*-patroon 63% van de gevonden relaties correct is. Hearst (1992,1998) geeft echter wel aan dat de methode moeilijkheden kent, met name op het gebied van metonymie (deel-geheel relaties), onderspecificatie en contextgevoeligheid. Daarnaast geeft Hearst (1998) aan dat het soort corpus van invloed kan zijn op de resultaten: krantentekst is vaak meer beïnvloed door de context dan encyclopedietekst. Ook bevat krantentekst vaak meer subjectieve oordelen, meningen, en metaforisch taalgebruik.

Morin en Jacquemin (2002) geven enkele aanpassingen van de methode van Hearst (1992,1998). Zo willen zij het proces nog verder automatiseren door in de derde stap van de methode de gemeenschappelijke kenmerken van de omgevingen automatisch te meten. Daarnaast geven ze aan dat de gevonden lexicaal-syntactische patronen en hyponiemparen tussentijds door een expert geëvalueerd moeten worden. Dit wordt door Hearst (1992,1998) niet expliciet genoemd.

Latent Semantic Analysis

Cederberg en Widdows (2003) maken ook gebruik van de methode van Hearst (1992), maar voegen hier Latent Semantic Analysis (LSA) aan toe. LSA is gebaseerd op het principe van Semantic Similarity: semantisch gelijke woorden hebben dezelfde distributie en komen in dezelfde contexten voor (Miller & Charles 1991).
Cederberg en Widdows meten voor elke door de methode van Hearst geïdentificeerde hyponiemrelatie in hoeverre het hyperniem en het hyponiem in dezelfde context voorkomen. Hiervoor hebben zij 1000 frequent in de context voorkomende woorden gekozen, waarbij zij voor elk van deze woorden bepalen hoe vaak deze in de context van het hyperniem of het hyponiem voorkomen. Vervolgens meten zij voor elk paar in hoeverre deze contexten overeen komen, en hoe groot de gelijkheid tussen hyponiem en hyperniem dus is. Van de 513 paren die zij met de methode van Hearst hebben geëxtraheerd, hebben zij de 100 paren met de hoogste gelijkheid handmatig geëvalueerd. Hiervan bleek 58% correct te zijn. Zonder toepassing van LSA was dit, gemeten aan de hand van 100 random gekozen paren, 40%.
Nadeel van deze methode is dat onjuiste hyponiemparen die wel semantisch verwant zijn (bijv. hyponiem ("hond","kat")) een hoge gelijkheid kennen

en dus als waarschijnlijk hyponiempaar worden aangemerkt.

Coördinatie van nomina

Naast LSA maken Cederberg en Widdows (2003) ook gebruik van algoritmes voor de identificatie van coördinaties van nomina om zo het aantal hyponiemrelaties uit te breiden.
Beschouw de volgende zin:

> This is not the case with sugar, honey, grape must, cloves and other spices which increase its merit.

Deze zin laat zien dat *clove*, net als *sugar*, *honey* en *grape must*, een soort van *spice* is. Hiervan uitgaand, kan uit de volgende zin

> Ships laden with nutmeg or cinnamon, cloves or coriander once battled the Seven Seas to bring home their precious cargo.

worden geconcludeerd dat ook *nutmeg*, *cinnamon* en *coriander* soorten van *spices* zijn. Deze inferentie is gebaseerd op het principe dat woorden die samen in lijsten voorkomen, vaak semantisch gelijke kenmerken hebben. Door voor elk gevonden hyponiem z'n 'soortgenoten' te zoeken, creëren Cederberg en Widdows (2003) nieuwe hyponiemrelaties.
Zij evalueerden deze methode door 45 correcte hyponiemparen als basis te kiezen, en voor elk hyponiem 10 semantisch gelijke woorden te zoeken m.b.v. de patronen voor coördinatie van nomina. Deze woorden vormen vervolgens nieuwe hyponiemparen met de bestaande hyperniemen. Van deze nieuwe paren was 46% correct. Door op deze paren ook weer LSA toe te passen, werd de precisie verhoogd naar 64%.
Het gebruik van coördinatie van nomina (zoals ook door Riloff & Shepherd 1997 en Roark & Charniak 1998) is geschikt om het aantal hyponiemparen uit te breiden, maar kent een groot nadeel: ambiguïteit van de *seed words*[1]. Doordat slechts één hyponiem per keer als input dient, is het bij ambigue hyponiemen onmogelijk de juiste betekenis te selecteren.

Het algoritme voor coördinatie van nomina van Widdows en Dorow (2002), waarmee semantisch sterk verwante woorden geïdentificeerd kunnen worden, kan wel goed omgaan met ambiguïteit, doordat het gebruik maakt van meerdere *seed words*. Ook deze methode is gebaseerd op het principe van

[1] Dit zijn de woorden die als input voor het algoritme gebruikt worden.

Semantic Similarity (Miller & Charles 1991). Aan de hand van lexicaal-syntactische patronen identificeren Widdows en Dorow (2002) nomina die voorkomen in coördinatie met de *seed words*. Vervolgens wordt voor elk van de gevonden nomina bekeken met hoeveel verschillende *seed words* er een coördinatierelatie is. De meest gelijke nomina zijn de nomina met de meeste verschillende relaties met de *seed words*. Er wordt dus gekeken naar *type frequency* in plaats van *token frequency*, waardoor de effecten van ambiguïteit geminimaliseerd worden:

> Suppose that we start with the seed-list *apple, orange, banana*. However many times the string "Apple and Novell" occurs in the corpus, the *novell* node will not be added to this list because it doesn't have a link to *orange, banana* or any of their neighbours except for *apple*.(Widdows & Dorow 2002, p. 1095)

De correct geïdentificeerde nomina kunnen vervolgens toegevoegd worden aan de *seed list*, en er kan opnieuw gezocht worden naar semantisch gelijke woorden.
Widdows en Dorow evalueerden deze methode door te vergelijken met WordNet: zij selecteerden categorieën (=hyperniemen) uit WordNet, en kozen uit elke categorie het meest prototypische hyponiem. Vervolgens zochten zij bij elk hyponiem twintig semantisch gelijke woorden die in dezelfde categorie horen. Deze hyponiemen werden vergeleken met de hyponiemen in desbetreffende categorie in WordNet, en daaruit is gebleken dat de methode een precisie van 82% haalt.
Nadeel van deze methode is echter dat steeds de meest waarschijnlijke en stabiele vormen worden gekozen, waardoor er veel correcte, maar minder frequent met de *seed words* gecoördineerde nomina niet opgenomen zullen worden. Dit zal dus ten koste gaan van de *recall*.

2.1.2 Het bouwen/uitbreiden van een ontologie

Lexicaal-syntactische patronen en coördinatie van nomina

Caraballo (1999) maakt net als Cederberg en Widdows (2003) gebruik van de methode van Hearst (1992,1998) en van coördinatie van nomina, maar doet dit in de omgekeerde volgorde: zij identificeert eerst gecoördineerde nomina, en clustert deze vervolgens *bottom up* tot een hiërarchie en voorziet de interne knopen van hyperniemlabels m.b.v. de methode van Hearst (1992,1998).

Caraballo maakt voor de identificatie van de nomina gebruik van het al eerder besproken principe van Semantic Similarity, en gaat er van uit dat nomina die in coördinatie voorkomen semantisch aan elkaar gerelateerd zijn. Zij geeft aan dat deze semantische gelijkheid ook geldt voor nomina en apposities[2]. Nadat zij met behulp van lexicaal-syntactische patronen voor coördinatie van nomina alle gewenste nomina geïdentificeerd heeft, creëert zij voor elk gevonden nomen een vector waarin opgeslagen wordt hoe vaak elk van de andere nomina voorkomt met desbetreffend nomen. Vervolgens berekent zij welke nomina het meest gelijk zijn aan elkaar en koppelt zij deze aan elkaar. Door de gekoppelde nomina met gezamenlijke ouder weer te linken aan een ander meest gelijk nomen of andere meest gelijke nominagroep, wordt *bottom up* een hiërarchische structuur opgebouwd.
Vervolgens probeert zij aan de hand van de methode van Hearst (1992,1998) voor elke interne knoop het juiste hyperniemlabel te vinden (waarbij twee andere correcte labels, indien aanwezig, ook worden opgeslagen). Het meest juiste label voor een interne knoop is het hyperniem met het hoogste aantal verschillende relaties met alle afstammelingen van de interne knoop. Omdat niet voor alle hyponiemen een hyperniem gevonden zal worden, zal niet elke interne knoop een hyperniem toegewezen krijgen. Vervolgens wordt de hiërarchie samengeperst: als een interne knoop geen hyperniemlabel heeft, of hetzelfde label heeft als z'n ouder, wordt de knoop verwijderd en worden z'n afstammelingen onder de ouder gehangen.

Deze methode kent enkele grote nadelen, die door Caraballo (1999) ook besproken worden: Ten eerste zit er veel overlap in de structuur. Labels die hoog in de boom voorkomen, worden vaak bij lagere interne knopen nog een keer gebruikt. Dit wordt veroorzaakt doordat labels toegekend worden op basis van alle afstammelingen. Het zou beter zijn de directe afstammelingen zwaarder mee te laten wegen.
Daarnaast kan de hyperniemtoekenning beter. Caraballo geeft aan dat het aantal lexicaal-syntactische patronen uitgebreid moet worden, en dat de hyperniemen een striktere selectie moeten ondergaan, zodat correctere hyperniemen aan de interne knopen kunnen worden toegewezen.

Caraballo evalueert de hiërarchie door tien hyperniemen te kiezen die minstens 20 hyponiemen hebben, en selecteert voor elk hyperniem 20 hyponiemen. Deze laat zij door drie beoordelaars evalueren. Uit deze evaluatie is

[2]Caraballo (1999, p. 120) geeft hierbij de voorbeelden "Boeing, a defense contractor, ..." en "James H. Rosenfield, a former CBS Inc. executive". Hieruit leidt zij af dat er een gelijkheidsrelatie bestaat tussen *Boeing* en *defense contractor*, en tussen *James H. Rosenfield* en *a former CBS Inc. executive*.

gebleken dat in 39% van de gevallen één van de beoordelaars de hyponiemrelatie als correct beoordeeld. Als niet alleen naar het beste hyperniemlabel, maar ook naar de twee andere opgeslagen labels wordt gekeken (d.w.z. één van de drie labels moet correct zijn), stijgt dit percentage naar 60,5%

Lexicaal-syntactische patronen en het Distributional Semantics model

Alfonseca en Manandhar (2002) combineren de methode van Hearst (1992,-1998) met het Distributional Semantics model en voegen zo nieuwe *synsets*, dit zijn sets van synonieme woorden, op de juiste plaats in een bestaande ontologie toe.
Het Distributional Semantics model (Rajman & Bonnet 1992, beschreven in Alfonseca & Manandhar 2002), dat sterk verwant is aan LSA, veronderstelt dat er sterke correlatie is tussen de semantiek van een woord en de set van contexten waarin het woord verschijnt. Aan de hand van de *topic signature*, dat is de lijst van woorden die in de context van een woord w voorkomen met de bijbehorende frequenties/gewichten, wordt elke onbekende synset vergeleken met de hyperniemen in de bestaande ontologie en wordt de gelijkheidswaarde berekend. Daarnaast wordt met behulp van lexicaal-syntactische patronen voor elk woord uit de synset en het hyperniem in de ontologie bekeken of er een hyponiemrelatie bestaat in het gebruikte tekstcorpus (aan de hand van een bewerking van de methode van Hearst (1992)).Op basis van de gevonden gelijkheidswaarde en hyponiemrelatie wordt bepaald waar de onbekende synset in de ontologie aangehecht zal worden.
Het algoritme loopt hiervoor *top-down* door de ontologie, en bekijkt steeds of de onbekende synset meer gelijkheid vertoont met het hyperniem of met een van de afstammelingen van het hyperniem. Als de gelijkheid met het hyperniem groter is, wordt de onbekende synset aangehecht als hyponiem. Als de gelijkheid met een of meer van de afstammelingen groter is, gaat het algoritme verder met nu de meest gelijke afstammeling als hyperniem.
Alfonseca en Manandhar geven aan dat de combinatie van de twee methodes zorgt voor dubbele objectiviteit. Doordat deze methode alleen bestaande hyperniemen in de ontologie met de onbekende synset vergelijkt, zullen de door de lexicaal-syntactische patronen onjuist gevonden hyperniemen nooit meegenomen worden in de analyse. Daarnaast helpen de lexicaal-syntactische patronen het top-down algoritme om te beslissen als er twee mogelijke labels zijn die in gelijke contexten voorkomen, zoals de concepten *man* en *vrouw*.
Deze methode kent echter ook een aantal nadelen. Ten eerste is tussen-

plaatsing van hyponiemen niet mogelijk. Stel dat er een vertakking is met als hyperniem *dier* en met als hyponiemen *Duitse Herder* en *Labrador*. De onbekende synset *hond* zal waarschijnlijk aangehecht worden als hyponiem van dier, waardoor *hond* een zusterrelatie krijgt met *Duitse Herder* en *Labrador*, terwijl deze laatste twee begrippen eigenlijk geclassificeerd zouden moeten worden als hyponiemen van *hond*. Hieruit komt voort dat de volgorde van aanhechting van nieuwe synsets een rol speelt: als *hond* eerder was aangehecht dan de twee hondenrassen, zou de hyponiemrelatie wel correct in de ontologie geplaatst zijn. Een tweede nadeel is dat een onbekende synset op de 'best mogelijke' plaats zal worden aangehecht. Als de basisontologie niet al te uitgebreid is, kan het voorkomen dat een synset onder een veel te algemeen hyperniem geplaatst worden. Doordat tussenplaatsing niet mogelijk is, kan dit later niet verder gespecificeerd worden. Tenslotte zal de methode van Hearst niet altijd een bijdrage kunnen leveren, omdat niet voor alle synsets hyperniemrelaties gevonden zullen worden.

Lexicaal-syntactische patronen en het samenvoegen van klasses

Rydin (2002) maakt voor het Zweeds ook gebruik van de methode van Hearst (1992,1998) om hyponiemparen te verzamelen. Vervolgens deelt ze deze hyponiemparen in in klasses, en bouwt ze met deze klasses een ontologie met hiërarchische structuur.
Deze indeling in klasses gebeurt in de volgende stappen (Rydin 2002):

1. Elk hyponiempaar vormt een eigen klasse
2. Twee klasses kunnen samengevoegd worden onder de volgende voorwaardes:
    - de hyperniemen zijn hetzelfde
    - er is een niet-lege intersectie tussen de hyponiemen van beide klasses

Vervolgens wordt met deze klasses een hiërarchische structuur gebouwd. De belangrijkste voorwaarde hierbij is:

> Als X een soort van Y is, en Y een soort van Z, dan kunnen deze gekoppeld worden als er in het lexicon een hyponiemrelatie te vinden is tussen X en Z.

Beschouw de volgende klasses:

Klasse 1: dier - konijn, cavia, hamster, kat
Klasse 2: dier - dwerghamster, muis, rat
Klasse 3: hamster - dwerghamster, Russische hamster

X = dwerghamster, Y = hamster, en Z = dier.
Er bestaat hier een hyponiemrelatie tussen X en Y, Y en Z, en ook X en Z. Koppeling van deze klassen levert op:

Klasse 1: dier - konijn, cavia, hamster, kat , muis, rat
Klasse 2: hamster - dwerghamster, Russische hamster

Door deze compositie van klasses toe te passen waar dat mogelijk is, ontstaat een hiërarchische structuur.

Uit handmatige beoordeling van 20% van de geëxtraheerde hyponiemparen bleek dat zo'n 92% correct was. Foutieve hyponiemparen werden volgens Rydin vooral veroorzaakt door incorrecte woordsoorttoekenning en betekenisverandering door PP-aanhechting. Opvallend is dat de resultaten van Rydin beduidend beter zijn dan die van Hearst (1992,1998). Rydin geeft hier echter geen verklaring voor. Bij evaluatie van de geconstrueerde hiërarchische structuur bleek dat de beoordeling van de paren in de hiërarchische structuur door menselijke beoordelaars niet makkelijk is, en dat er grote onderlinge verschillen waren. Het gemiddelde kwam uiteindelijk uit op 62,5%. In 82.2% van de gevallen vond één van de beoordelaars een hyponiempaar correct, en in 71.0% van de gevallen twee beoordelaars.
Rydin (2002) noemt als een van de belangrijkste nadelen dat er een groot tekort is aan hyponiemrelaties, waardoor veel synoniemen nog niet samengevoegd kunnen worden, en het lijkt alsof veel woorden meerdere betekenissen hebben. Daarnaast bleek dat lemmatisatie en stemming niet altijd voor goede resultaten zorgen: hoewel de meervoudsvorm *boys* wel een hyponiem is van *group*, geldt dit niet voor de gestemde en gelemmatiseerde vorm *boy*.
Er zijn echter ook enkele belangrijke voordelen van deze methode. Rydin (2002) geeft aan dat het een onafhankelijke methode is, en dat na de vergaring van de hyponiemparen geen extra informatie (uit bijv. een corpus) meer nodig is. Daarnaast kan een bestaande hiërarchie als basis worden genomen, en nieuwe relaties kunnen eenvoudig worden toegevoegd. Tussenplaatsing is hier dus, in tegenstelling tot de methode van Alfonseca en Manandhar (2002), wel mogelijk.

Hoofdstuk 3

Onderzoeksopzet

Het Nederlands lijkt, net als het Engels, lexicaal-syntactische patronen te kennen waarmee hyponiemrelaties kunnen worden geïdentificeerd:

> Huisdieren, zoals honden, katten en konijnen, zijn in ons hotel niet toegestaan.
> Patroon: NP_0 *zoals* $\{NP_1 NP_2 \ldots (en|of)\}$ NP_n
> $Voor\ alle\ NP_i\ 1 \leq I \leq n\ geldt:\ hyponiem\ (NP_i, NP_0)$

> Honden, katten, konijnen en andere soorten huisdieren, zijn bij ons niet welkom.
> Patroon: NP_1 $\{NP_2 NP_3 \ldots\}$ $(en|of)$ *andere* NP_n
> $Voor\ alle\ NP_i\ 1 \leq I \leq n\ geldt:\ hyponiem\ (NP_i, NP_n)$

In dit hoofdstuk zal een methode worden besproken waarmee deze hyponiemrelaties automatisch geïdentificeerd kunnen worden in grote tekstcorpora. Deze extractie levert uiteindelijk een lijst met hyponiemparen op, welke gebruikt kunnen worden voor de automatische uitbreiding van Nederlandse ontologieën.

In paragraaf 1 van dit hoofdstuk wordt bekeken in hoeverre de in hoofdstuk 2 besproken bestaande methodes voor automatische uitbreiding van ontologieën (na bewerking) bruikbaar kunnen zijn voor het Nederlands. Vervolgens worden in paragraaf 2 alle stappen van het onderzoek weergegeven en kort uitgelegd. Tenslotte worden in de derde paragraaf de beschikbare corpora en enkele technische specificaties van de corpora besproken.

3.1 Toepassing van bestaande methodes

In dit onderzoek is er voor gekozen om de automatische extractie van hyponiemparen te baseren op de methode van Hearst (1992,1998), omdat er in het Nederlands, net als in het Engels, Zweeds en Frans, lexicaal-syntactische patronen lijken te zijn waarmee hyponiemrelaties aangeduid kunnen worden. Hoewel de resultaten voor het Engels, Zweeds en Frans verschillen laten zien, blijven de resultaten veelbelovend en lijkt toepassing voor het Nederlands, dat syntactisch gezien lijkt op het deze talen, erg zinvol.

De lexicaal-syntactische patronen kunnen zowel handmatig als semi-automatisch worden verkregen, zoals beschreven door Hearst (1992,1998). Voor de identificatie van de Nederlandse patronen zullen beide methodes gebruikt worden. Het verder automatiseren van de methode voor automatische identificatie, zoals voorgesteld door Morin en Jacquemin (2003), lijkt niet noodzakelijk, aangezien het slechts zal gaan om een gering aantal lexicaal-syntactische patronen. De handmatige beoordeling van de gevonden patronen bij de semi-automische methode is niet veel werk, en deze beoordeling is waarschijnlijk accurater dan automatische evaluatie.

Cederberg en Widdows (2003) geven aan dat Latent Semantic Analysis (LSA) de precisie van de geïdentificeerde hyponiemparen kan verhogen. Hoewel deze methode als nadeel kent dat foutieve maar wel semantisch verwante hyponiemparen niet als foutief worden aangemerkt, worden de meeste foutieve en niet semantisch verwante hyponiemparen uit de resultaten gefilterd, waardoor de precisie toeneemt. Hoewel het gebruik van LSA in dit onderzoek niet aan bod zal komen, lijkt deze methode met behulp van de door Cederberg en Widdows (2004) ontwikkelde *Infomap NLP software* eenvoudig te kunnen worden geëvalueerd.

Uit de literatuur blijkt dat er 3 methodes zijn voor het automatisch bouwen van een hiërarchie: Caraballo (1999), Alfonseca en Manandhar (2002) en Rydin (2002).

De resultaten van de methode van Caraballo (1999), die gebruik maakt van coördinatie van nomina en het algoritme van Hearst, motiveren niet om deze methode voor het Nederlands te gebruiken. De output van het algoritme is een ontologie waarbij veel interne knopen geen label hebben. Knopen die wel een label hebben, hebben er vaak meerdere. Daarnaast komen labels vaak meerdere malen in een boom voor. Hierdoor lijkt deze ontologie niet geschikt voor praktische toepassingen.
Voor het uitbreiden van bestaande ontologie is deze methode niet geschikt,

omdat Caraballo (1999) de hiërarchische structuur bottom up construeert. Uitbreiden van bestaande ontologie lijkt niet mogelijk met deze methode.

Ook het toepassen van de methode van Alfonseca en Manandhar (2002) voor uitbreiding van Nederlandse ontologieën is niet zinvol. De grootste Nederlandse ontologie, EuroWordNet, is nog vrij beperkt qua woorden en relaties, waardoor synsets te snel onder algemeen hyperniem zullen worden geplaatst (zoals uitgelegd in hoofdstuk 2). Doordat tussenplaatsing onmogelijk is, kan dit later niet hersteld worden, en zal het hiërarchische karakter van het Nederlandse WordNet verloren gaan. Daarnaast zijn de resultaten die Alfonseca & Manandhar met deze methode hebben behaald ook niet erg overtuigend: slechts 37% van de synsets werd correct aangehecht.

De methode van Rydin (2002) lijkt het meest bruikbaar voor het Nederlands. De methode heeft als belangrijk voordeel dat een bestaande hiërarchie als basis kan worden genomen, en dat nieuwe relaties eenvoudig tussen geplaatst of aangehangen kunnen worden. Het grote nadeel van deze methode is echter wel dat een groot aantal hyponiemrelaties, en dus een groot corpus, benodigd is. Uit dit onderzoek zal blijken of uit de beschikbare corpora genoeg hyponiemparen kunnen worden geëxtraheerd. Bij voldoende hyponiemparen lijkt deze methode het meest geschikt voor uitbreiding van een Nederlandse ontologie zoals EuroWordNet.

Coördinatie van nomina tenslotte is erg geschikt om de bestaande hyponiemrelaties in WordNet uit te breiden. Het toepassen van de methode van Widdows en Dorow (2002) ligt voor de hand, omdat deze methode goed om kan gaan met ambiguïteit, en de precisie mede daardoor erg hoog is. Het uitbreiden van een hiërarchie m.b.v. coördinatie van nomina is echter een geheel op zichzelf staand proces, en het zal in dit onderzoek verder niet aan de orde komen.

3.2 Overzicht van alle onderzoeksstappen

Het eigen onderzoek is onder te verdelen in de volgende stappen:

1. Automatische extractie van hyponiemrelaties m.b.v. lexicaal-syntactische patronen.
   Deze stap valt uiteen in twee onderdelen: extractie uit vrije tekst en extractie uit een meer gestructureerde encyclopedie.

   - Extractie uit vrije tekst
     Hierbij is het opstellen van de lexicaal-syntactische patronen de

eerste stap. Deze patronen zullen zowel handmatig als semi-automatisch worden geïdentificeerd a.d.h.v. de door Hearst (1992,-1998) beschreven methodes.
Vervolgens worden de gevonden patronen gebruikt voor de extractie van hyponiemparen uit een ontwikkelcorpus. Korte evaluatie van deze paren wijst uit welke patronen geschikt zijn, en welke aanpassingen aan de patronen nodig zijn. Tenslotte wordt besproken waarom stemming van de hyponiemparen nodig is, en wat de invloed hiervan op de resultaten is.

- Extractie uit de encyclopedie
  Uit bestudering van het encyclopedie-corpus bleek dat de beschrijvingen van de begrippen in de encyclopedie veelal dezelfde structuur kennen: elke eerste zin van de beschrijving geeft in een overgroot deel van de gevallen weer tot welke supercategorie een begrip gerekend moet worden.
  Twee voorbeelden:

  > 2000-syndroom – *Computerprobleem* dat gepaard gaat met de overgang naar een nieuw millennium.
  > aal – *Europese palingen* ('Anguilla anguilla'), geslacht palingen.

  Deze supercategorieën zijn in feite de hyperniemen, en het verklaarde begrip zelf is het hyponiem. Om deze relaties te extraheren is handmatig één lexicaal-syntactisch patroon opgesteld, dat kort geëvalueerd wordt op het ontwikkelcorpus. Op basis van evaluatie van deze geïdentificeerde paren wordt het patroon aangepast.

2. Evaluatie van de lexicaal-syntactische patronen.
   De evaluatie van de lexicaal-syntactische patronen gebeurt op twee manieren:

   - Beoordeling van een steekproef door drie menselijke beoordelaars.
     Een steekproef van de resultaten van beide extractiemethodes zal beoordeeld worden door drie personen. Vervolgens wordt bekeken hoe hoog de precisie is, en hoe hoog de onderlinge overeenstemming tussen de boordelaars is.
   - Vergelijking met de hyponiemrelaties in EuroWordNet.
     Voor beide extractiemethodes wordt voor alle gextraheerde hyponiemparen bekeken of deze ook in EuroWordNet voorkomen,

en of er een hyponiemrelatie bestaat tussen hyponiem en hyperniem. Vervolgens wordt besproken wat de resultaten van deze vergelijking betekenen.

Tenslotte wordt voor beide extractiemethodes onderzocht wat de voornaamste oorzaken zijn van foutieve hyponiemparen.

3.3 Beschikbare corpora

Voor de extractie van hyponiemrelaties zijn tekstcorpora nodig. Voor dit onderzoek zijn de volgende corpora beschikbaar:

Volkskrantcorpus	851.236 zinnen
Delen van het Twente Nieuws Corpus (TwNC):	
AD, jaargangen 1994, 1995 en 2000	2.339.048 zinnen
Trouw, jaargangen 1999 en 2001	784.002 zinnen
NRC, jaargangen 1994 en 1995	1.802.450 zinnen
Spectrum Encyclopedie	358.195 zinnen

Het Volkskrantcorpus zal als ontwikkelcorpus dienen bij het opstellen van de lexicaal-syntactische patronen.

Alle zinnen in de corpora zijn geanalyseerd en ontleed m.b.v. Alpino. Alpino maakt computationele analyses van Nederlandse zinnen en wijst vervolgens de beste ontleding voor elke zin aan. De analyses zijn gebaseerd op een grammatica voor het Nederlands die getraind is op een syntactisch geannoteerd corpus (Bouma, van Noord & Malouf, 2001).

Alpino geeft voor elk woord in elke ontlede zin informatie over de woordsoort, de syntactische categorie en de functie van het woord in de zin. De ontlede zinnen worden samen met deze informatie elk in een apart document in XML-formaat opgeslagen.

Hieronder volgt een voorbeeld van deel van een door Alpino ontlede zin, opgeslagen in XML-formaat:

```
<node rel="top" cat="smain" hd="9">
 <node rel="su" cat="np" hd="2">
  <node rel="det" cat="detp" pos="det" hd="1" root="sommige" word="Sommige"/>
  <node rel="hd" pos="noun" hd="2" root="kaas" word="kazen"/>
  <node rel="mod" cat="cp" begin="3" end="7" hd="4">
   <node rel="cmp" pos="comp" begin="3" end="4" hd="4" root="zoals" word="zoals"/>
    <node rel="body" cat="conj" begin="4" end="7" hd="6">
     <node rel="cnj" cat="np" pos="noun" begin="4" end="5" hd="5" root="cantal" word="cantal"/>
     <node rel="crd" pos="vg" begin="5" end="6" hd="6" root="en" word="en"/>
     <node rel="cnj" cat="np" pos="noun" begin="6" end="7" hd="7" root="vacherin" word="vacherin"/>
    </node>
   </node>
  </node>
 <node rel="hd" pos="verb" begin="8" end="9" hd="9" root="krijg" word="krijgen"/>
 <node rel="obj1" cat="np" begin="9" end="18" hd="13">
 ...
 </node>
</node>
```

De boomstructuur van de volledige boom is weergegeven is figuur 3.1.

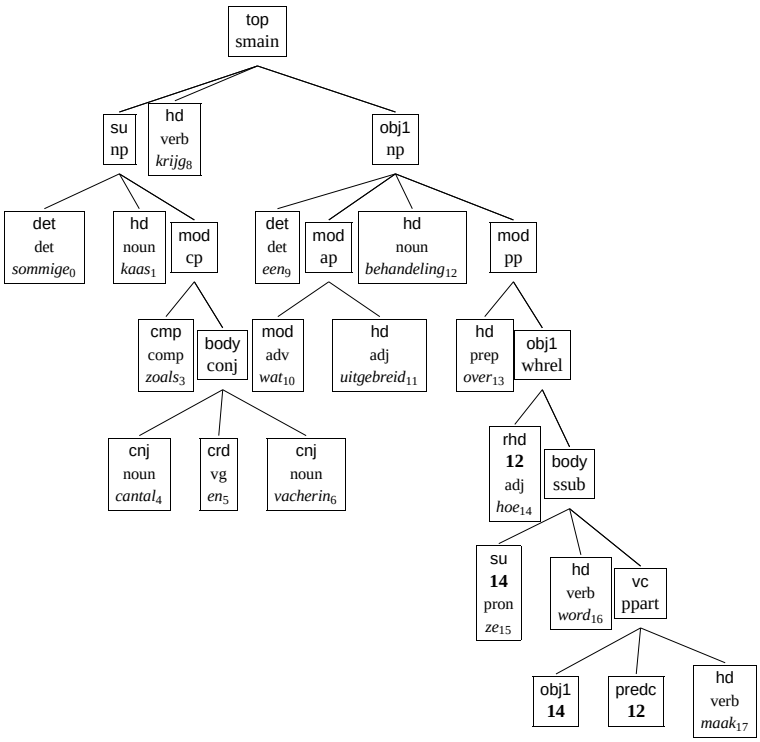
Doordat alle zinnen als boomstructuren worden opgeslagen, zijn *dependency* relaties tussen verschillende zinsdelen eenvoudig terug identificeren.
Zo zijn uit bovenstaand voorbeeld onder andere de volgende relaties te identificeren:

- Dependency relaties tussen hoofden
  Ondanks dat *kazen* en *krijgen* niet direct lineair op elkaar volgen, is af te leiden dat het nomen *kazen* het subject is van het werkwoord *krijgen*.

- Coördinatie-relaties tussen nouns
  Cantal en *vacharin* zijn verbonden door coördinatie.

Naast bovengenoemde informatie geeft Alpino ook de *root* van elk woord weer. Deze *root* is echter niet in alle gevallen gelijk aan de stam van een woord. Alpino maakt voor de bepaling van de *root* gebruik van een woordenboek dat stemvormen bevat. Voor woorden die in het woordenboek staan, wordt de juiste stam gegeven:

> *Kleinkinderen*, dat in het woordenboek staat, krijgt als root *kleinkind*.



Figuur 3.1: Boomstructuur voor de zin 'Sommige kazen zoals cantal en vacherin krijgen ...'

Voor onbekende woorden bekijkt Alpino, met behulp van het woordenboek, of ze een samenstelling zijn. Als dit het geval is, wordt de stam van het laatste woord als *root* gegeven:

> *Eenoudergezinnen* staat niet in het woordenboek, maar wordt herkend als een samenstelling.
> Van het laatste deel van deze samenstelling, *gezinnen*, wordt de stamvorm *gezin* als root voor het hele woord genomen.

Onbekende woorden die geen samenstelling lijken te zijn, hebben als *root* het hele woord:

> De root van *CDU'ers* wordt *CDU'ers*.

Hoofdstuk 4

Extractie van hyponiemparen m.b.v. lexicaal-syntactische patronen

In dit hoofdstuk worden twee methodes voor de extractie van hyponiemrelaties uit grote tekstcorpora besproken. De eerste methode extraheert m.b.v. verschillende lexicaal-syntactische patronen hyponiemrelaties uit vrije tekst. De tweede methode maakt gebruik van één lexicaal-syntactisch patroon voor de identificatie van hyponiemrelaties in de gestructureerde tekst van een encyclopedie.

Nadat de technische details voor de extractie zijn besproken, wordt de ontwikkeling van beide methodes besproken.

4.1 Technische specificaties

Voor de extractie van hyponiemrelaties wordt gebruik gemaakt van XSLT, een taal waarmee beschreven kan worden hoe delen van een XML-document getransformeerd kunnen worden naar, in dit geval, een tekstdocument.
In het XSLT-document worden alle lexicale en syntactische eisen waaraan de output moet voldoen opgesteld. Het XSLT-document analyseert vervolgens alle XML-documenten, en geeft alleen de zinnen of zinsdelen die voldoen aan de eisen als output.

Voorbeeld van een kleine XSLT-stylesheet:

```
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output method="text" encoding="ISO-8859-1"/>

<xsl:template match="top">
    <xsl:apply-templates select="./node[@rel='hd' and @pos='verb']"/>
</xsl:template>

<xsl:template match="node[@rel='hd' and @pos='verb']">
    <xsl:text>Hoofdwerkwoord: </xsl:text>
    <xsl:value-of select="@word"/>
</xsl:template>

</xsl:stylesheet>
```

Dit XSLT-document zoekt voor elke zin de knoop direct onder de *top*-knoop die voldoet aan de restrictie dat woordsoort een *verb* is, en de functie van zinsdeel *head* is. Als dit XSLT-document wordt toegepast op het in het vorige hoofdstuk beschreven XML-document zal de output zijn:

> Hoofdwerkwoord: krijgen

Bij de extractie van hyponiemrelaties zal voor elk lexicaal-syntactisch patroon een XSLT-document worden opgesteld, en deze XSLT-documenten selecteren vervolgens de gewenste hyponiemrelaties uit de corpora.

4.2 Extractie uit vrije tekst

4.2.1 Lexicaal-syntactische patronen voor hyponiemrelaties

Zoals beschreven door Hearst (1992, 1998), kunnen lexicaal-syntactische patronen worden gebruikt voor de identificatie van hyponiemrelaties in vrije tekst. Deze methode kan ook voor het Nederlands bruikbaar zijn, omdat het Nederlands ook lexicaal-syntactische patronen lijkt te kennen die hyponiemrelaties beschrijven, zoals:

> Huisdieren, zoals honden, katten en konijnen, zijn in ons hotel niet toegestaan.
> Patroon: NP_0 *zoals* $\{NP_1 NP_2 \ldots (en|of)\}$ NP_n
> Waarbij geldt: *voor alle* NP_i $1 \leq I \leq n$, *hyponiem* (NP_i, NP_0)
>
> Honden, katten, konijnen en andere soorten huisdieren, zijn bij

ons niet welkom.
Patroon: NP_1 $\{NP_2 NP_3 \ldots\}$ *en andere* NP_n
Waarbij geldt: *alle* NP_i $1 \leq I \leq n$ *geldt* : *hyponiem* (NP_i, NP_n)

De identificatie van deze lexicaal-syntactische patronen kan zowel handmatig als semi-automatisch gebeuren. Handmatig wil zeggen dat de patronen worden geïdentificeerd door de tekst door te kijken en patronen op te merken die hyponiemrelaties kunnen aanduiden. De semi-automatische identificatie wordt beschreven door het in hoofdstuk 2 besproken algirtme van Hearst (1992,1998).

In dit onderzoek is er voor gekozen om voor de extractie van hyponiemparen uit Nederlandse corpora een Nederlandse vertaling van de door Hearst gevonden patronen als basis te gebruiken. Dit leverde de volgende patronen op:

1. NP_0 *zoals* $\{NP_1, NP_2 \ldots, (en|of)\}$ NPn
   voor alle NP_i, $1 \leq i \leq n$ *geldt* : *hyponiem* (NP_i, NP_0)
   Duinplanten, zoals de addertong en de duingentiaan ...
   hyponiem("addertong","duinplant")
   hyponiem("duingentiaan","duinplant")

2. NP_1 $\{NP_2 NP_3 \ldots\}$ *en andere* NP_n
   voor alle NP_i $1 \leq I \leq n$ *geldt* : *hyponiem* (NP_i, NP_n)
   ... turnsters, volleyballers en andere sporters.
   hyponiem("turnster","sporter")
   hyponiem("volleyballer","sporter")

3. NP_1 $\{NP_2 NP_3 \ldots\}$ of *andere* NP_n
   voor alle NP_i $1 \leq I \leq n$ *geldt* : *hyponiem* (NP_i, NP_n)
   Malaria, dysenterie, cholera of andere tropische ziektes ...
   hyponiem("malaria","tropische ziekte")
   hyponiem("dysenterie","tropische ziekte")
   hyponiem("cholera","tropische ziekte")

4. NP_0 *inclusief* $\{NP_1, NP_2 \ldots, (en|of)\}$ NPn
   voor alle NP_i, $1 \leq i \leq n$ *geldt* : *hyponiem* (NP_i, NP_0)
   ... de hele familie, inclusief ooms en tantes.
   hyponiem("oom","familie")
   hyponiem("tante","familie")

5. NP_0 *vooral* $\{NP_1, NP_2 \ldots, (en|of)\}$ NPn
   voor alle NP_i, $1 \leq i \leq n$ *geldt* : *hyponiem* (NP_i, NP_0)
   Verkeersdeelnemers, vooral automobilisten en fietsers ...
   hyponiem("automobilist","verkeersdeelnemer")
   hyponiem("fietser","verkeersdeelnemer")

Vervolgens is geprobeerd om m.b.v. de semi-automatische methode van Hearst nieuwe patronen te ontdekken. Hiervoor is met behulp van bovenstaande lexicaal-syntactische patronen een (ongestemde) lijst van waarschijnlijke hyponiemparen gecreëerd. Vervolgens zijn in het ontwikkelcorpus[1] alle zinnen opgezocht die een van de paren bevatten, en is de tussenliggende context opgeslagen. Een voorbeeld:

hyponiem("Nederland","landen")

Verscheidene *landen* , **waaronder** *Nederland*, hebben bij elkaar al zo'n tweehonderd miljoen gulden beloofd , maar dat is volgens de Doema-leden volstrekt onvoldoende .
Context: waaronder

De export kwam niet op gang met *landen* **als** *Nederland* , aangezien de handelspartners van Duitsland verplicht waren met dollars te betalen - die de Nederlanders bijvoorbeeld toen nog niet hadden .
Context: als

Uit alle opgeslagen contexten zijn de meest voorkomende bekeken. Contexten die meer dan 50 keer in het ontwikkelcorpus voorkomen zijn:

- en andere
- zoals
- als
- en
- waaronder

[1] In dit hele onderzoek zal het Volkskrant 97 corpus, dat 851.236 zinnen telt, als ontwikkelcorpus worden gebruikt.

- of andere

Zoals verwacht bevinden de uit het Engels vertaalde patronen zich in dit lijstje. Daarnaast zijn *als*, *waaronder* en *en* mogelijke indicatoren van hyponiemrelaties te zijn.

Hoewel de context *als* hyponiemrelaties kan aanduiden, is deze toch niet geschikt als basis voor een lexicaal-syntactisch patroon:

> Bevriende landen krijgen uitgelegd hoe een *modeldemocratie* als *Duitsland* omgaat met leidend militair personeel.
> *Duitsland* als *modeldemocratie* geeft het goede voorbeeld aan andere landen.

Uit bovenstaande voorbeelden blijkt dat *als* wel een hyponiemrelatie aanduidt, maar dat het onduidelijk is welk woord het hyperniem vormt en welk woord het hyponiem. Omdat beide varianten van bovenstaand gebruik van *als* regelmatig voorkomen, kan deze context niet gebruikt worden als lexicaal-syntactisch patroon.

De context *waaronder* levert vaak goede hyponiemrelaties op:

> Maar een aantal *lidstaten* , waaronder *Duitsland* , heeft al laten doorschemeren dat de Europese Commissie niet te hard van stapel moet lopen .

> Hewlett Packard koos Heerenveen na onderzoek in acht landen en , toen de keuze op Nederland was gevallen , in dertien *steden* (waaronder *Groningen* , *Lelystad* en *Emmen*) .

> De aangesloten *landen* , waaronder *Nederland* , onderschrijven het OESO-lidmaatschap van Rusland ' op termijn ' .

> Moi's opponenten hebben steun gekregen van *mensenrechtenorganisaties* , waaronder *Amnesty International* dat vorige week waarschuwde dat de verkiezingen niets voorstellen zonder grondige hervorming van de grondwet en opheffen van de achterhaalde Public-Order-wetten.

Deze context levert het volgende lexicaal-syntactische patroon op

NP_0 *waaronder* $\{NP_1, NP_2 \ldots,$ $(en|of)\}$ NPn
voor alle NP_i, $1 \leq i \leq n$ *geldt* : *hyponiem* (NP_i, NP_0)

en zal evenals de uit het Engels vertaalde patronen verder worden geanalyseerd.

De context *en* geeft veelal coördinaties aan en is geen indicator van hyponiemrelaties.

4.2.2 Analyse van de opgestelde lexicaal-syntactische patronen

Hoewel de vijf uit het Engels vertaalde lexicaal-syntactische patronen en het automatisch geïdentificeerde lexicaal-syntactische patroon ook bruikbaar lijken voor het Nederlands, moet worden bekeken of zij naast de gewenste relaties niet teveel foutieve relaties opleveren. De patronen zijn daarom kort geëvalueerd op het ontwikkelcorpus.

Met behulp van de eerder beschreven XSLT-documenten zijn voor elk patroon alle zinnen die desbetreffend patroon bevatten geïdentificeerd, en zijn de hyponiemparen opgeslagen. Hierbij werd als restrictie genomen dat het hoofd-nomen[2] een nomen of een naam moet zijn, zodat vormen van persoonlijke voornaamwoorden worden uitgesloten. Van alle hyponiemparen per patroon zijn er door middel van een steekproef vijftig geselecteerd en beoordeeld door de auteur. De resultaten zijn te vinden in tabel 4.1.

Uit de resultaten blijkt dat patroon 1, 2 en 3 regelmatig voorkomen in het corpus en in een ruime meerderheid van de gevallen hyponiemrelaties aanduiden. Deze patronen zullen dan ook in het verdere proces worden gebruikt.

Hoewel patroon 4 117 paren oplevert, blijkt de relatie tussen de elementen vaak niet de gewenste te zijn: meestal is er geen sprake van een hyponiemrelatie, maar wordt een element dat niet tot de klasse van het hyperniem behoort er bij betrokken:

> Nu kost de *woning* inclusief *overdrachtsbelasting* 265 duizend gulden .

[2]Bij deze eerste analyse is gekozen om telkens alleen het hoofd-nomen van de NP op te slaan als hyponiem of hyperniem. In de volgende subparagraaf wordt deze keuze besproken.

	Patroon:	**Totaal aantal:**	**Totaal uniek:**	**Correct:**
1	NP_0 *zoals* $\{NP_1, NP_2 \ldots, (en\|of)\}$ NPn	761	753	64%
2	NP_1 $\{NP_2 NP_3 \ldots\}$ *en andere* NP_n	1581	1488	58%
3	NP_1 $\{NP_2 NP_3 \ldots\}$ *of andere* NP_n	155	155	58%
4	NP_0 *inclusief* $\{NP_1, NP_2 \ldots, (en\|of)\}$ NPn	117	117	28%
5	NP_0 *vooral* $\{NP_1, NP_2 \ldots, (en\|of)\}$ NPn	9	9	0%
6	NP_0 *waaronder* $\{NP_1, NP_2 \ldots, (en\|of)\}$ NPn	1269	874	52%

Tabel 4.1: Aantal voorkomens van lexicaal-syntactische patronen in Volkskrantcorpus en percentage hyponiemparen dat correct is blijkend uit steekproef (n=50).

> Ik kreeg polio , waardoor de linkerhelft van mijn *gezicht*, inclusief mijn *stembanden*, verlamd raakte.
> Maar velen hechten nog geloof aan de belofte van president Sali Berisha om de *beleggingen* inclusief *rente* terug te betalen .
> Het *opleidingsinstituut* van Cessna, inclusief de *loonkosten*, wordt gesubsidieerd door de staat Kansas .
> Geschatte *kosten*, inclusief achterstallig *onderhoud*: zes miljard dollar .

Patroon 5 bleek in het Volkskrantcorpus slechts enkele keren voor te komen, en duidde in alle gevallen geen hyponiemrelatie aan. De patronen 4 en 5 worden vanwege de slechte resultaten niet verder gebruikt voor de extractie van hyponiemparen.

Patroon 6 komt zeer regelmatig voor in het corpus, maar de analyse van de steekproef laat zien dat slechts zo'n 52% van de resultaten correct is. Omdat dit patroon door de hoge frequentie echter ook veel goede hyponiemparen zal kunnen opleveren, wordt het in het verdere proces wel gebruikt.

4.2.3 Aanpassing van de lexicaal-syntactische patronen

Om uit de zinsdelen die door de patronen geïdentificeerd zijn hyponiemrelaties te extraheren, moet bepaald worden welk deel van de NP aangeduid moet worden als hyperniem of hyponiem. Een voor de hand liggende optie, die ook gekozen is bij de eerste analyse van de lexicaal-syntactische patronen, is om alleen het hoofd-nomen van de NP te kiezen:

Sommige naaldbomen, zoals de spar of de den ...
hyponiem("spar","naaldboom")
hyponiem("den","naaldboom")

...schuchtere egels, nieuwsgierige eekhoorns en andere kleine bosdieren.
hyponiem("egel","bosdier")
hyponiem("eekhoorn","bosdier")

Ongewenste determinatoren, adjectieven en modificatoren (zoals aangehechte PP's) blijven hierdoor buiten beschouwing.
Daarnaast heeft selectie van de alleen het hoofd-nomen voordelen bij het opbouwen van de hiërarchie: het gaat om de opbouw van een algemene brede hiërarchie, en begrippen moeten eenduidig zijn en gemakkelijk aan elkaar gekoppeld kunnen worden. Modificaties zorgen voor een ongewenste complexiteit, die de koppeling van begrippen vaak in de weg staat.
De hyponiemparen

hyponiem("[den,spar]","inheemse naaldboom")
hyponiem("[spar,lariks]","bekende naaldboom")

kunnen niet worden gekoppeld, maar de paren

hyponiem("[den,spar]","naaldboom")
hyponiem("[spar,lariks]","naaldboom")

wel, en dit leidt tot:

hyponiem "([den,spar,lariks]","naaldboom")

Doordat de bestaande EuroWordNet-hiërarchie ook uit ongemodificeerde nomina bestaat, zal aanhechting van paren aan deze hiërarchie het best kunnen slagen bij gebruikmaking van ongemodificeerde nomina.

Uit de eerder beschreven resultaten van de steekproef bleek dat de keus voor alleen het hoofd-nomen vaak correcte resultaten opleverde. Nauwkeurige analyse van de geïdentificeede zinnen bracht echter naar voren dat enkele modificaties nodig zijn.

Ten eerste is gekozen om hoofd-nomina die gevolgd worden door een PP uit te sluiten. Uit de data bleek dat de PP in veel gevallen essentiële informatie

voor de hyponiemrelatie bevat, en dat zonder deze informatie de relatie niet tot stand kan worden gebracht[3]:

> *winnaars* [van circusprijzen] , zoals de Russische Borzovi Troupe en Gerd Seimoniet-Barum
> *middelen* [van buiten], zoals krachtvoer, kunstmest en bestrijdingsmiddelen
> *hulpmiddelen* [voor jeugdigen] zoals kinderduw-wandelwagens en speelvoertuigen
> wijnvlekken of andere *afwijkingen* [aan het gezicht]
> de imposante en strenge wolkenkrabbers of andere *iconen* [van de Nieuwe Wereld]

Ten tweede is uit de data gebleken dat in een aantal gevallen de hyponymierelatie pas tot stand kan worden gebracht of verbeterd kan worden, als het adjectief meegenomen wordt in de relatie.

> *weke* delen zoals kapsels en spieren
> *ontwrichte* schouders, *gebroken* benen en andere verwondingen
> Prednison en andere *pijnverlichtende* middelen

In de meeste gevallen is het echter zo, dat het adjectief geen toegevoegde waarde heeft en niet meegenomen moet worden. Omdat het onmogelijk is om met behulp van lexicaal-syntactische patronen te bepalen welke adjectieven wel en welke niet moeten worden meegenomen, worden de adjectieven buiten beschouwing gelaten. Hierop geldt echter één uitzondering: adjectieven die een eigennaam aanduiden (zoals een persoonsnaam of een plaatsnaam), worden wel meegenomen in de hyponiemrelatie. Dit is mogelijk omdat deze namen met een hoofdletter worden geschreven en daardoor gemakkelijk geïdentificeerd kunnen worden. Eigennamen geven in veel gevallen belangrijke extra informatie over de hyponiemrelatie:

[3]Het gaat hier uitsluitend om zinnen waarbij er binnen de NP een PP bestaat, en waarbij één van de elementen van de mogelijke hyponiemrelatie verwijst naar het hoofdnomen dat zich binnen de NP, maar buiten de PP bevindt. Als het hoofd-nomen onderdeel is van de PP, en een van de elementen van de hyponiemrelatie verwijst naar dat hoofdnomen, wordt de zin niet uitgesloten.
Voorbeeld van een correcte zin: In de moderne glastuinbouw draait alles om het telen [van tropische en subtropische *groenten*,] zoals *tomaat* en *komkommer*, die veel energie en hoge investeringen vragen .

Europese landen zoals Duitsland en Groot-Brittannië
Russische begrippen, zoals glasnost en perestrojka ...
Wigersma, Van Lunteren en andere *Hegeliaanse* filosofen ...

Om te zorgen dat alle nomina met adjectieven bij het bouwen van de hiërarchie gekoppeld kunnen worden aan hetzelfde nomen zonder adjectief, is uit elke adjectief-nomen combinatie de volgende hyponiemrelatie afgeleid, hetgeen 708 nieuwe relaties opleverde:

hyponiem(adjectief X nomen Y, nomen Y)

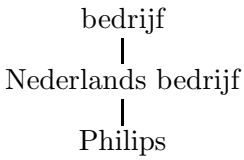
Voorbeeld:

hyponiem("Nederlands bedrijf","bedrijf")

Dit nieuwe hyponiempaar kan zorgen voor de juiste hiërarchie tussen de volgende paren:

hyponiem("Philips", "bedrijf")
hyponiem("Philips", "Nederlands bedrijf")

Combinatie van de drie hyponiemparen leidt tot:



Tenslotte zijn alle hyponiemparen waarbij hyponiem en hyperniem hetzelfde zijn verwijderd.

Het doorvoeren van de bovengenoemde wijzigingen levert de definitieve lexicaal-syntactische patronen op. Deze patronen zijn toegepast op alle beschikbare corpora, en leveren in totaal 17494 unieke hyponiemparen op. De hyponiemrelaties worden geëvalueerd in hoofdstuk 5.

4.2.4 Stemmen van de geïdentificeerde hyponiemparen

De bestaande EuroWordNet-hiërarchie bevat hyperniemen en hyponiemen in het enkelvoud. Omdat de verkregen hyponiemparen uit vrije tekst afkomstig zijn, bestaan deze uit zowel enkelvoudige als meervoudige woorden.

Om in een later stadium de bestaande hiërarchie te kunnen uitbreiden met de verkregen hyponiemparen, is het belangrijk om alle woorden in dezelfde vorm op te slaan. Vandaar dat is gekozen om alle verkregen paren met een stemmer terug te brengen naar enkelvoud.

Zoals besproken in paragraaf 3.3 geeft Alpino niet altijd de gewenste resultaten m.b.t. stemmen. Voor samenstellingen wordt slechts de stam van het laatste woord als *root* gegeven en bij onbekende woorden wordt het hele woord als *root* genomen. Deze *roots* beschouwen we als foutieve stamvormen[4]

In dit onderzoek is gekozen om een externe stemmer (Gaustad & Bouma 2002) te gebruiken die, net als Alpino, in eerste instantie de stam van het woord in een woordenboek opzoekt. Als het woord echter niet in het woordenboek staat, wordt als backup de Nederlandse versie van het op regels gebaseerde Porter-algoritme (Kraaij & Pohlman 1994) geraadpleegd. Hoewel deze stemmer niet foutloos stemt, wordt het grootste deel van de onbekende woorden correct gestemd. Dit in tegenstelling tot de Alpino-stemmer, waarbij we de stam van alle onbekende woorden (inclusief samenstellingen) als incorrect beschouwen.

Als input voor de externe stemmer dienen alle nomina die geen hoofdletter bevatten. Hierdoor wordt voorkomen dat eigennamen ten onrechte worden gestemd:

Philips ⇒ Philips en niet : Philips ⇒ Philip

De adjectieven die voorkomen in de paren worden, hoewel het om eigennamen gaat, wel gestemd, zodat alle adjectieven in dezelfde vorm opgeslagen worden. Het stemmen van adjectieven is eenvoudig en gaat vrijwel foutloos.

Enig nadeel is dat de congruentie tussen adjectief en nomen verloren kan gaan:

Nederlandse mannen ⇒ Nederlands man
Franse vrienden ⇒ Frans vriend

Dit probleem zou echter ook bestaan als het nomen wel en het adjectief niet werden gestemd:

[4]In enkele gevallen levert de foutieve *root* van een hyperniem dat een samenstelling is echter alsnog een goede hyponiemrelatie op:
Ongestemde hyponiemrelatie: hyponiem("honden","huisdieren")
Met Alpino gestemde hyponiemrelatie: hyponiem("hond","dier")

Europese landen ⇒ Europese land
Amsterdamse grachtenpanden ⇒ Amsterdamse grachtenpand

Het verlies van de congruentie levert voor het bouwen van de hiërarchie echter geen problemen op, omdat alle adjectieven wel consequent in dezelfde vorm staan, en eventuele koppeling dus mogelijk blijft[5].

Om te kijken in hoeverre het externe stemalgoritme correct werkt, is de stemmer getest op de verkregen hyponiemparen uit het Volkskrantcorpus.

Na stemming blijkt dat van de 3290 gestemde woorden er 2536 gestemd zijn m.b.v. het woordenboek. Van deze 2536 zijn er willekeurig 2 keer 100 gekozen en beoordeeld op correcte stemming. Uit de resultaten blijkt dat slechts 1% van de woorden incorrect is gestemd. Het gedeelte van de stemmer dat werkt m.b.v. het woordenboek hoeft daarom niet aangepast te worden.

Het backup-algoritme van de stemmer, de Nederlandse versie van het Porter-algoritme, kent regels voor het stemmen van alle bestaande woordsoorten. De te stemmen hyponiemparen bestaan echter alleen uit nomina, waardoor stemregels voor andere woordsoorten overbodig worden. Het is echter niet mogelijk om deze regels eenvoudig uit het algoritme te verwijderen, omdat deze niet per woordsoort gerangschikt zijn.

Uit analyse van de verkregen hyponiemparen uit het Volkskrantcorpus blijkt dat van de 4016 gestemde woorden er 754 zijn gestemd zijn m.b.v. het Porter-algoritme. Van deze 752 zijn er willekeurig weer 2 keer 100 gekozen en beoordeeld op correcte stemming. Hieruit blijkt dat 18% van de woorden foutief gestemd is.

De meest opvallende fout gemaakt door de Porter-stemmer is het incorrect stemmen van woorden met 'ge':

gezichten ⇒ zicht
geluidsrecorders ⇒ luidsrecorder
legerofficiers ⇒ lerofficier

Hoewel het zoals gezegd onmogelijk is om alle overbodige regels te verwijderen of aan te passen, zodat de stemmer geschikt is voor alleen nomina, was het wel mogelijk om de regel voor het stemmen van het pre- of infix

[5]Deze congruentie lijkt echter met enkele grammaticaregels eenvoudig herstelbaar, maar dit valt buiten het bereik van dit onderzoek.

'ge' te verwijderen. Het verwijderen van deze stemregel zorgt ervoor dat het foutenpercentage van de Porter-stemmer is verlaagd naar 10%.

Daarnaast is de regel voor koppeltekens aangepast: in het Porter-algoritme wordt deze verwijderd, maar in dit onderzoek worden koppeltekens in de stam, conform de Nederlandse spelling, behouden.

Als de resultaten uit de steekproeven samen worden genomen, kan worden geconcludeerd dat bij 3,1% van de gestemde woorden uit het Volkskrantcorpus de stemming incorrect lijkt te zijn.

4.3 Extractie uit een encyclopedie

Voor de eerder beschreven extractie van hyponiemparen uit vrije tekst met behulp van lexicaal-syntactische patronen wordt o.a. het Spectrum Enclyclopedie corpus gebruikt. De Spectrum Encyclopedie biedt echter nog een andere mogelijkheid om hyponiemparen te identificeren.

De Spectrum Encyclopedie bestaat uit enkele tienduizenden begrippen, ingangen genoemd, die één voor één beschreven en verklaard worden. Elke beschrijving bestaat uit één of meerdere zinnen. Opvallend is dat elke eerste zin van de verklaring in een overgroot deel van de gevallen beschrijft tot welke supercategorie een begrip gerekend moet worden.

Dit zijn enkele voorbeelden van ingangen met de eerste zin van de beschrijving:

> 2 Unlimited – Nederlands/Belgische house-act.
> 2000-syndroom – Computerprobleem dat gepaard gaat met de overgang naar een nieuw millennium.
> Aa en Hunze – Gemeente in de Nederlandse provincie Drenthe; 278,86 km2, 25.600 inw. (2002).
> Aachen, Hans von – Duits schilder van portretten, genrestukken, religieuze, allegorische en mythologische taferelen.
> aal – Europese palingen ('Anguilla anguilla'), geslacht palingen.

Uit deze zinnen kunnen de volgende hyponiemrelaties worden afgeleid:

> hyponiem("2 Unlimited"," Nederlands/Belgische house-act")
> hyponiem("2000-syndroom","computerprobleem")
> hyponiem("Aa en Hunze","gemeente")
> hyponiem("Aachen, Hans von","Duits schilder")
> hyponiem("aal","Europese paling")

In dit onderzoek is geprobeerd deze hyponiemrelaties automatisch te extraheren. Het hyponiem wordt altijd gevormd door de ingang, en het hyperniem wordt verkregen met behulp van een eenvoudige syntactisch extractieregel. Als basis voor deze regel dient de aanname het hyperniem wordt gevormd door het hoofd-nomen van eerste NP in de beschrijving[6].

Voor de extractie van de hyperniemen wordt het hoofd-nomen van de eerste NP in de eerste zin gezocht. Vaak is het echter zo dat de eerste zin geen NP bevat, omdat die zin een jaartal, pseudoniem of afkorting bevat. In dat geval wordt het hoofd-nomen van de eerste NP in de tweede zin van de beschrijving gezocht.

4.3.1 Analyse van de opgestelde syntactische extractieregel

De identificatie van het hoofd-nomen van de eerste NP leverde bij 53.587 van de 62.103 ingangen een resultaat op. Uit alle resultaten zijn door middel van een steekproef 100 hyperniemen en bijbehorende hyponiemen geselecteerd. Deze zijn beoordeeld door de auteur op correctheid van de relatie, zodat een beeld kan worden gevormd van de te verwachten resultaten. Hieruit blijkt dat in 72% van de gevallen een correcte hyponiemrelatie is geïdentificeerd.

4.3.2 Aanpassing van de extractieregel

De 100 steekproefparen zijn vervolgens gebruikt om te bekijken of er een relatie bestaat tussen de positie in de zin waarop de NP (die het hoofdonomen bevat) begint en de juistheid van het hyperniem. Tabel 4.2 geeft hiervan een overzicht:

Uit de resultaten in tabel 4.2 blijkt dat 88% van de NP's direct aan het begin van de zin staat. Het hoofd-nomen van deze NP's duidt in 80% van de gevallen het correcte hyperniem aan. 12% van de NP's begint niet direct aan het begin van de zin. Van deze 12% levert 83% het foutieve hyperniem op.
Het opstellen van de voorwaarde dat de eerste NP met hoofd-nomen aan het begin van de zin moet staan, levert 12% minder data op, maar verbetert het aantal correcte paren van 72% naar zo'n 80%. Gebaseerd op deze bevindingen is besloten de voorwaarde op te nemen in de extractieregel.

[6]De keus voor de selectie van slechts het hoofd-nomen is besproken in paragraaf 4.2.3

Positie:	**Hyperniem goed:**	**Hyperniem fout:**	**Totaal:**
0	70	18	88
1	0	1	2
2	1	3	4
3	0	1	1
4	1	2	3
5	0	2	2
≥ 6	0	1	1

Tabel 4.2: Overzicht van positie in de zin waarop de NP die het hyperniem bevat begint. Als de positie 0 is, betekent dit dat de NP aan het begin van de zin staat, als de positie 1 is, gaat er een woord aan de NP vooraf, enz.

Daarnaast bleek uit studie van de hyponiemparen met bijbehorende ingangen en beschrijvingen, dat een aantal woorden door Alpino vaak als hoofdnomen van de NP geannoteerd wordt, maar dit semantisch niet zijn. Het gaat hierbij om de volgende, in de Encyclopedie veel voorkomende, woorden:

> *deel (of delen)*:
> adamsappel – deel
> Het voelbare , naar voren gerichte deel van het schildkraakbeen (het grootste kraakbeenstuk van het strottenhoofd) .
>
> *aantal*:
> geslachtsziekten – aantal
> Een aantal ziekten te zamen , die door intiem lichamelijk contact worden overgebracht , dus vooral door geslachtsgemeenschap .
>
> *geheel*:
> apoplast - geheel
> Het geheel van samenhangende celwanden in de plant waardoor watertransport kan plaatsvinden.
>
> *soort (of soorten)*:
> welstaal - soort
> Verouderd soort koolstofstaal verkregen door in gesmolten ruwijzer het koolstofgehalte terug te brengen door toevoeging van ijzererts .

vorm (of vormen):
stekken - vorm
Vorm van ongeslachtelijke (vegetatieve) voortplanting bij planten .

(verzamel)naam:
structurele linguïstiek - verzamelnaam
Algemene en vandaar vrij vage verzamelnaam voor diverse wijzen van taalbestudering die teruggaan op de ideeën van de Zwitser Ferdinand de Saussure (1857-1913) .

type/typen/types:
neerquark - typen
Een van de zes typen quarks , de elementaire deeltjes waaruit o.a. protonen en neutronen bestaan.

Hyponiemparen waarbij een van deze woorden is geïdentificeerd als hyperniem zijn uitgesloten.

Ook is, net als bij de extractie van hyponiemparen uit vrije tekst, besloten om adjectieven die een eigennaam aanduiden en geïdentificeerd kunnen worden doordat ze met een hoofdletter beginnen, mee te nemen in het hyperniem.
Net als bij de extractie uit vrij tekst, zijn hier uit alle adjectief-nomen combinaties weer nieuwe relaties afgeleid van type:

hyponiem(Adj N, N)

Dit leverde 4978 nieuwe hyponiemparen op.

De bij de extractie uit vrije tekst gestelde voorwaarde, dat de NP niet gevolgd mag worden door een PP, wordt hier echter niet gesteld. Uit analyse van de steekproefresultaten is gebleken dat de beschrijvingen in de encyclopedie zodanig zijn geformuleerd, dat een op de NP volgende PP in het overgrote deel van de gevallen geen invloed uitoefent op de hyponiemrelatie.

Tenslotte zijn alle hyponiemparen waarbij hyponiem en hyperniem hetzelfde zijn verwijderd.

Het doorvoeren van de bovengenoemde wijzigingen levert de definitieve extractieregel op. Toepassing van deze regel op de Spectrum Enclyclopedie leverde 52063 hyperniemen en bijbehorende hyponiemen op. De evaluatie van deze hyponiemrelaties wordt besproken in hoofdstuk 5.

4.3.3 Stemmen van de geïdentificeerde hyponiemparen

De ingangen van de Spectrum Encyclopedie staan grotendeels in het enkelvoud. Dit geldt ook voor de geëxtraheerd hoofd van de eerste NP. Voor de meeste hyponiemparen zal stemming dus niet nodig zijn. Uit een steekproef van 100 paren bleek dat slechts 3% van de paren een begrip in het meervoud bevat. Aangezien dit foutenpercentage lager ligt dan het foutenpercentage van stemming, is besloten de uit de Spectrum Encyclopedie geëxtraheerde hyponiemparen niet te stemmen. Voordeel hiervan is dat stemfouten worden voorkomen en congruentie tussen adjectief en nomen wordt behouden[7].

[7]Aangezien de adjectieven van de uit vrije tekst geëxtraheerde hyponiemparen wel gestemd zijn, zullen deze mogelijk niet gekoppeld kunnen worden aan de ongestemde adjectieven uit de encyclopedie. Oplossing hiervoor is om deze ongestemde adjectieven bij het bouwen van de hiërarchie alsnog te stemmen en te gebruiken voor koppeling, terwijl de ongestemde vorm als weergavevorm gebruikt zal blijven worden.

Hoofdstuk 5

Resultaten en evaluatie

De extractie van hyponiemrelaties uit alle beschikbare corpora met vrije tekst leverde in totaal 17.494 gestemde unieke paren op en de extractie uit de Spectrum Encyclopedie leverde 52.063 ongestemde unieke paren op.

In dit hoofdstuk worden beide groepen geëxtraheerde paren geëvalueerd. Deze evaluatie is echter niet eenvoudig, omdat er geen Goldstandard-data (set van correcte data waarmee vergeleken kan worden) beschikbaar zijn. Daarom zijn twee alternatieve evaluatiemethodes bedacht:

1. van elk van de twee groepen geëxtraheerde paren wordt een steekproef genomen, en deze wordt beoordeeld door menselijke beoordelaars.

2. elk van de twee groepen geëxtraheerde paren wordt vergeleken met de bestaande hyponiemrelaties in EuroWordNet (Vossen 1998).

De eerste evaluatiemethode levert informatie op over de precisie van de data. Deze informatie wordt gebruikt voor vergelijking met resultaten van eerder vergelijkbaar onderzoek. Daarnaast wordt door nauwkeurige bestudering van de steekproef bepaald wat de oorzaak is van de foutieve hyponiemparen. De tweede evaluatiemethode laat zien hoe groot de overlap is tussen EuroWordNet en de geëxtraheerde paren. Hierdoor kan bepaald worden in hoeverre de geëxtraheerde hyponiemparen een aanvulling kunnen zijn op de bestaande EuroWordNet-ontologie .

Nadat eerst de resultaten van eerder relevant onderzoek kort worden gesproken, worden beide extractiemethodes m.b.v. de twee evaluatiemethodes zorgvuldig geanalyseerd.

5.1 Resultaten van eerder onderzoek

Eerder onderzoek naar de automatische extractie van hyponiemrelaties laat, hoewel de gebruikte methodes veel op elkaar lijken, een groot verschil in resultaten zien.

Hearst (1992) beschrijft 'een hoge kwaliteit' hyponiemrelaties geëxtraheerd uit vrije tekst in een encyclopedie m.b.v. het *such as*-patroon. Hierbij worden hoofd-nomina voorkomend in complexe NP's uitgesloten, waardoor problemen met bijvoorbeeld PP-aanhechting worden voorkomen. Daarnaast wordt bij de beoordeling van de relaties niet gekeken naar enkelvoud en meervoud.
Als belangrijke problemen beschrijft Hearst metonymia, onderspecificatie en contextgevoeligheid:

Metonymia:
hyponym("king","institution")
= hyponiem("koning","instituut")

Onderspecificatie:
hyponym("steatornis","species")
= hyponiem("olievogel","soort")

Contextgevoeligheid:
hyponym("aircraft","target")
= hyponiem("vliegtuig","doel")

Hearst (1998) evalueert het *or other*-patroon door het toe te passen op zes maanden tekst uit de New York Times, waarbij wederom alleen hoofd-nomina in eenvoudige NP's geselecteerd worden. Handmatige beoordeling van 200 hyponiemparen leverde een precisie van 63% op. Hearst geeft aan dat het soort corpus van invloed kan zijn op de resultaten: krantentekst kent vaak meer invloed door de context dan encyclopedietekst. Ook bevat krantentekst vaak meer subjectieve oordelen, meningen, en metaforisch taalgebruik, waardoor de resultaten minder goed kunnen zijn.

Morin en Jacquemin (2002) evalueren voor het Frans handmatig 884 paren, geëxtraheerd m.b.v. elf verschillende lexicaal-syntactische patronen, en halen een precisie van 79%. Zij geven echter niet aan welke criteria zij hebben gehanteerd bij selectie van hyperniem en hyponiem. Uit de gegeven

voorbeelden wordt echter wel duidelijk dat modificeerders, zoals aanwezige adjectieven en aangehechte PP's, meegenomen worden in de paren. Hierdoor wordt de hyponiemrelatie minder algemeen en minder geschikt voor automatische uitbreiding van een ontologie, maar is het foutenpercentage lager doordat hyperniemen en hyponiemen completer zijn. Een voorbeeld:

> Les *caractéristiques du site* telles la *pente*, le *sous-bois* et la *distance des usines* ...
> (De karakteristieken van het terrein, zoals de helling, het kreupelhout en de afstand tussen de fabrieken....)
>
> hyponiem("pente","caractéristiques du site")
> hyponiem("sous-bois","caractéristiques du site")
> hyponiem("distance des usines","caractéristiques du site")

Rydin (2002) stelt voor het Zweeds vijf lexicaal-syntactische patronen op, waarmee ze hyponiemrelaties identificeert in artikelen van een Zweeds gelemmatiseerd krantencorpus. Bij deze hyponiemrelaties extraheert ze steeds het hoofd-nomen van de NP als hyponiem en hyperniem. Na handmatige evaluatie van 20% van de hyponiemparen, blijkt 92% correct te zijn. Als problemen noemt ze vooral incorrecte woordsoorttoekenning en verandering van betekenis door PP-aanhechting. Daarnaast geeft ze aan dat er door stemming of lemmatisatie morfologische problemen kunnen ontstaan. Zo is de volgende hyponiemrelatie correct

> hyponiem("boys","group")

maar is deze na stemming/lemmatisatie niet meer correct:

> hyponiem("boy","group")

Hoewel Rydin bijzonder goede resultaten behaalt, verklaart ze niet waarom haar resultaten opvallend beter zijn dan die van bijvoorbeeld Hearst (1998).

Cederberg en Widdows (2003) gebruiken het British National Corpus (BNC 1994) voor de extractie van hyponiemparen m.b.v. de door Hearst opgesteld patronen. Bij evaluatie van de paren beoordelen zij de hele NP, en delen deze in in vijf categorieën:

4. De relatie tussen hyperniem en hyponiem is volledig correct
3. De relatie tussen hyperniem en hyponiem is correct na kleine modificatie, zoals het verwijderen van een lidwoord en andere voorafgaande woorden, of het veranderen van meervoud in enkelvoud.
2. De relatie tussen hyperniem en hyponiem is potentieel correct, maar (mogelijk computationeel lastige) wijzigingen zijn nodig.
1. De relatie tussen hyperniem en hyponiem is te algemeen of te context-specifiek.
0. De relatie tussen hyperniem en hyponiem is incorrect.

Cederberg en Widdows beschouwen alle paren in categorie 3 of 4 correct. Handmatige evaluatie van 100 willekeurig geselecteerde paren gaf een precisie van 40%. Als het voornaamste probleem noemen zij dat constructies waarvan aangenomen wordt dat ze hyponiemrelaties aanduiden, vaak voor andere doelen gebruikt worden. Daarnaast constateren zij veel problemen met contextgevoeligheid.

5.2 Evaluatie door middel van beoordeling van een steekproef

5.2.1 Methode

Om de precisie van de ontwikkelde extractiemethodes te bepalen, wordt uit beide groepen geëxtraheerde paren een steekproef van 200 paren genomen. Deze 200 paren worden vervolgens door 3 mensen[1] beoordeeld.

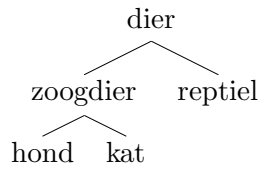
De paren worden beoordeeld als *goed* of *fout*. Hierbij gelden de volgende richtlijnen:

Paren zijn goed als:

- zij in een directe of indirecte hyponiemrelatie tot elkaar staan.

[1] Een studente Nederlands, een studente Kunstgeschiedenis en een website-beheerder met HBO-niveau

Een voorbeeld:



Directe relatie: hyponiem("reptiel","dier")
Indirecte relatie: hyponiem("hond","dier")[2]

- zij goed gespeld en gestemd zijn.
- zij beide in enkelvoud staan.

Alle andere paren zijn *fout*.

Uit de beoordeling van de hyponiemparen worden vervolgens de volgende statistische waardes afgeleid, gebaseerd op Rydin (2002):

Gemiddelde: Gemiddeld percentage paren dat door de beoordelaars als correct wordt beoordeeld

Tenminste één: Percentage paren dat door tenminste één van de beoordelaars als correct wordt beoordeeld.

Meerderheid: Percentage paren dat door tenminste twee van de beoordelaars als correct wordt beoordeeld.

Unaniem: Percentage paren dat door alle beoordelaars als correct wordt gezien.

Daarnaast wordt de *kappa waarde* (Carletta, 1996) gebruikt om te meten hoe groot de overeenstemming tussen de beoordelaars is. Uit deze waarde kan de moeilijkheidsgraad van de beoordelingstaak worden geschat.

De *kappa waarde* (K) meet per hyponiempaar de overeenstemming tussen de beoordelaars, en corrigeert deze voor de verwachte kans op overeenstemming:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

[2]Deze directe of indirecte relatie moet door de beoordelaars zelf worden ingeschat, zonder dat zij de beschikking hebben over dit soort boomstructuren.

Waarbij P(A) aangeeft in welk deel van de gevallen de beoordelaars het eens zijn, en P(E) aangeeft in welk deel van de gevallen te verwachten is dat de beoordelaars het bij toeval eens zijn. Als er geen overeenstemming is tussen de beoordelaars is de *kappa waarde* nul, en bij totale overeenstemming is deze één.

5.2.2 Resultaten

Voor de evaluatie van zowel de extractie van hyponiemparen uit vrije tekst als de extractie uit de Spectrum Encyclopedie is een steekproef (n=200) genomen uit het totaal aantal hyponiemparen. De resultaten van de beoordeling door de drie beoordelaars zijn te vinden in tabel 5.1 en 5.2:

Statistische maat:	**Percentage:**
Gemiddelde:	46,8%
Tenminste één:	60,5%
Meerderheid:	47%
Unaniem:	33%
Kappa waarde:	0,69

Tabel 5.1: Resultaten van de beoordeling van de steekproef ($n = 200$) van hyponiemparen geëxtraheerd uit vrije tekst

Statistische maat:	**Percentage:**
Gemiddelde:	84,8%
Tenminste één:	90%
Meerderheid:	87%
Unaniem:	78,5%
Kappa waarde:	0,70

Tabel 5.2: Resultaten van de beoordeling van de steekproef (n=200) van hyponiemparen geëxtraheerd uit de Spectrum Encyclopedie

De *kappa waardes* van 0.69 en 0.70 geven aan dat de overeenstemming tussen de beoordelaars vrij hoog is, maar dat de beoordelaars het toch bij een aanzienlijk aantal paren niet eens waren. De beoordelaars gaven aan dat de beoordeling niet voor alle paren eenvoudig was, en dat er meer gradaties van correctheid bestaan dan *goed* en *fout*. Doordat zij bij deze taken echter toch voor *goed* of *fout* moesten kiezen, zijn er onderlinge verschillen ontstaan.

Opvallend is het grote verschil tussen de resultaten van de extractie uit vrije tekst en de extractie uit de Spectrum Encyclopedie. Ondanks de eenvoud van de methode levert de extractie uit de Spectrum Encyclopedie opvallend goede resultaten op. De veel complexere methode van extractie uit vrije tekst laat minder goede resultaten zien. In de volgende paragraaf worden de oorzaken voor de foutieve hyponiemparen van beide methodes nader bekeken.

5.2.3 Analyse van de foutieve hyponiemparen

Extractie uit vrije tekst

Bij nadere beschouwing van de hyponiemparen in de steekproef, valt meteen op dat er opvallend veel paren zijn waarbij stemfouten de oorzaak zijn voor het afkeuren van de hyponiemrelatie: bij 23 paren is de hyponiemrelatie incorrect door een foutief gestemd hyponiem of hyperniem. Deze stemfouten zorgen voor een daling van de precisie van zo'n 11,5%[3]. Enkele voorbeelden:

> hyponiem("kind","inwonend")
> De indieners van het wetsvoorstel willen niet dat eventueel inkomen van *kinderen* en andere *inwonenden* meetelt voor het bepalen van de hoogte van de subsidie .
>
> hyponiem("kaneel","produk")
> Niet uit de hoofdplaats Colombo, maar uit Galle, dat over een betere rede beschikte, werden *kaneel* en andere *produkten* naar de Republiek verscheept .
>
> hyponiem("liefde","gevoelen")
> Uit vier korte interviews met jongeren blijkt dat ze niet met hun ouders maar met vrienden praten over *liefde*, seks en andere indrukwekkende *gevoelens*.
>
> hyponiem("Geissler","CDU'ers")
> In de jaren '80 beraamden *Geissler* en enkele andere *CDU'ers* een coup tegen Kohl, die zij een lichtgewicht vonden en met wiens politiek zij het niet eens waren.

[3]Bij de tussentijdse korte evaluatie van de hyponiemparen waren deze nog niet gestemd en is niet gekeken naar getal, waardoor de tussentijdse resultaten opvallend beter lijken dan de resultaten van deze steekproef.

Zoals besproken in paragraaf 4.2.4, wordt gemiddeld 3,1% van de woorden fout gestemd. Voor de steekproef, die in totaal uit 452 woorden bestaat, geeft dit een verwachting van 14 foutief gestemde woorden.
Dat er maarliefst 23 paren zijn die minstens één foutief gestemd woord bevatten, is te verklaren door het feit dat er opvallend veel woorden in de steekproef zitten die niet in het woordenboek van de stemmer voorkomen. Deze woorden worden vervolgens gestemd m.b.v. het Porter-algoritme, dat een veel hoger foutenpercentage kent dan de woordenboekstemmer. Hierdoor komt het percentage foutief gestemde woorden hoger te liggen.

Een bijkomend probleem dat ontstaat door het stemmen van de hyponiemparen is dat sommige relaties die in het meervoud wel kloppen, in het enkelvoud niet geldig zijn:

> hyponiem("vrouwenorganisatie","mensenrechtenactivist")
> Ik vraag mij af waarom *vrouwenorganisaties* (ook migrantenvrouwen), politici, schrijvers en andere *mensenrechtenactivisten* weinig doen om het leven van Nasrin te redden .

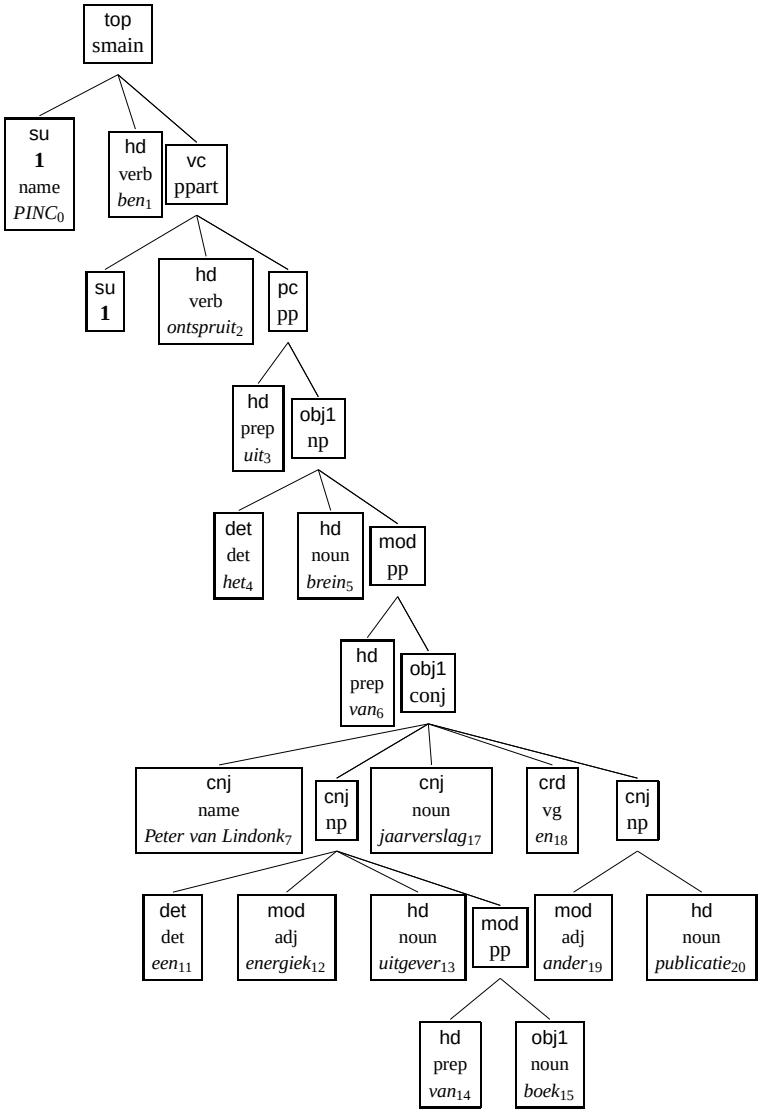
> hyponiem("computer","spul")
> Veiligheidsfunctionarissen van de VN zeiden dat civiel en militair personeel is betrapt op het stelen van magnetrons, koelkasten, waterpompen, *computers* en andere waardevolle *spullen*.

Naast problemen ontstaan door stemming, zijn er ook foutieve hyponiemparen veroorzaakt door fouten van de Alpino-ontleder. Een voorbeeld:

> hyponiem("Peter van Lindonk","publicatie")

> zin: PINC is ontsproten uit het brein van *Peter van Lindonk*, een energieke uitgever van gelegenheidsboeken, jaarverslagen en andere *publicaties*.

Deze zin is terug te vinden in figuur 5.1. Uit deze boomstructuur blijkt dat Alpino *Peter van Lindonk*, *jaarverslagen* en *andere publicaties* coördineert, waardoor *Peter van Lindonk* geïdentificeerd wordt als hyponiem van *publicaties*.



Figuur 5.1: Boomstructuur voor de zin 'PINC is ontsproten uit het brein van Peter van Lindonk, een energieke uitgever van gelegenheidsboeken, jaarverslagen en andere publicaties'.

Een derde duidelijke oorzaak voor foutieve hyponiemparen is contextgevoeligheid: een aantal hyponiemparen zijn niet correct omdat de zin waarin ze voorkomen of enkele voorafgaande zinnen nodig zijn voor juiste interpretatie:

> hyponiem("smaak","eigenschap")
> Hulpstoffen die aan voedingsmiddelen worden toegevoegd, om de geur, *smaak*, houdbaarheid of andere *eigenschappen* te verbeteren.
>
> hyponiem("Charles","landeigenaar")
> *Charles* en vier andere *landeigenaren* willen de indringers met hun vijftig karren, woonwagens en auto's weg hebben.
>
> hyponiem("2 Vandaag","samenwerkingsverband")
> Ondanks de op het eerste oog 'onmogelijke combinatie' is volgens de EO op Nederland 2 wonderlijk veel tot stand gekomen: *2 Vandaag*, een kinderblok en andere *samenwerkingsverbanden.*
>
> hyponiem("Lotus","kandidaat") Rietveld sprak in het diepste geheim met verschillende *kandidaten*, waaronder *Lotus*, en kwam ten slotte bij Novell terecht.
>
> hyponiem("petitie","initiatief")
> Met blokkades, *petities* en andere *initiatieven* proberen ze onder het motto 'Zonder auto natuurlijk' vooral gemeentebesturen te bewegen maatregelen te nemen.
>
> hyponiem("boeking","toepassing")
> De initiatiefnemers denken aan verschillende *toepassingen* zoals telewerken, *boekingen* of reserveringen en het raadplegen van informatiebanken .

De hierboven besproken fouten, ontstaan door foutieve stemming, foutieve ontleding of contextgevoeligheid, zijn niet of nauwelijks beïnvloedbaar door aanpassing van de lexicaal-syntactische patronen. Er zijn echter ook fouten die ontstaan door keuzes die gemaakt zijn bij het opstellen van de lexicaal-syntactische patronen.

Ten eerste zijn er problemen met PP-aanhechting. In het volgende voorbeeld

> *Winnaars* van circusprijzen, zoals de Russische *Borzovi Troupe* en *Gerd Seimoniet-Barum*...

zijn de nomina die in coördinatie na *zoals* voorkomen voorbeelden van *winnaars*. In het volgende voorbeeld, waarbij slechts de nomina in coördinatie zijn vervangen door andere nomina, is dit echter niet het geval:

> Winnaars van *circusprijzen*, zoals de *Nationale Circusprijs* en de *Herman Renz-Award*...

In dit voorbeeld zijn de nomina die in coördinatie na *zoals* voorkomen voorbeelden van *circusprijzen*. Voor een ontleder, zoals Alpino, die geen kennis heeft van de semantiek van de verschillende woorden in de zin, is het lastig om de juiste syntactische structuur toe te kennen aan de zinnen in de zojuist gegeven voorbeelden. Dit heeft tot gevolg dat de ontleding van NP's met aangehechte PP's niet altijd correct gebeurt.

Bij het opstellen van de lexicaal-syntactische patronen is besloten om alle hoofd-nomina, die gevolgd worden door een PP, uit te sluiten. Hoofd-nomina die voorkomen als onderdeel van een PP zijn echter wel meegenomen. Uit analyse van de steekproef blijkt dat dit, als gevolg van onjuiste ontleding, veel foutieve hyponiemparen oplevert:

> hyponiem("recht","douanemaatregel")
> De staten van een douane-unie vormen één grondgebied, voor zover het de heffing van deze *rechten* en andere *douanemaatregelen* betreft.

> hyponiem("première","Amerikaan")
> Maar liefst 22 werken van de *Amerikaan*, waaronder enkele *premières*, worden die dag uitgevoerd.

> hyponiem("Giovanni Fidanza","vertrek")
> Ook de Italiaanse ploegen Brescialat en Polti verschijnen met hun sterkste renners aan het *vertrek* zoals Eric Vanderaerden en *Giovanni Fidanza*.

> hyponiem("HFC","historie")
> Het liefst spelen ze tegen clubs met een rijke *historie* zoals *HFC*.

> hyponiem("Denemarken","dynamiek")

In landen met meer *dynamiek* zoals Japan, de VS en *Denemarken* is de kans om langdurig werkloos te raken veel kleiner.

Ten tweede is bij het opstellen van de lexicaal-syntactische patronen gekozen om adjectieven uit te sluiten (m.u.v. adjectieven beginnend met een hoofdletter). Dit heeft tot gevolg dat er enkele foutieve of minder correcte hyponiemrelaties ontstaan door het ontbreken van het adjectief:

hyponiem("oud-stalinist","zool")
Dat zag je al met de oproep van *oud-stalinisten* en andere *halve zolen* die namens hun kinderen het soevereine grondrecht opeisen op het gebruik van crack en heroïne.

hyponiem("nota","stuk")
Het gaat om twee drie meter *nota's* en andere *departementale stukken*.

hyponiem("bezuiniging","maatstaf")
Bij het akkoord ligt de klemtoon op *bezuiniging*, het terugdringen van het financieringstekort en andere *financieel-economische maatstaven*.

Tenslotte ontstaan er foutieve hyponiemparen doordat "hoeveelheids-woorden" als hoofd zijn aangemerkt door de parser en dus als hyperniem of hyponiem worden geselecteerd:

hyponiem("assortiment","drank")
Televisietoestellen, radio's, video's, kleren, horloges, een groot *assortiment* buitenlandse bieren, whisky's en andere sterke *drank*.

hyponiem("tractor","hoeveelheid")
India heeft Suriname een grote *hoeveelheid* agrarische werktuigen geschonken, waaronder *tractoren*, graafmachines, pelmolens, dorsmachines en waterpompen.

hyponiem("doosje","presentje")
Bloemen, flessen drank, *doosjes* bonbons en andere *presentjes* worden hem aangereikt.

Uit bovenstaande foutenanalyse wordt duidelijk dat het verbeteren van de stemmer en het verbeteren van de ontleding door Alpino de precisie behoorlijk kan verhogen. Hoewel stemming en ontleding buiten het bereik van deze

scriptie liggen, zijn beide van groot belang voor het succes van de extractiemethode. Het is dus noodzakelijk om bij extractie van hyponiemrelaties m.b.v. lexicaal-syntactische patronen de beschikking te hebben over een goede ontleder en stemmer.

De fouten die ontstaan door contextgevoeligheid zijn niet eenvoudig op te lossen door aanpassing van de gebruikte lexicaal-syntactische patronen. In de meeste gevallen zijn deze hyponiemparen niet volledig fout, maar ontbreekt een context voor juiste interpretatie. Aangezien de syntaxis van correcte en contextgevoelige hyponiemparen hetzelfde is, is dit probleem niet op te lossen door het aanpassen van de patronen. De fouten die ontstaan door contextgevoeligheid zijn wellicht te verminderen door het gebruik van een minder contextgevoelig corpus: in dit onderzoek bestond 94,2% van de data uit krantentekst, en slechts 5,8% uit encyclopedietekst. Zoals aangegeven door Hearst (1998), is het waarschijnlijk dat encyclopedietekst minder contextgevoelig is. Encyclopedietekst is voor het Nederlands echter niet in grote aantallen beschikbaar, waardoor krantencorpora dus ook nodig zullen zijn voor het uitbreiden van een ontologie.

De fouten met de PP-aanhechting kunnen verbeterd worden door striktere selectie van hyponiemparen: naast de restrictie dat hoofd-nomina die gevolgd worden door een PP niet worden geselecteerd, kan als restrictie worden opgenomen dat ook hoofd-nomina die onderdeel zijn van een PP niet worden geselecteerd. Dit heeft echter wel tot gevolg dat ook een groot aantal correcte hyponiemparen niet meer geselecteerd wordt, waardoor de recall flink zal dalen.

Ook de fouten die ontstaan door uitsluiting van adjectieven kunnen eenvoudig worden voorkomen door het aanpassen van de lexicaal-syntactische patronen. Maar ook hier moet bekeken worden of dit het gewenste resultaat oplevert: door het toestaan van adjectieven in hyponiemparen zal het foutenpercentage omlaag gaan, maar verdwijnt wel het algemene karakter van de hyponiemparen die in een later stadium in een bestaande algemene ontologie geplaatst worden.

Het probleem met de hoeveelheidswoorden is wel eenvoudig op te lossen, omdat al deze fouten uit zinnen met dezelfde syntactische constructie komen. Door een extra regel toe te voegen die er voor zorgt bij alle hoofd-nomina die gevolgd worden door een modificerende NP het hoofd van die modificerende NP wordt geselecteerd i.p.v. het hoofd-nomen van de hoofd-NP, is het probleem in de meeste gevallen opgelost. In de nieuwste versie van Alpino worden hoeveelheidswoorden voorzien van een aparte markering

(AMOUNT), waardoor deze constructies eenvoudig te herkennen zijn.

Extractie uit de Spectrum Encyclopedie

Nauwkeurige analyse van de foutieve hyponiemparen in de steekproef laat zien dat de meeste foute paren (13 van de 20 paren die door alle beoordelaars als fout worden beoordeeld) ontstaan doordat het hyponiem of het hyperniem niet in het enkelvoud staat. Voorbeelden:

> hyponiem("chromosoomafwijkingen","afwijkingen")
> Aangeboren *afwijkingen* die ontstaan als het chromosomenmateriaal in de lichaamscellen van de foetus defect, overmatig aanwezig of onvolledig is.
>
> hyponiem("gladiolen","planten")
> *Planten* met grote bloemen in allerlei kleuren, vooral gekweekt als tuinplant en snijbloem in vele soorten en variëteiten.
>
> hyponiem("geldzuivering","maatregelen")
> *Maatregelen* om het teveel aan geld (overtollige koopkracht) weg te nemen.
>
> hyponiem("kelp","wiervegetaties")
> Op rotsen groeiende *wiervegetaties*, vooral van grote soorten bruinwieren; vormen belangrijke broed- en schuilplaats voor talrijke dieren.

Er is bewust gekozen om de hyponiemparen niet te stemmen, omdat verondersteld wordt dat dit minder foutieve paren oplevert dan wel stemmen. Het foutenpercentage van 6,5% is aanzienlijk lager dan de 11,5% stemfouten die bij de steekproef van hyponiemparen uit vrije tekst werd verkregen.

Bij de overige fouten is de oorzaak in de meeste gevallen een onderspecificatie: het hyperniem is niet specifiek genoeg. Enkele voorbeelden:

> hyponiem("sterrenstelsel","verzameling")
> Grote *verzameling* van sterren, die door de onderlinge zwaartekracht bij elkaar worden gehouden.
>
> hyponiem("subcultuur","complex")

Complex van heel eigen normen en waarden, naast of zelfs tegenovergesteld aan de heersende cultuur (tegencultuur).

hyponiem("Gurkha","benaming")
Benaming voor de Nepalese soldaten in het Brits-Indische leger, die met name faam verwierven in de Sepoy-opstand (1857-1858).

Deze hyperniemen zouden uitgesloten kunnen worden door een restrictie die alle hoofd-nomina gevolgd door een PP uitsluit. In de Spectrum Encyclopedie wordt het eerste hoofd-nomen echter in veel gevallen gevolgd door een PP, waardoor ook veel goede relaties zouden worden uitgesloten. Dit leidt tot een drastische verlaging van de recall. Omdat overspecificatie slechts in zeer weinig gevallen voorkomt, is dit geen verstandige restrictie.

5.2.4 Vergelijking met de resultaten van eerder onderzoek

De in dit onderzoek genoemde oorzaken voor foutieve hyponiemparen bij extractie uit vrije tekst, zoals problemen met PP-aanhechting, contextgevoeligheid en morfologische problemen worden ook in de eerdere onderzoeken genoemd.
Ook het probleem met onderspecificatie, dat naar voren komt bij de evaluatie van de foutieve hyponiemparen bij extractie uit de Spectrum Encyclopedie, wordt in de eerdere onderzoeken genoemd. Onderspecificatie speelt bij de extractie uit vrije tekst geen of een kleine rol, omdat gekozen is om alle hoofd-nomina van NP's die gevolgd worden door een PP uit te sluiten en hiermee de voornaamste oorzaak voor onderspecificatie is geëlimineerd. Bij de extractie uit de Spectrum Encyclopedie is deze restrictie echter niet doorgevoerd.

De in de literatuur genoemde problemen met metonymia of andersoortige relaties dan hyponiemrelaties komen uit de analyse van de steekproef niet naar voren. Het is echter onwaarschijnlijk dat deze relaties totaal niet voorkomen, omdat sommige lexicaal-syntactische patronen deze relaties wel kunnen selecteren:

hyponiem("neus","gezicht") ⇒ metonymia
hyponiem("mond","gezicht") ⇒ metonymia
Enkele delen van het *gezicht*, zoals de *neus* en de *mond* ...

hyponiem("dochter","zoon")
Hij belde zijn oudste *dochter* en drie andere *zonen* ...

Uit nauwkeurige vergelijking van de besproken evaluatiemethodes blijkt dat er verschillende evaluatiecriteria worden gebruikt, waardoor het vergelijken van de resultaten niet eenvoudig is. Zo evalueren Hearst (1998), Morin en Jacquemin (2002) en Cederberg en Widdows (2003) ongestemde hyponiemparen en wordt er niet naar enkelvoud of meervoud gekeken. In dit onderzoek worden echter bij de methode voor extractie uit vrije tekst gestemde hyponiemparen geëvalueerd, en zorgen de foutief gestemde hyponiemparen voor de verlaging van de precisie met 11,5%.

Daarnaast worden verschillende definities voor 'correct' gebruikt. Terwijl in dit onderzoek alleen een zeer strikte definitie van correct wordt gebruikt, keuren Cederberg en Widdows (2003) ook hyponiemparen goed die pas na kleine modificaties correct zijn. Morin en Jacquemin (2002) kijken niet alleen naar het hoofd-nomen, maar selecteren ook eventuele modificeerders. Hoewel de hyponiemrelatie hierdoor minder algemeen wordt en minder geschikt wordt voor automatische uitbreiding van een ontologie, neemt het foutenpercentage wel af doordat niet slechts één woord maar een exactere beschrijving wordt geselecteerd.

Ondanks het verschil in evaluatiemethodes en -criteria kan toch kan ruwweg worden geconcludeerd dat met een precisie van 47%, uitgaande van de meerderheid van de beoordelaars en de strikte beoordelingscriteria, de resultaten bij extractie uit vrije tekst vergelijkbaar zijn met eerdere onderzoeken.
Er is echter één onderzoek dat ondanks de grote gelijkenis met de in deze scriptie beschreven methode voor extractie uit vrije tekst opvallend betere resultaten kent: Rydin extraheert gelemmatiseerde hyponiemparen uit een groot krantencorpus, en selecteert hierbij net als in dit onderzoek slechts het hoofd-nomen van de NP. Na handmatige evaluatie van 20% van de paren haalt Rydin (2002) een precisie van 92%, hetgeen opvallend hoger is dan de resultaten van extractie uit vrije tekst bij dit onderzoek en de resultaten van de andere onderzoeken. Zoals al genoemd geeft Rydin echter geen verklaring voor deze resultaten, waardoor verdere vergelijking met de in dit onderzoek ontwikkelde methodes niet mogelijk is.

De precisie van de extractie uit de Spectrum Encyclopedie is met 87% erg hoog, en is gezien het hoge aantal geëxtraheerde relaties een erg goede methode. Enig nadeel is dat het hier om een heel specifiek corpus gaat, en dat de ontwikkelde methode niet onbeperkt op willekeurige tekstcorpora toe te passen is.

In de volgende paragraaf worden de geëxtraheerde hyponiemparen van beide methodes vergeleken met de bestaande hyponiemrelaties in de EuroWordNet-

ontologie, en wordt er bekeken in hoeverre de geëxtraheerde paren kunnen bijdragen aan de uitbreiding van de bestaande ontologie.

5.3 Evaluatie door vergelijking met bestaande hyponiemrelaties in EuroWordNet

5.3.1 Methode

Voor de evaluatie van de verkregen hyponiemparen gebruikt Hearst (1992) informatie uit WordNet. WordNet is een handmatig gebouwde thesaurus die zo'n 34.000 nomina bevat. Voor alle nomina zijn synoniem- en hyponiemrelaties opgeslagen.

Hearst beschrijft de drie mogelijke uitkomsten van vergelijking van een hyponiem(N0,N1) met de hyponiemrelaties in WordNet:

- Verify: N0 en N1 komen beide in WordNet voor en er bestaat een hyponiemrelatie (direct of indirect) tussen beide.
- Critique: N0 en N1 komen beide in WordNet voor, maar de hyponiemrelatie tussen beide niet. De hyponiemrelatie tussen N0 en N1 wordt als nieuwe relatie geadviseerd[4].
- Augment: N0, N1 of beide komen niet in WordNet voor. Toevoeging van het hyponiempaar aan WordNet wordt geadviseerd.

Deze methode geeft echter geen uitsluitsel over de vraag of een geëxtraheerde relatie goed of fout is: elk hyponiempaar wordt door Hearst correct verondersteld en toegevoegd aan WordNet.
Bij hyponiemparen die als *critique* worden aangemerkt zijn er twee mogelijkheden:

- de relatie is correct, maar nog niet opgenomen in WordNet.
- de relatie is niet correct en is daar niet in WordNet te vinden.

[4]Hierbij moet opgemerkt worden dat ontbrekende relaties -gesteld dat deze correct zijn- niet simpelweg aan WordNet toegevoegd kunnen worden, omdat niet bekend is of het hyponiem en hyperniem dezelfde semantische betekenis hebben als de gevonden begrippen in WordNet.

Voor een hyponiempaar dat als *augment* wordt aangemerkt geldt grotendeels hetzelfde:

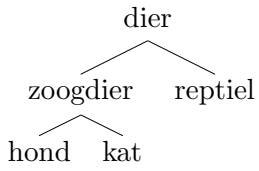
- de relatie is correct, maar een van beide of beide begrippen zijn nog niet in WordNet opgenomen.
- de relatie is niet correct en is daarom niet in WordNet te vinden.

Er vanuit gaande dat WordNet correcte relaties bevat, is alleen bij paren die als *verify* worden aangemerkt met zekerheid te zeggen dat deze correct zijn.

Toch is het zinvol om de vergelijking met WordNet te maken om een beeld te krijgen van in hoeverre de gevonden resultaten al bestaan in WordNet en in hoeverre de gebruikte methode aanvullingen op WordNet kan leveren.

Voor het Nederlands kan deze vergelijking gemaakt worden m.b.v. het Nederlandse deel van EuroWordNet, dat 54.439 nomina bevat. Voor deze nomina zijn net als bij WordNet synoniem- en hyponiemrelaties opgeslagen.

Elk hyponiempaar is vergeleken met WordNet en aangemerkt met een van de door Hearst opgestelde uitkomsten: *verify, critique* of *augment*. Daarnaast is voor alle paren die als *verify* aangemerkt zijn bekeken hoe groot de afstand tussen hyponiem en hyperniem is, m.a.w. of er sprake is van een directe hyponiemrelatie of dat er tussenliggende begrippen zijn.



Directe relatie: hyponiem("reptiel","dier")
Indirecte relatie: hyponiem("hond","dier")

5.3.2 Technische specificaties

Voor de vergelijking van de hyponiemparen met de hyponiemrelaties in EuroWordNet is gebruik gemaakt van de logische programmeertaal Prolog. Zowel alle relaties in EuroWordNet als de geëxtraheerde hyponiemparen zijn vertaald naar Prologpredikaten.
Vervolgens is een Prolog-programma geschreven waarmee de hyponiemrelaties in Prolog kunnen worden opgezocht en geeft het programma één van de

drie eerder beschreven uitkomsten. Daarnaast wordt voor alle relaties die *verify* opleveren bekeken hoe groot de afstand is tussen hyponiem en hyperniem en wordt deze opgeslagen. Als er meerdere 'paden' zijn waardoor hyponiem en hyperniem verbonden zijn, wordt de lengte van het kortste pad opgeslagen.

5.3.3 Resultaten

Extractie uit vrije tekst

Van alle 17.494 paren die gevonden werden in vrije tekst, zijn er 6104 waarvan zowel hyponiem als hyperniem terug te vinden zijn in EuroWordNet. Van deze 6104 paren hebben er 1206 een hyponiemrelatie in EuroWordNet. Deze resultaten zijn terug te vinden in tabel 5.3.

Uitkomst:	**Aantal:**	**Percentage:**
verify:	1206	6,89%
critique:	4898	28,00%
augment:	11400	65,17%

Tabel 5.3: Uitkomsten van vergelijking hyponiemparen verkregen uit vrije tekst en EuroWordNet.

Uit tabel 5.3 blijkt dat voor slechts 19,8% (1206 van 6104) van alle hyponiemparen waarvan beide delen in EuroWordNet te vinden zijn er ook een relatie bestaat tussen hyperniem en hyponiem. Voor de overige 80,2% betekent dit, zoals al eerder besproken, dat de relatie correct is, maar nog niet opgenomen in WordNet, of dat de relatie niet correct is.

Bij nadere beschouwing van de paren die als *critique* worden aangemerkt blijkt dat er een grote groep paren is met correcte relaties, die een aanvulling kunnen vormen voor EuroWordNet. Enkele voorbeelden:

hyponiem("acrobaat","circusartiest")
hyponiem("alcohol","drug")
hyponiem("ambassade","instelling")
hyponiem("AOW","uitkering")
hyponiem("arend","roofdier")
hyponiem("arts","hulpverlener")
hyponiem("asperge","tuinbouwprodukt")
hyponiem("aspirine","pijnstiller")

hyponiem("autodiefstal","misdrijf")
hyponiem("autoradio","audio-apparatuur")
hyponiem("belastingfraude","onrechtmatigheid")

Er is echter ook een grote groep paren te vinden waarbij de relaties niet correct zijn. Veelal is de relatie afhankelijk van de context, zoals bij de volgende voorbeelden:

hyponiem("aannemersbedrijf","dekmantel")
kan een geldige hyponiemrelatie zijn in de context van bijvoorbeeld fraudepraktijken, waarbij een aannemersbedrijf als dekmantel voor illegale activiteiten dient. Zonder deze contextinformatie is de hyponiemrelatie minder voor de hand liggend.

hyponiem("aardewerk","bewijsmateriaal")
kan een geldige hyponiemrelatie zijn in de context van bijvoorbeeld opgravingen, waarbij aardewerk het bewijs kan leveren dat een bepaalde nederzetting ooit op een bepaalde plaats gevestigd was. Ook hier is de relatie zonder de contextinformatie minder waarschijnlijk.

Daarnaast zijn er ook paren die niet correct zijn doordat de selectie van alleen het hoofd van de NP niet voldoende is. Een voorbeeld:

hyponiem("been","verwonding")
hier is een *been* niet een soort van *verwonding*, maar een *gebroken been* wel.

Tenslotte zijn er onder de niet correcte hyponiemparen paren die begrippen bevatten die wel met elkaar te maken hebben, maar een andere relatie dan een hyponiemrelatie hebben. Enkele voorbeelden:

hyponiem("barkeeper","bar")
hyponiem("bijtring","baby")
hyponiem("borst","kliniek")
hyponiem("boter","melk")
hyponiem("brandweerman","zwaailicht")
hyponiem("broer","zus")
hyponiem("dochter","zoon")

Afstand:	**Aantal:**	**Percentage:**
1	587	48,51%
2	370	30,58%
3	161	13,31%
4	53	4,38%
5	17	1,40%
>5	18	1,49%

Tabel 5.4: Afstand in WordNet tussen hyponiem en hyperniem

Van de 1210 paren die als *verify* worden aangemerkt, is de afstand tussen hyponiem en hyperniem in WordNet als volgt:

Uit tabel 5.4 is af te lezen dat meer dan de helft van de gevonden hyponiemparen geen directe hyponiemrelatie bevat. Het hyponiem *auto* bevat bijvoorbeeld vier verschillende hyperniemen met verschillende afstanden tot het hyponiem:

hyponiem("auto","voertuig")	afstand=2
hyponiem("auto","transportmiddel")	afstand=3
hyponiem("auto","vervoermiddel")	afstand=3
hyponiem("auto","voorwerp")	afstand=4

In EuroWordNet zijn deze relaties terug te vinden in één boomstructuur[5]:

```
[iets/1]
   [object/1]
      [ding/1,voorwerp/1]
         [transportmiddel/1,vervoer/2,vervoermiddel/1]
            [vehikel/1,voertuig/1]
               [motorrijtuig/1,motorvoertuig/1]
                  [auto/1,automobiel/1,kar/3,statusblik/1,wagen/2]
```

Hyponiemparen met een grote afstand bevatten vaak een zeer algemeen hyperniem:

hyponiem("wekker","ding")	afstand=6
hyponiem("poster","voorwerp)	afstand=9
hyponiem("boeddha","object")	afstand=8

[5]Het nummer achter elk woord onderscheidt de verschillende betekenissen van ambigue woorden door toekenning van verschillende nummers.

Uit de vergelijking van de geëxtraheerde paren met EuroWordNet kan geconcludeerd worden dat veel van de geëxtraheerde hyponiemen en hyperniemen nog niet in EuroWordNet voorkomen, en dat deze een goede aanvulling kunnen vormen op de bestaande ontologie, mits de precisie van de hyponiemparen verhoogd wordt. Van de hyponiemparen waarvan beide delen wel in EuroWordNet voorkomen, is slechts bij een klein deel ook een hyponiemrelatie aanwezig. Uit de vergelijking blijkt dat dit zowel te wijten is aan het ontbreken van relaties in EuroWordNet, als aan foutieve relaties tussen de geëxtraheerde hyponiemen. Hier komen, net als bij evaluatie van de steekproef, de problemen met andersoortige relaties, contextgevoeligheid en onvolledige hyponiemen/hyperniemen weer naar voren. De hyponiemrelaties die wel in EuroWordNet terug zijn te vinden, zijn veelal indirect. Dit hoeft echter geen probleem te zijn, omdat deze relaties goed van pas kunnen komen bij de uitbreiding van een bestaande ontologie zoals EuroWordNet, zoals bijvoorbeeld bij de methode beschreven door Rydin (2002, besproken in paragraaf 2.1).

Extractie uit de Spectrum Encyclopedie

Van alle 52.063 paren die gevonden werden in de Spectrum Encyclopedie, zijn er 3569 waarvan zowel hyponiem als hyperniem terug te vinden zijn in EuroWordNet. Dit is een erg laag aantal vergeleken met de resultaten van de hyponiemparen die uit vrije tekst geëxtraheerd zijn. Een van de oorzaken hiervoor is eenvoudig aan te wijzen: 68% van de hyponiemparen uit de Spectrum Encyclopedie bevat een of meerdere eigennamen (bij de hyponiemparen geëxtraheerd uit vrije tekst is dit maar 42,2%). Aangezien EuroWordNet slechts 1745 eigennamen bevat, zal het overgrote deel van de hyponiemparen met eigennamen uit de Spectrum Encyclopedie als *augment* worden aangemerkt.
Daarnaast kan het soort tekst van invloed zijn: in dit geval gaat het om een encyclopedietekst, die veel specifieke en niet-alledaagse begrippen bevat. Mogelijkerwijs levert dit een ander type hyponiemparen op dan in EuroWordNet zijn te vinden.

Van de 3569 paren waarvan het hyponiem en hyperniem wel in EuroWordNet gevonden worden, hebben er 1679 een hyponiemrelatie in EuroWordNet. Deze resultaten zijn terug te vinden in tabel 5.5.

Dit aantal van 1674 paren (=46,9%) is opvallend hoger dan de 19,8% van de uit vrije tekst geëxtraheerde hyponiemparen. Dit lijkt aan te duiden dat de kwaliteit van de hyponiemrelaties beduidend beter is.

Uitkomst:	**Aantal:**	**Percentage:**
verify:	1674	3,22%
critique:	1895	3,63%
augment:	48566	93,28%

Tabel 5.5: Uitkomsten van vergelijking hyponiemparen verkregen uit de Spectrum Encyclopedie en EuroWordNet

De afstand tussen hyponiem en hyperniem van de 1674 in WordNet bestaande relaties is te vinden in tabel 5.6:

Afstand:	**Aantal:**	**Percentage:**
1	1208	72,16%
2	326	19,47%
3	96	5,73%
4	30	1,79%
5	9	0,54%
>5	5	0,30%

Tabel 5.6: Afstand in WordNet tussen hyponiem en hyperniem

Opvallend is dat er bij deze hyponiemparen veel meer directe hyponiemrelaties zijn dan bij de paren geëxtraheerd uit vrije tekst (respectievelijk 72,2% en 48,5%). Mogelijkerwijs komt dit doordat men in een encyclopedie elk begrip zo precies mogelijk probeert te beschrijven en geen gebruik probeert te maken van algemene termen.

Nadere beschouwing van de 1895 paren die als *critique* worden aangemerkt laat zien dat er een grote groep paren is met correcte relaties, die een aanvulling kunnen vormen voor EuroWordNet. Enkele voorbeelden:

hyponiem(”aardappelziekte”,”schimmelziekte”)
hyponiem(”acajou”,”boom”)
hyponiem(”acne”,”huidaandoening”)
hyponiem(”adhesie”,”kracht”)
hyponiem(”adressenbestand”,”databank”)
hyponiem(”androsteron”,”geslachtshormoon”)
hyponiem(”autisme”,”ontwikkelingsstoornis”)
hyponiem(”basalt”,”stollingsgesteente”)
hyponiem(”basso continuo”,”basstem”)
hyponiem(”benzinemotor”,”verbrandingsmotor”)

Bij de paren die als *critique* worden aangemerkt, is slechts een zeer klein deel echt niet correct is. Voornaamste oorzaak hiervoor is dat het hyperniem niet te beschrijven is in één begrip waardoor onderspecificatie optreedt. Enkele voorbeelden:

hyponiem("acidimetrie","onderdeel")
Onderdeel van de analytische chemie.

hyponiem("afplatting","mate")
Mate waarin een hemellichaam van de bolvorm afwijkt.

hyponiem("agnosticisme","leer")
De *leer* dat men omtrent het wezen der niets met zekerheid kan zeggen.

hyponiem("ademhaling","opname")
Opname van zuurstof en afgifte van kooldioxide.

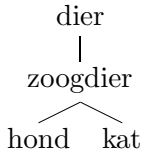
hyponiem("afschrijving","vaststelling")
Vaststelling van de waardevermindering van vaste activa in de tijd.

hyponiem("alcoholmisbruik","inname")
De buitensporige *inname* van alcohol.

Uit de vergelijking van de geëxtraheerde paren met EuroWordNet kan geconcludeerd worden dat de hyponiemparen een uitstekende aanvulling kunnen vormen op de bestaande ontologie. In tegenstelling tot EuroWordNet, bevatten de uit de Spectrum Encyclopedie geëxtraheerde hyponiemparen veel eigennamen, welke voor Question Answering erg belangrijk kunnen zijn. Hoewel van weinig van de geëxtraheerde paren beide delen in EuroWordNet terug te vinden zijn, is de kwaliteit van de paren waarbij dit wel zo is hoog. Het niet aanwezig zijn van een relatie tussen het hyponiem en het hyperniem lijkt grotendeels te wijten te zijn aan de incompleetheid van EuroWordNet, en niet aan de kwaliteit van de geëxtraheerde relaties (hoewel onderspecificatie voor enkele foute relaties zorgt). Gezien het feit dat de afstand tussen hyponiem en hyperniem in de meerderheid van de gevallen erg klein is, kan geconcludeerd worden dat de hyponiemrelaties goed gespecificeerd zijn.
Door de hoge precisie en het grote aantal hyponiemrelaties lijken de uit de Spectrum Encyclopedie geëxtraheerde relaties uitermate geschikt voor

het uitbreiden van de EuroWordNet-ontologie. De methode van Rydin (2002) lijkt hiervoor echter minder geschikt dan voor uitbreiding met paren geëxtraheerd uit vrije tekst. Voor plaatsing in de bestaande ontologie zijn er bij de methode van Rydin minstens twee voorkomens van elk hyperniem en hyponiem nodig:

> De volgende hyponiemrelaties
>
> hyponiem("zoogdier","dier")
> hyponiem(["hond"|"kat"],"dier")
> hyponiem("hond","zoogdier")
>
> leveren bij samenvoeging de volgende boomstuctuur op:



De hyponiemen geëxtraheerd uit de Spectrum Encyclopedie zijn echter uniek, en daarnaast is de overlap met paren in de EuroWordNet-ontologie zoals blijkt uit de vergelijking, met name door de vele eigennamen, klein.
Voor de uitbreiding moet wellicht gezocht worden naar een combinatie van bestaande methodes of een geheel nieuwe methode.

Hoofdstuk 6

Conclusie en discussie

In deze scriptie zijn twee methodes voor de automatische extractie van hyponiemrelaties uit grote tekstcorpora beschreven. Hoewel beide methodes gebruik maken van lexicaal-syntactische patronen voor de extractie van hyponiemparen, verschillen zij qua implementatie, toepasbaarheid op corpora en resultaten.

In de eerste paragraaf worden de resultaten van beide methodes nogmaals kort besproken, waarbij enkele aanbevelingen voor verder onderzoek worden gedaan. Daarnaast worden in paragraaf 2 enkele suggesties gegeven voor het automatisch uitbreiden van de bestaande EuroWordNet-ontologie met de bestaande data.

6.1 Automatische extractie van hyponiemrelaties uit grote tekstcorpora

Het doel van dit onderzoek was om te bepalen of met behulp van lexicaal-syntactische patronen hyponiemrelaties uit vrije tekst kunnen worden geëxtraheerd. Hiervoor zijn twee methodes ontwikkeld:

- Extractie van hyponiemparen uit vrije tekst met behulp van lexicaal-syntactische patronen, gebaseerd op Engels onderzoek door Hearst (1992,1998).
- Extractie van hyponiemparen uit het gestructureerde Nederlandse Spectum Encyclopedie-corpus.

De methode voor extractie van hyponiemparen uit vrije tekst met behulp van enkele lexicaal-syntactische patronen is gebaseerd op de voor het Engels ontwikkelde methode van Hearst (1992,1998). De Nederlandse methode extraheert uit een corpus van 6.134.931 zinnen 17.494 hyponiemparen, en haalt hierbij een precisie van 47%.
In het totale foutenpercentage hebben fouten die ontstaan zijn door de buiten het bereik van dit onderzoek liggende stemmer en Alpino ontleder een groot aandeel. Het gebruik van een goede stemmer en ontleder kunnen de resultaten bij verder onderzoek aanzienlijk verbeteren.
De eerder genoemde problemen met contextgevoeligheid, incorrecte hyponiemrelaties na stemming, PP-aanhechting en uitsluiting van adjectieven worden ook in vergelijkbaar onderzoek voor het Engels (Hearst 1992, Hearst 1998, Cederberg & Widdows 2003), Frans (Jacquemin & Morin 2002) en Zweeds (Rydin 2002) genoemd.
De ontstane fouten door contextgevoeligheid lijken lastig te voorkomen. Een oplossing kan zijn om een minder contextgevoelig corpus, zoals een encyclopediecorpus, te gebruiken. Deze corpora zijn voor het Nederlands echter aanzienlijk minder beschikbaar dan de grote groep krantencorpora.
Problemen met PP-aanhechting en uitsluiting van adjectieven kunnen worden voorkomen door het gebruik van striktere lexicaal-syntactische patronen. Hoewel dit de precisie zal verhogen, gaat het ten koste van de *recall*. Omdat er voor het Nederlands geen eerdere resultaten beschikbaar zijn, is in dit onderzoek gekozen voor een hoge *recall* door het gebruik van vrij soepele lexicaal-syntactische patronen. Hierdoor is een goede basis verkregen voor verder onderzoek, doordat duidelijk is geworden hoe de patronen kunnen worden aangescherpt. Verder onderzoek zal moeten uitwijzen wat het effect is van de toepassing van striktere lexicaal-syntactische patronen op de precisie en *recall*, en waar een goede balans ligt tussen beide.
Naast het gebruik van striktere patronen ter verhoging van de precisie, geven Cederberg en Widdows (2003) aan dat het gebruik van LSA de precisie kan verhogen. Met behulp van de door Cederberg en Widdows (2004) ontwikkelde *Infomap NLP software*lijkt deze methode in verder onderzoek relatief eenvoudig te kunnen worden geëvalueerd.

De methode voor extractie van hyponiemparen uit de gestructureerde Spectrum Encyclopedie geeft ondanks de eenvoud van de methode opvallend goede resultaten: uit 62.103 encyclopedie-ingangen worden 52.063 hyponiemparen geëxtraheerd met een precisie van 87%. Deze extractiemethode is echter specifiek ontwikkeld voor de structuur van de Spectrum Encyclopedie en daardoor niet toepasbaar op andere corpora. Deze eenmalige extractie

kan echter bij uitbreiding van de bestaande EuroWordNet-ontologie wel zorgen voor bijna een verdubbeling van het aantal nomen-hyponiemrelaties. De meeste foute hyponiemparen ontstaan door het niet stemmen van de paren waardoor deze in het meervoud blijven staan en door onderspecificatie. Het verbeteren van de accuratesse van de stemmer zou het foutenpercentage van de hyponiemrelaties behoorlijk kunnen verlagen. Hoewel in dit onderzoek de externe stemmer werd verkozen boven de interne Alpino-stemmer, zou de Alpino-stemmer met enkele kleine aanpassingen erg bruikbaar kunnen zijn voor verder onderzoek: door bij samenstellingen het eerste deel/de eerste delen van de samenstelling en de *root* van het laatste deel te combineren, ontstaat in de meeste gevallen een goede stam. Hierdoor zal de accuratesse behoorlijk stijgen en heeft de Alpino-stemmer, mede door z'n goede taalkundige kennis, veel potentie om geschikt te zijn voor het stemmen van hyponiemparen. Het probleem met onderspecificatie lijkt lastiger te voorkomen en is bovendien onafhankelijk van de gebruikte stemmer.

6.2 Automatische uitbreiding van de EuroWordNet-ontologie

De automatische extractie van hyponiemrelaties is een eerste stap in het proces van het automatisch uitbreiden van een ontologie. Voor het Nederlands is één grote algemene ontologie beschikbaar: EuroWordNet. Er is echter nog geen onderzoek gedaan naar de automatische uitbreiding van deze ontologie. Gebaseerd op de bestudeerde literatuur en bevindingen uit eigen onderzoek worden in deze laatste paragraaf enkele aanbevelingen gedaan voor de automatische uitbreiding van EuroWordNet in toekomstig onderzoek.

Uit vergelijking van beide groepen geïdentificeerde hyponiemrelaties met de bestaande relaties in de EuroWordNet is gebleken dat beide groepen een goede aanvulling kunnen vormen op de bestaande hyponiemrelaties in EuroWordNet, omdat de overlap slechts gering is.
De beide extractiemethodes leveren totaal verschillende soorten data op. De extractie uit vrije tekst levert een brede selectie van hyponiemrelaties op, variërend van zeer (situatie-)specifieke hyponiemparen en eigennamen tot zeer algemene veel voorkomende hyponiemrelaties. De geëxtraheerde hyponiemparen uit de Spectrum Encyclopedie bevatten veel eigennamen, zoals plaatsnamen en namen van historische figuren. Daarnaast worden er veel zeer specifieke directe hyponiemrelaties geëxtraheerd.
Samenvoeging van deze nieuwe relaties met de bestaande relaties in Eu-

roWordNet levert een grote ontologie met een grote diversiteit aan informatie op, hetgeen bruikbaar is voor toepassing in Question Answering.

Voordat enkele suggesties worden gegeven voor het automatisch uitbreiden van de Nederlandse EuroWordNet-ontologie, moeten eerst twee opmerkingen worden geplaatst over de opslag van de data. Ten eerste moeten de geëxtraheerde hyponiemrelaties niet als tweedelig paar worden opgeslagen (zoals steeds is beschreven in voorgaande hoofdstukken), maar als reeksen, zodat de zusterrelatie tussen hyponiemen behouden blijft. Voorbeeld:

> Sommige naaldbomen, zoals de spar of de den ...
>
> Niet:
> hyponiem("spar","naaldboom")
> hyponiem("den","naaldboom")
>
> Maar:
> hyponiem(["spar","den"],"naaldboom")

Ten tweede moet voor een hyponiempaar waarbij ook een adjectief opgeslagen is, zoals

> hyponiem("Philips","Nederlands bedrijf")

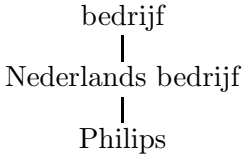
en de daarvan afgeleide hyponiemrelatie

> hyponiem("Nederlands bedrijf","bedrijf")

de koppeling tussen beide hyponiemparen ook worden opgeslagen, zoals bijvoorbeeld door toevoeging van een unieke identificatiecode aan elk hyponiem en hyperniem, waarbij identieke begrippen dezelfde code krijgen:

> hyponiem([1234,"Philips"],[9876,"Nederlands bedrijf]")
> hyponiem([9876,"Nederlands bedrijf"],[4567,"bedrijf"])

Hierdoor wordt de volgende relatie vastgelegd in de data:



Deze unieke identificatiecodes zijn nodig om ambigue woorden van elkaar te onderscheiden. Door aan de twee voorkomens van *Nederlands bedrijf* dezelfde identificatiecode toe te kennen, wordt expliciet vastgelegd dat deze begrippen hetzelfde zijn, en dat deze probleemloos aan elkaar gekoppeld kunnen worden in latere stadia van het onderzoek. Voor ambigue woorden, zoals het Nederlandse woord *bank*, krijgt elke betekenis van het woord een andere identificatiecode.

Het automatisch uitbreiden van een bestaande ontologie is niet eenvoudig. Caraballo (1999), Alfonseca en Manandhar (2002) en Rydin (2002) beschrijven alledrie totaal verschillende methodes voor het bouwen of uitbreiden van een ontologie.
Uit vergelijking is gebleken dat de methode van Rydin (2002) het meest geschikt lijkt voor het automatisch uitbreiden van de Nederlandse EuroWordNet-ontologie. Voordeel van deze methode is dat relaties relatief eenvoudig zonder extra benodigde semantische kennis aan- of tussengehangen kunnen worden. Nadeel is dat een groot aantal hyponiemrelaties vereist is.
Met de bestaande 54.439 hyponiemparen in EuroWordnet en de 69.557 geëxtraheerde hyponiemrelaties lijkt echter een goede basis aanwezig te zijn voor het slagen van dit algoritme.
Hierbij moet echter wel opgemerkt worden dat alle 52.063 hyponiemen geëxtraheerd uit de Spectrum Encyclopedie uniek zijn (doordat zij zijn afgeleid van unieke encyclopedie-ingangen), waardoor koppeling van hyponiemrelaties volgens de methode van Rydin (2002) bemoeilijkt wordt. Hiervoor zijn immers meerdere voorkomens van elk te koppelen hyperniem en hyponiem nodig.

Het lijkt de moeite waard om in toekomstig onderzoek te bekijken of de methode van Rydin, wellicht in combinatie met elementen uit andere methodes, bruikbaar is voor de uitbreiding van de EuroWordNet-ontologie. Ter afsluiting van dit onderzoek wordt kort een suggestie voor een mogelijk uitbreidingsalgoritme gegeven.

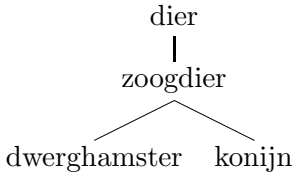
Naast het algoritme voor koppeling van hyponiemrelaties gebruikt in de methode van Rydin (2002), lijkt coördinatie van nomina (o.a. Widdows & Dorow 2002, Cederberg & Widdows 2003) een bruikbare en betrouwbare methode voor de uitbreiding van EuroWordNet. Combinatie van kennis over hyponiemrelaties en coördinatierelaties kan mogelijkerwijs zorgen voor een goed algoritme voor koppeling van hyponiemrelaties.
Dit algoritme voegt dan, net als Rydin (2002), hyponiemparen samen tot klasses, en probeert vervolgens een klasse in de bestaande hiërarchie te plaatsen op basis van het volgende principe:

> Als X een soort van Y is, en Y een soort van Z, dan kunnen deze gekoppeld worden als er in het lexicon/de hiërarchie een hyponiemrelatie te vinden is tussen X en Z.

Een simpel voorbeeld:

> Klasse 1: hamster - dwerghamster

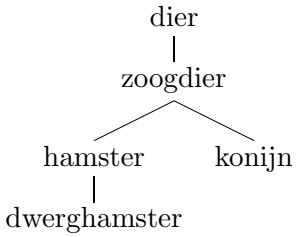
kan in de volgende bestaande hiërarchie geplaatst worden



als ook de volgende klasse aanwezig is:

> Klasse 2: zoogdier - hamster

Dit levert de volgende hiërarchische structuur op:

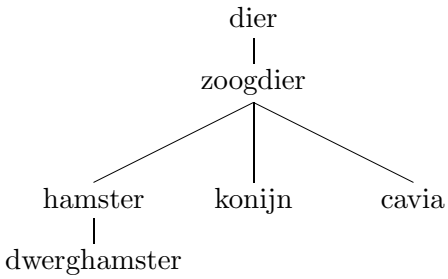


Als een klasse echter niet m.b.v. dit principe van Rydin in de ontologie geplaatst kan worden, wordt vervolgens gekeken naar coördinatie van het hyponiem.
Weer een simpel voorbeeld:

> Klasse 3: zoogdier - cavia

Deze klasse kan in de laatstgegeven hiërarchische structuur geplaatst worden onder het hyperniem *zoogdier*, als in één van de beschikbare corpora een coördinatierelatie kan worden gevonden tussen *hamster-cavia* of *konijn-cavia*.

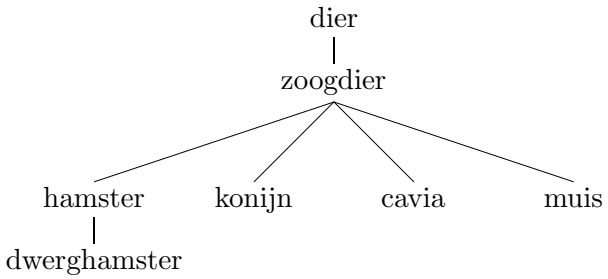
Dit levert de volgende hiërarchische structuur op:



Met behulp van deze methode kunnen -theoretisch- veel klassen met hyponiemrelaties op een betrouwbare manier in de bestaande ontologie worden geplaatst. Als een klasse met deze methode niet in de ontologie kan worden geplaatst, kan eventueel nog teruggevallen worden op LSA (Cederberg & Widdows 2003):

Klasse 4: zoogdier - muis

Deze klasse kan in de hiërarchische structuur geplaatst worden onder het hyperniem *zoogdier*, als m.b.v. LSA berekend is dat er een waarschijnlijke gelijkheid is tussen *muis* en één van de hyponiemen *konijn*, *hamster* of *cavia*:



De hierboven beschreven methode is slechts bedoeld als suggestie voor een basismethode voor automatische uitbreiding. Verder onderzoek moet uitwijzen in hoeverre deze basismethode bruikbaar is om de geëxtraheerde hyponiemrelaties in de bestaande EuroWordNet-ontologie te plaatsen, en welke aanpassingen en aanvullingen noodzakelijk zijn.

Bibliografie

[1] Alfonseca, Enrique and Suresh Manandhar. 2002. Improving an ontology refinement method with hyponymy patterns. In *Third International Conference on Language Resources and Evaluation*, p. 235 - 239, Las Palmas, Spain.

[2] Bouma, Gosse, Gertjan van Noord and Robert Malouf. 2001. Alpino: Wide-coverage Computational Analysis of Dutch. In W. Daelemans, K. Sima'an, J. Veenstra, and J. Zavrel, editors, *Computational Linguistics in The Netherlands 2000*, p. 45 - 59.

[3] British National Corpus. 1994. http://www.natcorp.ox.ac.uk

[4] Caraballo, Sharon. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *37th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, p. 120 - 126.

[5] Carletta, Jean. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2): 249 - 254.

[6] Cederberg, Scott and Dominic Widdows. 2003. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In *Seventh Conference on Computational Natural Language Learning (CoNLL-2003)*, Edmonton, Canada, June 2003, pages 111-118.

[7] Cederberg, Scott and Dominic Widdows. 2004. *Infomap NLP software*, verkregen van http://sourceforge.net/projects/infomap-nlp/ op 01-04-2004.

[8] Curran, James R. and Marc Moens. 2002. Improvements in Automatic Thesaurus Extraction. In *Proceedings of the Workshop on Unsupervised Lexical Acquisition*, Philadelphia, PA, USA, p. 59 - 67.

[9] Deerwester, Scott, Susan Dumais, George Furnas, Thomas Landauer and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391 - 407.

[10] Feldbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge MA.

[11] Gaustad, Tanja and Gosse Bouma. 2002. Accurate Stemming of Dutch for Text Classification. In *Computational Linguistics in the Netherlands 2001*, p. 104 - 117.

[12] Hahn, Udo and Klemens Schnattinger. 1998. Towards Text Knowledge Engineering. In *AAAI/IAAI*, p. 524-531.

[13] Hastings, P.M. 1994. *Automatic Acquisition of Word Meanings from Context*. University of Michigan, Dissertation.

[14] Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING*, Nantes, France.

[15] Hearst, Marti A. 1998. *WordNet: An Electronic Lexical Database*, chapter 5, Automated discovery of WordNet relations, p. 131-152.

[16] Kraaij, W. and R. Pohlman. 1994. Porter's stemming algorithm for Dutch. In L. Noordman and W. de Vroomen (eds), *Informatiewetenschap 1994: Wetenschapelijke bijdragen aan de derde STINFON Conferentie*, Tilburg, p. 167-180.

[17] Manning, Christopher D. and Hinrick Schütze. 1999. Chapter 8.5 Semantic Similarity. In *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, p. 294-308.

[18] Miller, George A. and William G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1-28.

[19] Morin, Emmanuel and Christian Jacquemin. 2003. Automatic acquisition and expansion of hypernym links. *Computer and the Humanities*, forthcoming.

[20] Rajman, M. and A. Bonnet. 1992. Corpora-based Linguistics: New Tools for Natural Language Processing. In *1^{s}t Annual Conference of the Association for Global Strategic Information*, Germany. Bad Kreuznach.

[21] Riloff, Ellen and Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. In Claire Cardie and Ralph Weischedel, editors, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, p. 117 - 124. Association for Computational Linguistics, Somerset, New Jersey.

[22] Roark, Brian and Eugene Charniak. 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *COLING-ACL*, p. 1110 - 1116.

[23] Rydin, Sara. 2002. Building a hyponymy lexicon with hierarchical structure. In *Proceedings of the SIGLEX Workshop on Unsupervised Lexical Acquisition*, ACL'02, Philadelphia, Pennsylvania, pp. 26-33.

[24] Schütze H. 1998. Automatic Word Sense Discrimination. In *Computational Linguistics*, 24(1):97-124.

[25] Vossen, P. (eds) 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht.

[26] Widdows, Dominic and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *19^{t}h International Conference on Computational Linguistics*, p. 1093 - 1099, Taipei, Taiwan.