Internet als corpus voor onderzoek naar woordbetekenis

Christian Huijsmans
28 juli 2003
Faculteit der Letteren
Afstudeerrichting Bedrijfscommunicatie & Digitale Media

Internet als corpus voor onderzoek naar woordbetekenis

"The corpus of the new millennium is the Web" (A. Kilgarriff)

Christian Huijsmans
28 juli 2003

Inhoudsopgave

INHOUDSOPGAVE **3**

VOORWOORD **5**

1. INLEIDING **6**

1.1 WORD SENSE DISAMBIGUATION 6
1.1.1 Wat is WSD? *6*
1.1.2 Positie van WSD binnen NLP *7*
1.2 ONDERZOEKSVELD 8
1.2.1 Geschiedenis van het onderzoek *8*
1.2.2 Lerende systemen als empirisch onderzoek binnen NLP *9*
1.2.3 General vs. small scale *10*
1.2.4 SENSEVAL *10*
1.3 VRAAGSTELLING 11
1.3.1 Algemene onderzoeksvraag *11*
1.3.2 Deelvragen *11*
1.3.3 Hypotheses *12*
1.3.4 Doelen van het onderzoek *13*
1.4 OPBOUW VAN DE SCRIPTIE 13

2. SENSES EN HIËRARCHIEËN **15**

2.1 SENSES 15
2.1.1 Het kiezen van senses *15*
2.1.2 Probleem 1: Verschillen tussen woordenboeken *16*
2.1.3 Probleem 2: Homonymie en polysemie *16*
2.1.4 Probleem 3: Granularity *17*
2.1.5 Probleem 4: Data-sparseness probleem *18*
2.2 HIËRARCHIEËN ALS MOGELIJKE OPLOSSING 18
2.2.1 Voordelen van hiërarchieën *19*
2.2.2 Een nadeel van hiërarchieën *19*
2.2.3 Voorbeeld van een hiërarchie *20*
2.2.4 Verdere toepassingen van hiërarchieën *20*
2.3 ENTROPIE 21
2.4 CONTEXT 22
2.5 HET INTERNET ALS CORPUS 23

3. METHODE **24**

3.1 DATAVERZAMELING 24
3.1.1 Werkwijze voor het verzamelen van instanties en context *24*
3.1.2 Criteria voor het verzamelen van instanties en context *24*
3.2 OPBOUW VAN HET INTERNET-CORPUS 26
3.2.1 Voorbereidende handelingen verricht op het corpus *26*
3.2.2 Het taggen van betekenissen *26*
3.3 DE TWEE DATASETS 27
3.3.1 Het kinderboeken-corpus *27*
3.3.2 Het Internet-corpus *28*

3.4 MATERIAAL 28
3.4.1 Zoekmachine Google *28*
3.4.2 Memory Based Tagger *29*
3.4.3 MBWSD-D en classificatie *29*

4. RESULTATEN 30

4.1 ENTROPIE: RESULTATEN 30
4.2 ACCURAATHEID: RESULTATEN 31
4.3 VERBANDEN TUSSEN ENTROPIE EN ACCURAATHEID 33

5. CONCLUSIES 36

5.1 DE BETEKENISVERDELING 36
5.2 DE BETEKENISTOEKENNING 37
5.3 ENTROPIE ALS VOORSPELLER VAN ACCURAATHEID 37
5.4 INTERNET GESCHIKT ALS CORPUS? 38

6. DISCUSSIE 39

6.1 DISCUSSIE OVER ONDERZOEKSOPZET 39
6.1.1 Menselijke annotator als bottleneck *39*
6.1.2 De schaal van het onderzoek *40*
6.1.3 Toekomstige verbeteringen *40*
6.2 GESCHIKTHEID VAN TEKSTEN OP INTERNET 41
6.2.1 Het zoeken van tekst op Internet *42*
6.2.2 Specifieke Internet-aspecten van verdeling van betekenissen *42*
6.2.3 Kwaliteit van de teksten *43*
6.3 DISCUSSIE MOGELIJKHEDEN HIËRARCHIEËN 44

SLOTWOORD 46

LITERATUUR: 47

Voorwoord

Binnen de Universiteit van Tilburg is door de onderzoeksgroep ILK binnen de Faculteit der Letteren onderzoek verricht naar de mogelijkheid om met behulp van een geannoteerd corpus bestaande uit teksten afkomstig uit willekeurige kinderboeken voorspellingen te doen over de betekenis van polyseme woorden. Polyseme woorden zijn woorden die meerdere betekenissen hebben.
Onderzoek met behulp van dit corpus (bestaande uit kinderboeken) leverde een lijst met polyseme woorden op, die tijdens het testen met een WSD-systeem moeilijk te disambigueren bleken. Een voorname oorzaak hiervan was dat bij veel woorden er een gebrek was aan trainingsmateriaal voor een bepaalde betekenis.
Op de kennis die bij het hiervoor genoemde onderzoek verworven is, kan voortgebouwd worden door voor een hoeveelheid willekeurig gekozen woorden uit het kinderboeken-corpus die problemen leverden bij het automatisch toekennen van een betekenis, een nieuwe dataverzameling met instanties van juist die woorden op te zetten, maar dan met het Internet als bron voor voorbeelden van instanties waarin deze woorden gebruikt worden. Voor deze steekproef zal gekeken worden of het Internet een beter - zo niet ander - aanbod aan woordbetekenissen geeft in vergelijking met de betekenissen die gegeven worden in wat we van nu af aan het 'kinderboeken-corpus' zullen noemen.
Dit exploratieve onderzoek zal uiteindelijk onderzoekers in het onderzoeksveld van word sense disambiguation een aantal aanbevelingen geven over de mogelijkheid om willekeurige teksten op Internet te gebruiken om een corpus op te bouwen voor het onderzoek naar word sense disambiguation, oftewel het bepalen van woordbetekenissen. Om deze aanbevelingen te geven zal duidelijk gemaakt worden wat de problemen zijn van het gebruik van teksten afkomstig van een open domein als het Internet, alsook de verbreding van het aantal betekenissen of verandering in de set mogelijke betekenissen die kan voortkomen uit het gebruik van een grotere hoeveelheid instanties als testmateriaal voor een WSD-systeem.

1. Inleiding

In deze inleiding wordt eerst besproken wat onder word sense disambiguation verstaan kan worden. Vervolgens wordt in het kort beschreven hoe het onderzoeksveld eruit ziet. Hieruit blijkt dat het hedendaagse onderzoek zich vooral richt op de betekenisverdeling die geleverd wordt in woordenboeken of die ontstaat uit corpora die specifiek zijn geworven uit bepaalde bronnen.
Omdat deze aanpak maar beperkte mogelijkheden biedt wil dit onderzoek het gat vullen dat te vinden is in het bestaande onderzoek, door de mogelijkheid te onderzoeken of een corpora met willekeurige teksten afkomstig van een bron waaruit iedereen kan putten, namelijk het World Wide Web, geschikt zou zijn om te gebruiken binnen WSD-onderzoek.
Aan het einde van dit inleidende hoofdstuk worden de algemene onderzoeksvraag en de verwachtingen voor dit onderzoek behandeld en wordt een overzicht gegeven van wat er in de verdere hoofdstukken besproken zal worden.

1.1 Word sense disambiguation

Een groot aantal woorden in veel talen heeft meer dan één betekenis. Wanneer deze talen goed gedocumenteerd zijn kunnen deze betekenissen teruggevonden worden in bijvoorbeeld woordenboeken.
Wanneer een ambigu of polyseem woord voorkomt in een boek of in een conversatie kan over het algemeen maar één van de betekenissen van toepassing zijn. Mensen ondervinden meestal geen problemen met het begrijpen van deze woorden en zijn het niet bewust dat ze aan het bepalen zijn welke betekenis een ambigu woord zou moeten hebben (Kilgarriff & Palmer, 2000). Voor computers echter is disambigueren een moeilijke taak. Binnen het onderzoek naar Word Sense Disambiguation wordt er naar gestreefd om computers toch goed te laten presteren bij het disambigueren van polyseme woorden.

1.1.1 Wat is WSD?

De Engelse, en meest gebruikte, aanduiding voor het vakgebied waaronder deze scriptie valt is Word Sense Disambiguation (WSD). In (Hoste, Daelemans, Hendrickx & van den Bosch, 2002) wordt WSD omschreven als "the task ... to assign a sense label to a word in context". Een ruimere definitie kan gevonden worden in een inleidend artikel van Ide & Véronis (Ide & Véronis, 1998): "in general terms word sense disambiguation involves the association of a given word in a text or discourse with a definition or meaning (sense) which is distinguishable from other meanings potentially attributable to that word".
Sommige definities geven aan hoe deze taak volbracht zal moeten worden, bijvoorbeeld door het gebruik maken van de context rond een woord: “Word Sense Disambiguation is the problem of assigining a sense to an ambiguous word, using its context” (Karov & Edelman, 1998).

Ide en Véronis (Ide & Véronis, 1998) onderscheiden twee stappen binnen de subtaak van WSD. De eerste stap is het vaststellen van alle verschillende betekenissen van een woord die relevant zijn ten opzichte van de tekst (sense discrimination). De tweede stap is vervolgens het toekennen van de correcte betekenis aan elke instantie van het woord (sense labelling). Verderop in deze scriptie wordt toegelicht dat beide stappen de nodige problemen met zich mee zullen brengen en dan voornamelijk de eerste stap.
WSD is geen doel op zich, maar eerder een belangrijk onderdeel binnen Natural Language Processing systemen (Kilgariff & Rosenzweig, 2000). Hierop komen we in de volgende paragraaf nog terug. Door juist ook voor verschillende talen en genres WSD-systemen te

ontwikkelen wordt gehoopt het probleem van de lexicale ambiguïteit, dat nu verre van opgelost is, in de toekomst wel degelijk op te lossen. Binnen de SENSEVAL competitie (zie paragraaf 1.2.6) wordt gestreefd naar het bouwen van WSD-systemen die zo goed mogelijk presteren op het disambigueren.

1.1.2 Positie van WSD binnen NLP

Het algemene onderzoeksveld waarbinnen dit onderzoek valt is dat van Natural Language Processing. Dit ruime gebied bestrijkt allerlei applicaties die, wanneer aan elkaar gekoppeld, taal kunnen begrijpen en/of kunnen produceren. Het doel van dit onderzoeksveld is om computers uiteindelijk net zo goed met taal om te kunnen laten gaan als mensen.
De applicaties binnen dit onderzoeksveld zijn onder andere: discourse analysis (gespreksanalyse), machine translation (automatische vertaling) en semantic analysis (betekenis analyse) (Brill & Mooney, 1997). Door het internationale karakter van het onderzoeksveld worden deze applicaties meestal met hun Engelstalige termen genoemd. Deze zullen dan ook in deze scriptie gebruikt worden.
Laatstgenoemde applicatie, semantic analysis, probeert een deel van taal te begrijpen aan de hand van de betekenis van woorden en heeft twee subtaken: WSD en semantic parsing. De subtaak WSD houdt zich, zoals hiervoor ook al is vermeld, voornamelijk bezig met het beslissen welke van de mogelijke betekenissen correct is in een bepaalde context. De andere subtaak van semantic analysis, semantic parsing, houdt zich meer bezig met semantische rollen, zoals het bepalen van agent of instrument binnen een zin. Dit zal verder niet besproken worden in deze scriptie.
Ook Kilgarriff & Palmer (Kilgarriff & Palmer, 2000) noemen een aantal applicaties waarbinnen WSD een belangrijke stap is:

- Machine Translation: een ambigu woord in een bepaalde taal heeft meestal meerdere betekenissen in een andere taal. Een voorbeeld hiervan is het Nederlandse woord ‘blik’, dat in het Engels vertaald kan worden als ‘glance’ of ‘can’ (beans), terwijl omgekeerd ook het Engelse woord ‘can’ twee verschillende Nederlandse betekenissen kan krijgen (‘kan’ en ‘blik’). WSD helpt bij de keuze voor de correcte vertaling;
- Information Retrieval: bij het terugvinden van documenten worden documenten teruggevonden met de niet-bedoelde betekenis als topic. Een voorbeeld hiervan is het Engelse woord ‘jaguar’ dat als zoekwoord zowel informatie over het dier als over de auto kan geven. Door middel van WSD kan bepaald worden welke info de gebruiker wil ontvangen en kan een IR-systeem een beter resultaat teruggeven doordat een hogere precision behaald wordt;
- Text-to-speech systems: hier worden fouten gemaakt in de uitspraak van woorden die hetzelfde geschreven zijn, maar ondertussen niet hetzelfde uitgesproken worden. Een voorbeeld hiervan is het Engelse woord ‘suspect’ dat in de betekenis als zelfstandig naamwoord anders uitgesproken wordt dan in de betekenis als werkwoord.

Dit zijn een aantal voor de hand liggende voorbeelden. Voor bijna alle Natural Language Processing taken is WSD een deelprobleem en kan een goede uitvoering ervan naar alle waarschijnlijkheid een substantiële verbetering opleveren. Wilks en Stevenson (Wilks & Stevenson, 1996) noemen het dan ook een “intermediate task”, die noodzakelijk is om de meeste NLP taken met succes te volbrengen. Brill & Mooney (Brill en Mooney, 1997) noemen de hoger liggende applicaties, zoals speech recognition, machine translation en information extraction ook wel high level taken binnen NLP, terwijl WSD een low level taak is, net als part-of-speech tagging.
WSD is vooral essentieel voor applicaties die dienen de taal te begrijpen, zoals message understanding en man-machine communicatie (Ide & Véronis, 1998). En soms biedt het ook assistentie bij applicaties die niet per se als taak hebben om taal te begrijpen, zoals de

hiervoor al genoemde machine translation en information retrieval en daarbij ook nog content analysis, grammatical analysis, speech synthesis en tekst processing (correctie van spelling, grammatica en stijl).
WSD is niet de enige subtaak die plaatsvindt binnen de NLP-applicaties. Een ander bekend voorbeeld is part-of-speech tagging, dat overigens ook gebruikt wordt binnen dit onderzoek. In tegenstelling tot WSD worden bij POS tagging al zeer goede resultaten behaald. Het WSD probleem is grotendeels onopgelost en heeft daarom nu een meer prominente plaats gekregen dan voorheen in onderzoek naar hoe verbeteringen binnen de NLP taken zijn door te voeren.

1.2 Onderzoeksveld

In de volgende paragrafen wordt een overzicht gegeven van de wijze waarop er in de loop der tijd onderzoek is gedaan naar de oplossing van het probleem van word sense disambiguation en hoe nieuwe technieken in de toekomst bij kunnen dragen aan nog betere resultaten.

1.2.1 Geschiedenis van het onderzoek

Veel onderzoek naar WSD aan het eind van de jaren '50 vertoont overeenkomsten met het huidige onderzoek. Toen al zochten onderzoekers in de aanwezigheid van directe context en de probabiliteit voor een bepaalde betekenis die dat met zich meebracht, een oplossing voor het WSD probleem (zie Madhu & Lytle, 1965). Weaver (Weaver, 1955) wees al op die mogelijkheid door een metafoor te geven voor wat wij later de windowgrootte van de context noemen: "But if one lengthens the slit in the opaque mask, until one can see not only the central word in question, but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word…" Jaren later bleek dat N niet groot hoefde te zijn; een contextgrootte van 2 of 3 woorden aan beide zijden van het centrale woord zou al zeer goede resultaten opleveren.
Halverwege de jaren '60 verminderde de interesse in een statistische vorm van taalonderzoek echter en veranderde de trend richting het onderzoek naar formele linguïstische regels, mede dankzij de grote populariteit van de transformationele theorieën van Noam Chomsky. Over het empirisch onderzoek dat gedaan werd naar WSD in deze periode van 10 à 15 jaar is eigenlijk niets meer terug te vinden, waaruit geconcludeerd kan worden dat wanneer er enig empirisch onderzoek is verricht, het waarschijnlijk weinig heeft opgeleverd.
Begin jaren '80 werd het onderzoek naar NLP minder gefocust op systemen die alle taken van NLP moesten beheersen en werd er weer meer gefocust op de subtaken van deze systemen. Daarnaast kwamen grote elektronische woordenboeken (dictionaries) op de markt, waaruit werd getracht om automatisch kennis te extraheren voor WSD-onderzoek. Hoewel deze machine readable dictionaries (MRD's) volgens Ide en Véronis (Ide & Véronis, 1998) de nodige tekortkomingen hadden, leverden zij uiteindelijk toch veel onderzoeksresultaten op. In de loop van de tachtiger jaren hebben grote corpora de positie van deze woordenboeken overgenomen als de grootste bron van informatie voor het disambigueren van woorden. De corpora bleken completer te zijn dan de verschillende dictionaries die in omloop waren; in corpora kwamen bijvoorbeeld relaties (co-occurences) tussen woorden voor die in woordenboeken nog over het hoofd werden gezien. Vooral nieuwe technieken zorgden voor meer mogelijkheden om de grote corpora te creëren en op te slaan die op dit moment nog steeds volop worden gebruikt bij het trainen en testen van WSD-systemen.

1.2.2 Lerende systemen als empirisch onderzoek binnen NLP

Het merendeel van het WSD-onderzoek, zoals dit heden ten dage uitgevoerd wordt is empirisch van aard. Al vanaf het begin dat empirische technieken gebruikt werden voor NLU en NLP, in de jaren vijftig, was duidelijk dat niet alleen lexicale en grammaticale informatie nodig was, maar ook semantische en pragmatische informatie, alsook algemene wereldkennis. Tot in de jaren '70 werden echter systemen ontwikkeld die zich alleen richtten op NLU van een beperkt domein (Brill & Mooney, 1997). Pas eind jaren '80 heeft het onderzoeksparadigma zich in feite weer verschoven richting de empirie. De focus verschoof van rationalistische methoden gebaseerd op regels naar empirische methoden gebaseerd op het verkrijgen van data uit grote corpora. Tegelijk ontstond de mogelijkheid om het menselijke handwerk te vervangen door automatisch lerende systemen.
Empirische methoden zelf bieden volgens Armstrong-Warwick (Armstrong-Warwick, 1993) vier oplossingen voor langdurige problemen binnen NLP, die ook zeer van toepassing zijn op het onderzoek naar WSD:

1. **Acquisition:** Veel kennis wordt direct uit data gewonnen zonder kostbaar handwerk;
2. **Coverage:** Een eventueel zeer grote trainingset kan uitgebreid onderzocht worden, wanneer gezocht wordt naar bepaalde verschijnselen;
3. **Robustness:** Het biedt een flexibele aanpak waarbij ruis gemakkelijker omzeild kan worden en er altijd tenminste een redelijk goede keuze gemaakt kan worden bij het classificeren;
4. **Extensibility:** Er kan gemakkelijk nieuwe data toegevoegd worden vanuit bijvoorbeeld andere domeinen om zo het systeem uit te breiden.

Daar boven op noemen Church en Mercer (Church & Mercer, 1993) drie ontwikkelingen die mede bijgedragen hebben tot de terugkeer van het empirische onderzoek, dus ook binnen onderzoek naar WSD:

1. Er is meer rekenkracht verkrijgbaar en computers zijn relatief goedkoper geworden;
2. Er is meer data verkrijgbaar in de vorm van grote corpora voor het trainen en testen van systemen;
3. Onderzoekers willen hun techniek meer gebruiken voor toepassingen in plaats van alleen voor onderzoek.

Empirische methoden zelf zijn op te delen in twee paradigma's: het onderzoek met behulp van probabilistische modellen, en onderzoek met behulp van symbolische lerende methoden (Brill & Mooney, 1997). Onder het laatstgenoemde paradigma valt het WSD-systeem dat in deze studie onderzocht wordt. Deze (symbolisch) lerende systemen die ontwikkeld worden kunnen omschreven worden als systemen die kennis die automatisch verkregen is kunnen onthouden of omzetten in regels; Op basis van de gelijkenissen met de instanties die het systeem al eerder heeft gezien of op basis van het toepassen van geleerde regels op nieuwe instanties kan het systeem vervolgens correcte keuzes maken.
Een voordeel van deze laatstgenoemde symbolische lerende methode is dat de verkregen kennis meer transparant is voor mensen dan grote matrices met probabiliteiten, waardoor meer inzicht wordt gegeven in het proces van NLU en ontwikkelaars het systeem gemakkelijker aan kunnen passen (Brill & Mooney, 1997).
Binnen de empirische methoden bestaan twee manieren om systemen te laten leren, gebaseerd op twee verschillende typen data: supervised leren en unsupervised leren. Het eerste maakt gebruik van door mensen geannoteerde tekst, het tweede van ruwe tekst zonder annotaties, dat beschouwd kan worden als een complexere methode. In dit onderzoek wordt gebruik gemaakt van supervised data. Voor supervised data geldt dat het moeilijk en tijdrovend is om te annoteren.

1.2.3 General vs. small scale

Binnen het onderzoek naar WSD zijn op dit moment twee paradigma's terug te vinden, namelijk de groep die streeft naar general WSD en de groep die WSD bedrijft door middel van small scale WSD (Wilks, 2000). Wilks & Stevenson (Wilks & Stevenson 1998) claimen dat binnen general WSD zo'n 95 % van de woorden in een volledige tekst correct geclassificeerd kunnen worden, en dat niet verwacht kan worden dat dit percentage nog zal stijgen: de overige woorden kunnen alleen behandeld worden middels nog niet bestaande, geavanceerde kunstmatige intelligentie, of lexicale word sense disambiguation, oftewel small scale WSD. Het argument om juist door te gaan met general WSD is dat dit ook de werkelijke taak is, namelijk de betekenis vinden van alle woorden binnen volledige teksten of andere vormen van uitingen.
Dit onderzoek is een voorbeeld van small scale of lexicale WSD. Er wordt alleen getest op de verschillende betekenissen van 30 woorden, niet op de betekenissen van de woorden binnen een volledige tekst. Vanwege het exploratieve karakter van dit onderzoek, waarbij de mogelijkheid onderzocht wordt om op een andere wijze een corpus te creëren, namelijk middels teksten afkomstig van het Internet, is het ook niet de bedoeling om gegevens te verzamelen over het disambigueren van een volledige tekst.

1.2.4 SENSEVAL

SENSEVAL is een open, internationale evaluatieve toets voor WSD-systemen en werd gestart in 1998. Bij de eerste editie deden 23 onderzoeksgroepen mee aan taken in het Engels, Frans of Italiaans (Kilgarriff & Palmer, 2000). Ieder deelnemend team krijgt voor een taal dezelfde standaardset van zinnen waarop getest kan worden. Op deze manier wordt er door iedereen gewerkt binnen hetzelfde kader, waardoor onderzoeksresultaten met elkaar vergeleken kunnen worden. De organisatie van SENSEVAL kan vergeleken worden met de TREC competities (http://trec.nist.gov/) binnen het onderzoeksgebied van Information Retrieval & Extraction.
Het uiteindelijke doel van deze samenwerking is een zo goed mogelijk WSD-systeem te ontwikkelen, dat geïmplementeerd kan worden in het geheel van NLP-taken. Doordat iedereen binnen SENSEVAL met dezelfde gegevens werkt is het mogelijk een consistente vergelijking op te bouwen om tot verbeteringen van de systemen te komen. Hierdoor helpt het de doelstellingen voor WSD-onderzoek verder te definiëren en verscherpt het de focus op bepaalde onderdelen van het WSD-onderzoek (Kilgarriff & Palmer, 2000).
Uiteraard doet een verscheidenheid aan systemen mee aan SENSEVAL. Aan het begin van SENSEVAL stonden twee groepen tegenover elkaar. Aan de ene kant stonden de voorstanders van supervised learning (zoals in dit onderzoek wordt toegepast), aan de andere kant de voorstanders van 'unsupervised' systemen. Beide groepen bleken echter veel ideeën van elkaar te hebben gebruikt om zo goed mogelijke resultaten te behalen (Kilgarriff & Palmer, 2000).
Naar aanleiding van SENSEVAL is getoond dat WSD als een losse subtaak van NLP gezien kan worden, maar er is nog niet bewezen dat een goed uitgevoerde WSD de overall performance van een NLP applicatie, zoals Information Retrieval of Machine Translation kan verbeteren (Kilgarrif & Palmer, 2000).
Het is binnen SENSEVAL mogelijk voor iedere deelnemer om een andere taal te kiezen dan de bovengenoemde drie talen. Bij de tweede SENSEVAL competitie deden dan ook al meerdere talen mee (zie Edmonds, 2002). Dit betekent echter wel dat een deelnemer die een andere taal onderzoekt zelf voor data moet zorgen als dat nog niet aanwezig is.
Het WSD-systeem waarvoor het corpus van Internetteksten dat opgebouwd is voor dit onderzoek als testmateriaal dient, is deelnemer binnen de Nederlandse taal aan SENSEVAL. Met toevoeging van het extra trainingsmateriaal in de vorm van het Internet-corpus zou verwacht kunnen worden dat het systeem beter gaat presteren in een test. Het ligt echter buiten de breedte van dit onderzoek om dit verder te bespreken.

1.3 Vraagstelling

Nu besproken is waar dit onderzoek geplaatst kan worden binnen het onderzoeksveld, is het tijd om te bespreken wat de vraagstelling van dit onderzoek inhoudt. Een aantal elementen hieruit zijn hiervoor al besproken, maar in deze paragraaf zal de vraagstelling uitgebreider aan bod komen.

1.3.1 Algemene onderzoeksvraag

Zoals hiervoor al is vermeld, is het een kostbare zaak om geannoteerde (supervised) data te verkrijgen voor onderzoek naar word sense disambiguation. Daarnaast bieden corpora vaak een eenzijdig beeld van het betekenisgebruik van bepaalde woorden, omdat de teksten afkomstig zijn uit één bepaalde bron, bijvoorbeeld een krant. Door deze werkwijze wordt er te weinig trainings- of testmateriaal verkregen voor minder frequent gebruikte betekenissen van een polyseem woord. Dit zorgt dan ook voor niet-optimale resultaten van WSD-systemen. Dit exploratieve onderzoek wil een verkenning uitvoeren naar de mogelijkheid om teksten afkomstig van het World Wide Web te gebruiken als data voor WSD-onderzoek.
Daarom luidt de algemene onderzoeksvraag:

Kan het Internet (webpagina's, forums) dienen als een taalkundig interessant "corpus" voor het doen van WSD-onderzoek?

1.3.2 Deelvragen

Om tot een antwoord op de algemene onderzoeksvraag te komen zullen een aantal zaken diepgaand onderzocht moeten worden. In dit onderzoek zal op een aantal punten een vergelijking gemaakt worden tussen het kinderboeken-corpus (bestaande uit tekst afkomstig uit kinderboeken) dat gebruikt is om het WSD-systeem te trainen, en het Internet-corpus waarop dit systeem getest wordt. Om dit gestructureerd te kunnen bespreken zijn een aantal deelvragen opgesteld.

De eerste deelvraag is:

- Wijkt de betekenisverdeling van polyseme woorden binnen het Internet-corpus af van de klassenverdeling voor de betekenissen binnen het kinderboeken-corpus?

De verschillende betekenissen voor een polyseem woord kunnen gezien worden als klassen waarbij binnen iedere klasse een bepaalde hoeveelheid voorbeelden van het gebruik van het polyseem woord in die betekenis hoort. Hoe evenwichtiger deze voorbeelden verspreid zijn over de diverse betekenissen, hoe hoger de entropie zal zijn voor een bepaald polyseem woord. Een grotere hoeveelheid verschillende betekenissen kan in combinatie met een gelijkmatige verdeling zorgen voor een nog hogere entropie. Entropie is "an estimation of the information chaos in the frequency distribution of the senses" (Hoste et al., 2002), oftewel vertaald naar het Nederlands: de hoeveelheid chaos in de verdeling van de betekenissen van een woord in de data.
Om het antwoord op deze eerste deelvraag te krijgen zal de entropie van de gekozen 30 woorden voor beide corpora uitgerekend moeten worden. Vervolgens zal nagegaan worden of de entropie van deze twee corpora significant correleert. Een verdere bespreking van het begrip 'entropie' en wat een onderzoek hiernaar mogelijk kan opleveren is te vinden in paragraaf 2.3.

Wanneer de entropie voor het Internet-corpus sterk afwijkt voor bepaalde woorden ten opzichte van het materiaal waarmee het WSD-systeem getraind is, namelijk het kinderboeken-corpus (zie verder paragraaf 3.3.1), zal dit problemen op kunnen leveren voor dit WSD-systeem. Dit zou kunnen betekenen dat de classificatieresultaten met het WSD-systeem tegenvallen ten opzichte van de resultaten met het kinderboeken-corpus.

De tweede deelvraag is:

- Krijgen de instanties van de polyseme woorden en hun context uit het verworven Internet-corpus, zoals ze teruggegeven zijn door een zoekmachine, vaker of minder vaak automatisch door een op het kinderboeken-corpus getraind WSD-systeem hun correcte betekenis toegewezen dan dat het geval is wanneer dit WSD-systeem getest wordt op een kinderboeken-corpus?

Voor elk van de 30 onderzochte polyseme woorden zal aan de hand van de instanties in het Internet-corpus door het WSD-systeem getest worden hoeveel procent van de testinstanties de correcte betekenis toegekend krijgen. Dit levert voor alle 30 polyseme woorden een score op die de accuraatheid van het WSD-systeem laat zien op de disambigueringstaak voor een bepaald woord. De resultaten op het kinderboeken-corpus zijn al bekend. Doordat de twee corpora zowel verschillen wat betreft bron als waarschijnlijk ook betekenisverdeling zal een vergelijking van de resultaten de nodige kennis op kunnen leveren.
Volgens Hoste, Daelemans, Hendrickx en van den Bosch (Hoste, et al., 2002) zal de accuraatheid toenemen als het aantal verschillende betekenissen afneemt. Voor de distributie van betekenissen en de complexiteit ervan (waarvan entropie een objectieve schatting is) geldt volgens hen hetzelfde, in een sterkere mate. Een lage entropie komt namelijk veelal voor met een hoge accuraatheid, terwijl een hoge entropie vaak leidt tot een lage accuraatheid. De eerste en de tweede deelvraag hebben daarom veel verband met elkaar.

1.3.3 Hypotheses

Voordat de deelvragen beantwoord zullen worden is het mogelijk alvast de verwachtingen te formuleren over de resultaten die behaald zullen worden binnen dit onderzoek.
Het WSD-systeem van ILK is getraind op kinderboeken en verwacht daarom bij verwerking van nieuwe data dezelfde data als die geleverd werden in het kinderboeken-corpus. Dit WSD-systeem wordt nu toegepast op de voor dit onderzoek opgebouwde database van zinnen afkomstig van het Internet. Verwacht wordt dat het systeem een minder hoge accuraatheid zal behalen op het materiaal afkomstig van het Internet dan op het kinderboeken-corpus. De verwachting is tevens dat het corpus met teksten van het Internet gaat afwijken in distributies van woordbetekenissen ten opzichte van het kinderboeken-corpus. De meest aannemelijke oorzaak hiervoor zal zijn dat er op het Internet minder sprake is van het woordgebruik dat voorkomt in kinderboeken, maar juist eerder het woordgebruik dat gericht is op volwassenen.
Rekening houdend met het bovenstaande zijn de hypotheses als volgt te formuleren:

Hypothese 1: De betekenisverdeling en de entropie van het Internet-corpus verschillen wezenlijk van die van het kinderboeken-corpus. Een mogelijke reden hiervoor is dat het Internet-corpus meer voorbeelden aanbiedt van verschillende betekenissen die niet voorkomen in het kinderboeken-corpus. De disambigueringstaak bij het kinderboeken-corpus zal gemakkelijker kunnen worden geacht dan bij het complexere Internet-corpus.

Hypothese 2: De polyseme woorden in het Internet-corpus zullen percentueel vaker een incorrecte betekenis toegewezen krijgen dan de polyseme woorden in het kinderboeken-corpus. Het WSD-systeem is tenslotte getraind met de data uit het kinderboeken-corpus en kan beschouwd worden als een betekenistoekenner met een beperkte vorm van

achtergrondinformatie van een basisschoolleerling. Het Internet-corpus zal meer woorden bevatten die meerdere verschillende betekenissen hebben, vaak metaforisch of gebruikt op een taalniveau dat meer 'volwassen' te noemen is. Het valt te verwachten dat het WSD-systeem vooral moeilijkheden zal hebben met de instanties van woorden in een figuurlijke betekenis die geboden worden door het Internet-corpus.

1.3.4 Doelen van het onderzoek

Het doel van dit onderzoek is in de eerste plaats om een klein corpus samen te stellen dat voor een aantal geselecteerde woorden 100 grammaticale zinnen bevat waarin de betreffende woorden voorkomen. Deze zinnen zijn afkomstig van het Internet en worden teruggegeven door een zoekmachine. Dit corpus wordt gebruikt binnen dit onderzoek, maar kan uiteraard ook dienen als aanvulling binnen verdere onderzoeken waarbij verschillende corpora bijvoorbeeld gecombineerd kunnen worden.
In de tweede plaats is het doel van dit onderzoek om een evaluatie te geven van de mogelijkheden die WSD-onderzoek door middel van een op Internet verworven corpus kan bieden. Gestreefd wordt naar aanbevelingen die onderzoekers in het hele veld, en in de eerste plaats diegenen die verbonden zijn aan de ILK onderzoeksgroep aan de Universiteit van Tilburg, in de toekomst in staat kunnen stellen om een oordeel te vellen over de stelling dat het Internet wel of niet geschikt is als bron om een corpus op te bouwen voor woordbetekenisonderzoek. De resultaten op de accuraatheid van toewijzing van de correcte betekenis aan een instantie binnen het Internet-corpus kan hier extra informatie bij verschaffen.
In de derde plaats kunnen de resultaten binnen dit onderzoek nieuwe wegen openen naar verbeteringen van WSD-systemen in het algemeen. De mogelijke problemen die het WSD-systeem tegen zal komen tijdens het testen op data afkomstig van het Internet kunnen hiervoor aanwijzingen geven.

1.4 Opbouw van de scriptie

In dit eerste hoofdstuk is een overzicht gegeven van wat het onderzoeksgebied van WSD precies inhoudt en wat de vraagstelling en doelstellingen van het hier besproken onderzoek zijn. In het verdere verloop van deze scriptie zal blijken dat er getracht wordt om zowel theoriëen aan te dragen als empirische toetsen uit te voeren. Hierdoor wordt gezocht naar een synthese tussen beiden, die uit zal monden in een antwoord op de hiervoor behandelde onderzoeksvragen.

In hoofdstuk 2 zal dieper ingegaan worden op wat precies 'senses' oftewel betekenissen zijn en de moeilijkheden die onderzoekers tegenkomen bij voornamelijk de eerste stap binnen het proces van word sense disambiguation. Er zal gezocht worden naar een mogelijke oplossing in de vorm van hiërarchieën. Verder wordt dieper ingegaan op de rollen die context en entropie in dit onderzoek zullen spelen.
In hoofdstuk 3 zal de onderzoeksmethode besproken worden, waarbij dieper ingegaan wordt op de dataverzameling en de handelingen die verricht zijn bij de opbouw van de twee corpora, en dan met name het Internet-corpus. Verder zal dieper worden ingegaan op het gebruikte WSD-systeem.
Vervolgens wordt in het vierde hoofdstuk een overzicht gegeven van de behaalde resultaten van het WSD-systeem en de entropiecijfers van de onderzochte polyseme woorden. Verder zal de focus liggen op het zoeken naar een verband tussen entropie en accuraatheid.
In hoofdstuk 5 zal een conclusie getrokken worden uit de resultaten en zal een antwoord gegeven worden op de onderzoeksvragen die in het eerste hoofdstuk werden geïntroduceerd.
Tot slot zal in het zesde en laatste hoofdstuk een discussie gegeven worden over de invloed van de onderzoeksmethode op de resultaten van dit onderzoek. Verder wordt er

ruimschoots aandacht gegeven aan de mogelijkheden en onmogelijkheden van het gebruik van tekstfragmenten afkomstig van het Internet voor onderzoek naar woordbetekenissen. Ook een mogelijke toepassing van hiërarchieën wordt verder behandeld in dit laatste hoofdstuk.

2. Senses en hiërarchieën

In dit hoofdstuk wordt meer theoretische achtergrond besproken van dit onderzoek. In de eerste paragrafen van dit hoofdstuk wordt ingegaan op een aantal onderwerpen uit de theoretische onderbouwing van dit onderzoek, namelijk wat betekenissen precies zijn en welke problemen er te verwachten zijn bij het kiezen van betekenissen tijdens de eerste stap van WSD-onderzoek.
Vervolgens wordt dieper ingegaan op het begrip entropie en zal verder uitgelegd worden welke invloed dit kan hebben op de prestaties van een WSD-systeem. Dit kan tevens beschouwd worden als een inleiding op de eerste paragraaf van hoofdstuk 4, waarin de entropiecijfers voor de woorden binnen dit onderzoek besproken zullen worden.
Daarna zal de rol van context beschreven worden binnen de werkwijze van een WSD-systeem. Tot slot zal een schets gegeven worden van de mogelijkheden van het Internet als corpus voor WSD-onderzoek.

2.1 Senses

Wat zijn betekenissen of 'senses' nu precies? Hierover wordt al sinds mensenheugenis gediscussieerd in allerlei onderzoeksgemeenschappen. Vanaf de tijd van Aristoteles is ook het nodige hierover op schrift gesteld. Ook in het onderzoeksgebied van WSD wordt hierover gediscussieerd. Er is volgens Ide en Véronis (Ide & Véronis, 1998) geen algemene definitie te geven voor het woord 'betekenis'.
Hoewel deze scriptie niet de plaats is om een weergave te geven van deze discussie is het wel van toepassing om aan te geven dat er twee opvattingen zijn over wat een betekenis precies is. De ene opvatting zegt dat een betekenis een losse entiteit is, waarvan de inhoud uit zichzelf bepaald is. De andere opvatting meent dat betekenis alleen kan ontstaan in interactie met een context, waarbij het gebruik van een woord de betekenis ervan bepaald. De filosoof Wittgenstein wijst op de aanwezigheid van die tweede opvatting: "Don't look for the meaning, but for the use" (Wittgenstein, 1953).
Het ziet er naar uit dat een betekenis in het geval van WSD-onderzoek een definitie toegekend zal moeten worden die beide hiervoor besproken opvattingen zal moeten bevatten. Tenslotte bepaalt in het geval van polyseme woorden zowel het gebruik als de inhoud die het woord als losse entiteit heeft uiteindelijk de betekenis van het woord. Voor monoseme woorden kan dan gesteld worden dat het gebruik van het woord niet van invloed hoeft te zijn op de betekenis.

2.1.1 Het kiezen van senses

Kilgariff en Palmer (Kilgarriff & Palmer, 2000) noemen een groot probleem bij het doen van WSD-onderzoek het vaststellen en benoemen van de verschillende betekenissen of senses die een polyseem woord kan hebben. Dit is de eerste van de twee stappen van word sense disambiguation, namelijk sense discrimination, en wordt door veel onderzoekers als de belangrijkste stap beschouwd. En juist deze stap brengt de meeste moeilijkheden met zich mee. De kans dat sense discrimination resulteert in één groot puzzelwerk is dan ook groot. Kilgarriff (Kilgarriff, 1997) stelt dan ook dat: "To know what a word sense S1 is, is to know which uses of the word are part of S1 and which are not (…). If we are to know what word senses are, we need operational criteria for distinguishing them."
In de komende paragrafen komen vier grote problemen bij het vaststellen van betekenissen aan de orde. Uiteindelijk wordt naar een mogelijke oplossing toe gewerkt, in de vorm van hiërarchieën waarbij betekenissen in een boomstructuur geplaatst worden.

2.1.2 Probleem 1: Verschillen tussen woordenboeken

In veel onderzoeken wordt gebruik gemaakt van een betekenisindeling die gegeven wordt door (elektronische) woordenboeken of thesauri, zoals WordNet (zie Chodorow, Leacock & Miller, 2000), Roget, Hector (zie Ellman, Klincke & Taite 2000) en LDOCE (zie Wilks & Stevenson, 1997a en Wilks & Stevenson, 1997b). Wanneer er gebruik wordt gemaakt van de betekenisindeling die aangegeven wordt in woordenboeken, wordt het vaststellen van betekenissen in de eerste plaats bemoeilijkt door het feit dat niet vast staat welk woordenboek nu unaniem gebruikt kan worden als bron die de woordbetekenissen aan kan leveren.
Uiteindelijk zullen hier goede afspraken over gemaakt moeten worden, aangezien er veel verschillen zijn tussen de verschillende woordenboeken in een taal, waarbij de één meer betekenissen geeft dan de ander en ook de kwaliteit van de vastgestelde betekenissen kan verschillen. Daarnaast is het zo dat bij het indelen van betekenissen in woordenboeken keuzes worden gemaakt door de lexicografen, terwijl alternatieve keuzes zeker zo goed zouden zijn (Kilgarriff, 1997).
Sommige onderzoekers noemen woordenboeken overigens zelfs verre van geschikt als uitgangspunt voor betekenisverdelingen. Kilgarriff (Kilgarriff 1993) meent dat de betekenissen die gevonden worden in een groot corpus niet worden gedekt door de betekenissen die gegeven worden in een woordenboek. Ook Wilks (Wilks, 1997) heeft kritiek op het gebruik van woordenboeken. Volgens hem kunnen 20 % van de betekenissen in teksten niet verbonden worden met een vastomlijnde betekenis uit een bepaald woordenboek.
Volgens Véronis (Véronis, 2000) zijn traditionele woordenboeken alleen gericht op het defninieren van betekenissen van woorden, maar niet op de toepassing van die betekenissen op bijvoorbeeld een syntactische wijze of binnen collocaties. Ze geven niet die bepaalde ‘clues’ aan die gebruikt kunnen worden door een WSD-systeem en dit alles maakt ze niet geschikt als bron voor betekenissen voor een WSD-systeem. Hetzelfde geldt volgens hem ook voor WordNet.

2.1.3 Probleem 2: Homonymie en polysemie

In de tweede plaats zorgt het verschil tussen homoniemen en polyseme woorden voor problemen bij het vaststellen van betekenissen. Bij homonymie is er sprake van min of meer twee verschillende "woorden" die toevallig hetzelfde geschreven worden. Bij polysemie heeft een enkel woord meerdere betekenissen. Mensen hebben geen problemen om de betekenissen van homoniemen uit elkaar te houden, met polyseme woorden hebben ze meer moeite, vooral als ze in een abstracte betekenis gebruikt worden of in een bepaalde context.
De verschillen in betekenissen van polyseme woorden zijn een stuk moeilijker aan te duiden dan bij homoniemen en zijn vaak zelfs onvoorspelbaar: een woord kan in een bepaalde context een betekenis hebben die zelden gebruikt wordt en niet terug te vinden is in een woordenboek, terwijl andere woorden in die context wel vaak gebruikt worden en in sommige woordenboeken wel gelexicaliseerd zijn.
Cruse (Cruse, 1986) maakt een onderscheid tussen instanties waarin de context een bepaalde betekenis selecteert en instanties waar de context de betekenis moduleert. In het eerste geval is er sprake van twee verschillende betekenissen (homonymie), terwijl in het tweede geval de betekenissen duidelijk onderling overeenkomsten hebben (polysemie). Bij de volgende twee naar het Nederlands vertaalde zinnen maakt hij duidelijk dat de context één van de verschillende betekenissen selecteert:

> "Heb je het geld al naar de bank gebracht?"
> "De oude man zat op de bank in het park."

In de volgende zinnen worden verschillende aspecten van een betekenis gemoduleerd door de context:

"Jan verft zijn fiets rood." (frame)
"Jan maakt zijn fiets droog voordat hij wegrijdt in de regen." (zadel)
"Jan's fiets werd gisteren gestolen." (geheel)

Een computer zal bij het automatisch disambigueren tussen deze verschillende typen betekenissen - ook al zijn ze erg zeldzaam en worden ze maar één keer gebruikt in een enorm corpus – toch iets relevants moeten doen.

2.1.4 Probleem 3: Granularity

Het derde probleem bij het bepalen van de betekenissen die een polyseem woord kan hebben is het vaststellen van de graad van fijnheid, in het Engels granularity, waarin de betekenissen onderverdeeld worden. Dit hangt samen met het hiervoor genoemde probleem van het verschil tussen homonymie en polysemie en is met name gerelateerd aan deze tweede categorie.
Verschillende auteurs (o.a. Véronis, 2000) hebben opgemerkt dat de betekenisverdeling die te vinden is in woordenboeken vaak te fijn is voor NLP taken. Een grote hoeveelheid betekenissen met onderling kleine verschillen kan namelijk voor een explosie van mogelijke betekenissen leiden als er meerdere polyseme woorden in één zin staan (Ide & Véronis, 1998).
Wilks en Stevenson (Wilks & Stevenson, 1997b) menen dat het geven van specifieke betekenissen voor bijvoorbeeld het woord ‘fiets’ (zie de vorige paragraaf) absurd is. Volgens hen zou je dat ook kunnen doen voor de andere 250 onderdelen van een fiets! Dit heeft volgens hen dan ook meer te maken met knowledge processing, maar niets met de taak van WSD. Volgens hen houdt de indeling van betekenissen dan ook op bij de indeling die gebruikt wordt in het in hun onderzoek toegepaste woordenboek.
Het onderscheid tussen de fijne betekenissen is vaak moeilijk te zien, zelfs voor ervaren menselijke taggers als lexicografen (zie paragraaf 6.1.1). Ook is een zeer grote dataset nodig binnen een supervised methode om een WSD-systeem al deze betekenissen te kunnen leren. Een gemiddelde menselijke lezer van teksten kan de fijne verschillen tussen die betekenissen vaak zelf ook niet onderscheiden.

En toch zal het handig zijn wanneer een WSD-systeem sommige genuanceerde betekenissen voor een polyseem woord kan onderscheiden. Binnen de dataset van de 30 polyseme woorden die in dit onderzoek onderzocht werden, zijn een aantal van dit soort gevallen terug te vinden. Zo kunnen bij het woord “staart” de volgende zeer verwante maar toch verschillende betekenissen <staart_dier> en <staart_mensenhaar> regelmatig teruggevonden worden binnen het corpus. Een verdere bespreking van de resultaten zal volgen in hoofdstuk 4.
Kunnen deze zeer fijne betekenissen samengevoegd worden? Met het combineren van betekenissen die in een woordenboek gegeven worden kom je niet ver. Deze zijn namelijk in de meeste gevallen niet gedetailleerd genoeg of juist veel te gedetailleerd als WSD gezien wordt als subtaak binnen een grotere NLP taak.
Tevens staan er in elk woordenboek wel woorden die gewoonweg niet goed uitgewerkt zijn, omdat er in grote hoeveelheden tekst gewoonweg door menselijke taggers meer betekenissen onderscheiden kunnen worden dan dat er in een woordenboek staan, waarbij sommige woorden duidelijk meer gebruikt blijken te worden in de realiteit dan lexicografen bij het samenstellen van het woordenboek hadden ingeschat en die bij hen dus buiten de boot waren gevallen.
De oplossing die ruim aan de voorhand van de bespreking van het empirische gedeelte van dit onderzoek gegeven wordt is daarom deze: laat menselijke taggers zo gedetailleerd mogelijk taggen, waarbij ze ruim de gelegenheid krijgen om subtiele verschillen in de betekenis aan te wijzen. Bouw vervolgens uit de testervaring die door een WSD-systeem is verkregen middels dit supervised corpus voort, door voor polyseme woorden hiërarchieën

samen te stellen, met daarin de mogelijkheid om fijne betekenissen te plaatsen binnen een klein deel van de hiërarchische boom. Over deze hiërarchieën gaat paragraaf 2.2.

2.1.5 Probleem 4: Data-sparseness probleem

Het vierde probleem bij het vaststellen van betekenissen hangt niet zozeer samen met betekenissen zelf, maar is eerder een achterliggend probleem dat hierop betrekking heeft. Naast het feit dat het moeilijk is om handmatig een groot trainingscorpus te taggen is het vaak ook zeer tijdrovend en daardoor ook zeer duur. Dit is dan ook de reden waarom er nog niet veel handmatig getagde corpora verkrijgbaar zijn (Ide & Véronis, 1998).
Het gevolg hiervan is dat er te weinig trainingsmateriaal voor een grote hoeveelheid betekenissen aanwezig is om een systeem voldoende te kunnen trainen. Dit wordt ook wel het data-sparseness probleem genoemd. Enorme hoeveelheden getagde tekst blijken nodig te zijn om alle betekenissen van een polyseem woord te representeren, vooral ook omdat sommige betekenissen erg zeldzaam zijn. Hoewel op dit moment een kleine hoeveelheid kleine tot redelijk grote corpora voor alle onderzoekers voorhanden is, is het gewenst om dit probleem zo spoedig mogelijk op te lossen.

Voor dit data-sparseness probleem bieden similarity-based methods op dit moment de beste oplossing (Ide & Véronis, 1998). Bij similarity methods worden woorden die veel met elkaar te maken hebben (zoals synoniemen) aan elkaar gekoppeld, zodat tijdens het disambigueren niet per se een verband hoeft te zijn tussen focuswoord en een specifiek woord in de context, maar ook met een daaraan gekoppeld woord dat gelijkenis vertoond wat betreft de betekenis met het contextwoord, en die dus wel voorkomt in de training set. Dat gekoppelde woord kan dus vervolgens gebruikt worden om het focuswoord te disambigueren. Karov & Edelman (Karov & Edelman, 1998) vermelden met deze methode een resultaat van 92 % correcte WSD, zonder dat een enorm corpus nodig was. Deze methode lijkt dus duidelijk (één van) de methode(s) voor de toekomst.

2.2 Hiërarchieën als mogelijke oplossing

Bij een hiërarchische sense inventory, ook wel kortweg een hiërarchie of een hiërarchische betekenislijst genoemd, worden betekenissen die door de context geselecteerd worden (zie vorige paragraaf) bovenin in een knoop geplaatst, terwijl betekenissen die door de context worden gemoduleerd dieper in de hiërarchie in één (van de) tak(ken) geplaatst worden.
Op deze manier ontstaat er een omgekeerde boomstructuur, waarbij betekenissen die veel van elkaar verschillen bovenin in het smalle gedeelte op de eerste niveaus terug te vinden zijn, terwijl (sub-) betekenissen die onderling veel gemeen hebben, maar weinig overeenkomsten hebben met andere betekenissen, in dezelfde tak in één van de onderste niveaus in het brede gedeelte van de boom terug te vinden zijn.

2.2.1 Voordelen van hiërarchieën

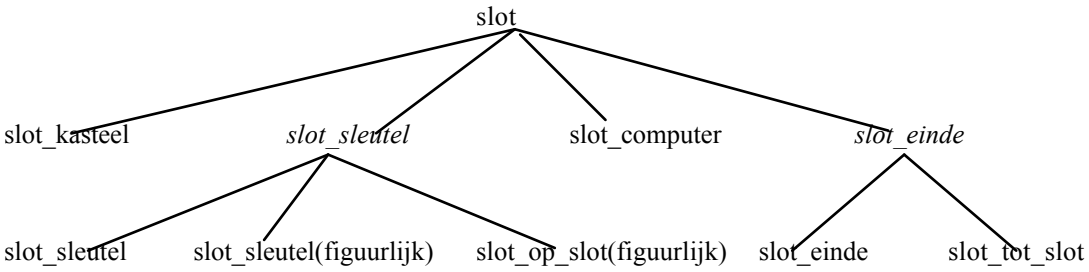
Een hiërarchie van betekenissen biedt een aantal voordelen, die verderop in dit onderzoek aan te blijken sluiten bij de ontwikkelingen die het gebruik van instanties afkomstig van het Web met zich meebrengt.
Het eerste voordeel van zo'n hiërarchie is dat bij het maken van een keuze voor een bepaalde betekenis een WSD-systeem niet per se de precieze subbetekenis hoeft te raden, maar bijvoorbeeld een betekenis die hoger te vinden is in een bepaalde tak van de hiërarchie, om betekenisgerichte problemen binnen bepaalde NLP-applicaties al te voorkomen. Het maakt de kans op een goed gefundeerde keuze in feite groter bij sommige applicaties waarin de eisen voor WSD niet op het hoogste niveau zijn.
Het tweede voordeel is dat voor zeldzame betekenissen waar zeer moeilijk trainingsmateriaal voor te vinden is toch nog een potentiële plaats aan te wijzen is binnen een hiërarchie. Hierdoor vallen zeldzame betekenissen niet per definitie buiten de boot; er hoeft minder vaak een noodmaatregel aangesproken te worden, die bijvoorbeeld voor zeldzame betekenissen de meest voorkomende betekenis kiest bij gebrek aan voorbeelden in het geheugen.
Het derde voordeel is dat een goede prestatie van een WSD-systeem op de lange termijn ook gegarandeerd is, doordat het mogelijk is nieuw gevormde betekenissen die door taalvernieuwing ons (schriftelijk) taalgebruik binnen zijn gedrongen ook een goede plaats te geven binnen een hiërarchie.

2.2.2 Een nadeel van hiërarchieën

Een groot nadeel van hiërarchieën komt tegelijkertijd met het eerste grote voordeel ervan. Tussen de betekenissen die twee eindpunten vormen die ver uit elkaar liggen in de omgekeerde boomstructuur kunnen toch verbanden te trekken zijn. Dit maakt het opbouwen van zo'n omgekeerde boomstructuur tot een ingewikkelde taak. Een besluit op een hoger niveau tijdens het keuzetraject kan namelijk leiden tot een onmogelijkheid om voor een bepaalde betekenis te kiezen die in een bepaald geval wel van toepassing was.
Een voorbeeld hiervan is het Engelse woord 'orange' dat twee betekenissen heeft die ver uit elkaar liggen in een hiërarchische boomstructuur, namelijk <orange_vrucht> aan de ene kant en <orange_kleur> aan de andere kant van de boomstructuur, waarbij <orange_kleur> geen eindbetekenis is maar nog ingedeeld kan worden in twee sub-betetekenissen, namelijk <orange_kleur_zichtbaar> en <orange_kleur_abstract>. Hoewel deze betekenissen zichtbaar gescheiden zijn in de hiërarchie liggen ze intuïtief dicht bij elkaar. De naam van de vrucht is tenslotte verbonden met de kleur en andersom.
Doordat gebruik wordt gemaakt van een betekenishiërarchie zal waarschijnlijk gekozen worden voor één van de twee betekenissen terwijl gevallen aan te wijzen zijn waarin de andere betekenis ook van toepassing was geweest. Een voorbeeld van zo'n geval is de zin: "The cyclist's head turned orange when he climbed the steep hill".
Voor dit nadeel bieden hiërarchieën geen oplossing. In dit geval valt het te verwachten dat een associatief netwerk beter werkt. Wanneer binnen een hiërarchie een vertakking plaatsvindt, dan komen deze vertakkingen nergens meer samen, wat een zekere beperking oplevert. Bij een associatief netwerk kunnen vertakkingen wel weer samengevoegd worden en is het mogelijk om probleemgevallen zoals het hiervoor genoemde voorbeeld correct te disambigueren.

2.2.3 Voorbeeld van een hiërarchie

In deze paragraaf zal een voorbeeld gegeven worden van de toepassing van een hiërarchische betekenislijst op één van de 30 polyseme woorden waarop onderzoek is uitgevoerd, namelijk het woord 'slot', dat beschikt over een voor het Internet-corpus gemiddeld aantal van 7 betekenissen, waarvan sommige dichter bij elkaar liggen dan andere.
Figuur 1 laat zien dat sommige betekenissen op een lager niveau ingedeeld kunnen worden in een tak van de omgekeerde boomstructuur. Andere indelingen binnen een dergelijke boomstructuur zijn ook mogelijk. Zo kunnen voor een aantal polyseme woorden de betekenissen ingedeeld worden in de categorieën "letterlijk", "figuurlijk" en eventueel "meta", om hiermee op het eerste niveau al aan te geven of een woord letterlijk gebruikt wordt of figuurlijk. Met een figuurlijk gebruik van een woord hebben mensen al meer moeite om er een betekenis aan toe te kennen, dus verwacht kan worden dat dit voor een computer al zeker het geval zal zijn.



Figuur 1: Een mogelijke betekenishiërarchie van het polyseme woord "slot"

2.2.4 Verdere toepassingen van hiërarchieën

Wanneer de structuur van een hiërarchische inventory wordt overgenomen in een WSD-systeem, bijvoorbeeld in een decision list, kan nog een stap verder worden gegaan dan alleen een hiërarchische opsomming geven van alle betekenissen. Yarowski (Yarowski, 2000) schrijft dat een belangrijk criterium in een hiërarchische (betekenis) lijst, zoals een tag die aangeeft tot welke syntactische categorie een woord behoort, oftewel een part-of-speech-tag, al kan bepalen in welk onafhankelijk pad van de boomstructuur de betekenis gezocht kan worden: "we would wish to divide the control flow of the decision procedure into relatively independent paths specialized for the modelling needs of each side of the splitting partition".
Juist de POS-tag is zeer geschikt om boven in de hiërarchische boom een vertakking te vormen. Op het tweede niveau kunnen de verschillende inflexievormen als vertakkingen gebruikt worden en op het derde en laagste niveau kan vervolgens ruimte vrij gemaakt worden voor diverse zegswijzen en uitdrukkingen waarin het focuswoord voorkomt. Een groot deel van zijn ideeën vallen echter buiten het bereik van dit onderzoek. Eén goed punt heeft hij echter aangestipt dat ook zeer bruikbaar is in dit onderzoek, namelijk de mogelijkheid om collocaties te betrekken in een hiërarchische inventory.
Door collocaties op een laag niveau in de hiërarchie te betrekken wordt op meerdere fronten voordeel behaald: Niet alleen het focuswoord zelf maar ook de andere woorden binnen de uitdrukking of collocatie worden direct van de goede betekenis voorzien. Mocht het classificeren van de precieze collocatie niet slagen, dan kan de classificatie het polyseme woord mogelijk nog een betekenis toewijzen die zich bevindt in de goede tak van de hiërarchie. Een voorbeeld van zo'n collocatie is 'tot slot', dat ook verwerkt is in de hiërarchie in figuur 1.

2.3 Entropie

In deze paragraaf zal meer informatie gegeven worden over het begrip entropie, dat geïntroduceerd werd in de eerste deelvraag van dit onderzoek. Zoals daar al in het kort werd aangegeven wordt entropie hier beschouwd als de hoeveelheid informatiechaos in de verdeling van de betekenissen van een polyseem woord. De binnen dit onderzoek gehanteerde formule voor entropie is:

$$H(D) = -\sum_i p_i \log_2 p_i$$

De entropie voor een betekenisverdeling is hoog wanneer er sprake is van een gelijkmatige verdeling. Wanneer er sprake is van twee betekenissen voor een polyseem woord die allebei in 50 % van de gevallen voorkomen in het corpus dan is er sprake van de maximale entropie van 1.0. Wanneer één van de betekenissen in 95 % van de gevallen voorkomt dan is er sprake van een lage entropie.
Bij een aanwezigheid van meer dan 2 verschillende betekenissen kan de entropie uitgedrukt worden door een getal groter dan 1, zoals straks in de resultaten in paragraaf 4.1 te zien zal zijn. Wanneer bij meer dan 2 betekenissen het entropiecijfer net onder 1 is wil dat nog niet betekenen dat er sprake is van een bijna maximale chaos.
Voor een groot aantal woorden in het kinderboeken-corpus geldt dat de verdeling van de betekenissen ruwweg vergeleken kan worden met een Zipf-curve (Zipf, 1935): veel betekenissen komen amper voor, terwijl een aantal betekenissen zeer vaak voorkomen. Een voorbeeld daarvan wordt gegeven door Chodorow (Chodorow et al., 2000) voor het Engelse woord “bank”. Van elke 100 instanties kwamen in hun corpus er 78 voor met de betekenis <bank_financial_institution>, terwijl de overige 22 instanties 8 andere betekenissen representeerden. Er is hier wel degelijk sprake van meerdere verschillende betekenissen, maar toch een vrij lage entropie.

Hoe hoger de entropie zal zijn, hoe moeilijker het zal zijn voor een WSD-systeem om een goede accuraatheid te behalen tijdens een WSD-test. Melamed (Melamed, 1997) wijst hier ook op: “...semantic entropy can be interpreted as semantic ambiguity. On this interpretation, it can predict the difficulty of disambiguating the sense of a given word.”
Hoewel de hoeveelheid polysemie, die in (Hoste et al., 2002) wordt omschreven als “the number of senses of a word-POS combination” in principe ook aan kan tonen hoe moeilijk de WSD-taak zal worden, is entropie volgens Kilgariff en Rosenzweig (Kilgarriff & Rosenzweig, 2000) een betere maatstaf. Zij kwamen tot deze conclusie door een grote hoeveelheid polyseme woorden op te splitsen in hun syntactische categorie. Het bleek dat Nouns gemiddeld een hogere polysemie hadden, maar dat Verbs een hogere entropie bezaten. Omdat het WSD-systeem meer moeite had met het voorspellen van de betekenis van Verbs, was volgens hen entropie een betere voorspeller van de moeilijkheid van een WSD-taak.
Overigens geldt voor dit onderzoek ook het vermoeden dat van de 30 hier onderzochte polyseme woorden Verbs lager zullen scoren op accuracy dan Nouns. Een groot aantal woorden binnen de hier onderzochte set hebben overigens zowel een betekenis als Noun als ook een betekenis als Verb. Het maken van onderscheid tussen syntactische categorieën is echter impliciet als taak meegenomen door het onderzochte WSD-systeem (zie paragraaf 3.4.3). Met behulp van een andere set woorden waarbij duidelijk sprake is van polyseme woorden die alleen in de vorm van één bepaalde syntactische categorie voorkomen kan dit verder onderzocht worden.

2.4 Context

Het analyseren van de context is volgens Ide en Véronis (Ide & Véronis, 1998) de enige manier om de betekenis van een polyseem woord te achterhalen. Dit is volgens hen dan ook de reden dat bijna al het onderzoek naar WSD gebaseerd is op de informatie die de context rond het focuswoord aanlevert voor de disambiguering.
Vooral voor methoden die gebruik maken van veel data, zoals in dit onderzoek het geval is, is context belangrijk. Context voorziet het lerende systeem namelijk van trainingsmateriaal in de vorm van de tekstuele omgeving van een woord om de huidige context mee te vergelijken om een keuze te maken voor een bepaalde betekenis.
Context wordt volgens het hiervoor geciteerde artikel van Ide en Véronis op twee manieren gebruikt:

- Bag-of-words approach: context wordt beschouwd als de woorden in een bepaald window, zonder dat ze verder worden behandeld als hebbende een bepaalde positionele syntactische relatie met het focuswoord;
- Relational information: context in een bepaald window wordt wel behandeld als hebbende een bepaalde dieper liggende syntactische of positionele relatie met het focuswoord.

In dit onderzoek wordt context alleen behandeld door middel van een relationeel-positionele aanpak. Er wordt hier dus in de eerste plaats gekeken naar een klein window rond het focuswoord om de betekenis aan de hand daarvan te voorspellen. Deze kleine hoeveelheid context wordt ook wel microcontext genoemd, en bevindt zich dus het liefst in de naaste omgeving van het focuswoord. Deze aanpak is goedkoper dan andere wijzen (vaak een combinatie met andere technieken) om een WSD-systeem op te bouwen, vraagt tevens om minder ingewikkelde handelingen en kan dezelfde goede resultaten behalen binnen sommige NLP-applicaties.
Om het in dit onderzoek gebruikte WSD-systeem te voorzien van context worden in het corpus de instanties met de volledige zin waarin ze voorkomen, plus de voorgaande en de daaropvolgende zin, in het corpus geplaatst. Uiteindelijk zal het WSD-systeem echter alleen gebruik maken van de context van 4 woorden rondom het focuswoord, waarbij 2 woorden links van het focuswoord en 2 woorden rechts van het focuswoord staan. Zo'n raamwerk van 4 woorden lijkt klein, maar geeft al een goed, zo niet het beste resultaat (Veenstra, van den Bosch, Buchholz, Daelemans & Zavrel, 2000).
Over de ideale grootte van de window is echter nog steeds discussie tussen WSD-onderzoekers onderling. Sommigen maken gebruik van een windowgrootte van 6, met 3 contextwoorden aan de linkerkant en 3 contextwoorden aan de rechterkant. In dit onderzoek is echter gekozen voor de windowgrootte die zeer vaak, zo niet het meest voorkomt binnen de onderzoeken naar WSD. Op deze manier kan er ook een grotere vergelijkingsbasis gegeven worden met andere onderzoeken in het veld. Daarnaast is dezelfde windowgrootte ook toegepast op de WSD-test met het kinderboeken-corpus, waarbij geconcludeerd werd dat deze windowgrootte bij dat corpus het beste resultaat behaalde (zie Hendrickx et al., 2002).

2.5 Het Internet als corpus

Het World Wide Web, het deel van het Internet waarin communicatie verloopt via het http-protocol en voornamelijk voor informatieve (non-profit en commerciële) doeleinden wordt gebruikt, geeft vrij toegang tot kolossale hoeveelheden tekst in allerlei vormen en talen (Kilgarriff, 2001). Het tekstuele materiaal is terug te vinden op webpagina's, die naast tekst ook bestaan uit afbeeldingen en in de broncode voorzien zijn van HTML-tags. De teksten gaan over allerlei uiteenlopende onderwerpen en bevatten allerlei typen 'taal'. Dit komt doordat de auteurs van tekstueel materiaal op het Internet zeer uiteenlopend zijn, en niet alleen journalisten of andersoortige professionele schrijvers zijn, zoals bijvoorbeeld in een corpus bestaande uit krantenartikelen, of zoals in dit geval kinderboeken.
Als andere corpora, zoals kranten-corpora, de basis zijn waarin de diverse betekenissen van een polyseem woord ontdekt kunnen worden, zullen over het algemeen minder snel alle mogelijke en onvoorspelbare betekenissen gevonden kunnen worden dan in een corpus gebaseerd op tekst afkomstig van het Web. Het Web als grootst mogelijke corpus en een bron die up-to-date blijft kan het meest in staat worden geacht om zeldzame maar ook gloednieuwe betekenissen van polyseme woorden in teksten over de meest uiteenlopende topics in zichzelf terug te vinden. Web crawling wordt ook door diverse auteurs expliciet genoemd als een middel om corpora te vergroten (zie onder andere Leacock, Chodorow & Miller, 1998).
Er zijn ook nadelen verbonden aan het Internet als corpus. Om een zo groot mogelijk corpus te hebben zullen de teksten gebruikt worden in de vorm waarin ze online staan. Het Web verandert echter constant en bevat duplikaten en gedeeltelijk duplicaten (Kilgarriff, 2001). Daarnaast moet de taal op zo'n pagina nog worden vastgesteld (soms is er sprake van een mengsel van verschillende talen). Vervolgens zouden eigenlijk alle teksten geclassificeerd moeten worden (chat-tekst, wetenschappelijk artikel enzovoorts). Er verdwijnen ook regelmatig webpagina's. Het laatstgenoemde nadeel kan soms ook als voordeel gezien worden, doordat hopeloos verouderde content verwijderd wordt, waardoor een tijdsgebonden betekenis van een polyseem woord verdwijnt. Volgens het hiervoor geciteerde artikel van Kilgarriff moeten veel te veel parameters gesteld en handelingen verricht worden op een verzameling teksten van het Web om er onderzoek op uit te voeren. Om deze redenen zou het Internet niet geschikt zijn als onderzoeksomgeving. Verdere nadelen van het gebruik van Internetteksten voor het opbouwen van een corpus komen terug in de discussie (zie hoofdstuk 6), waarin de praktijkervaringen met het verzamelen van instanties voor dit onderzoek uitvoerig zullen worden beschreven.
Maar als we rekening houden met het feit dat juist de anarchie van het Internet, samen met de onmetelijke grootte ervan, beschouwd kunnen worden als de belangrijkste kenmerken van dit Web, dan kan deze informatiebron ook als zeer nuttig beschouwd worden voor het creëren van onderzoeksdata. In ieder geval zijn de bovengenoemde problemen van het Internet omzeild door de wijze waarop in dit onderzoek de steekproef van Internetteksten is opgebouwd. De tekst in het corpus dat gecreëerd is voor dit onderzoek is tenslotte handmatig opgezocht aan de hand van strenge criteria en wordt niet online, maar vanuit een offline bestand geopend, waarin het ook is opgeslagen. Hierdoor blijft de inhoud van het corpus vaststaan en gaan geen teksten verloren doordat pagina's offline gaan. Om uiteindelijk gebruik te maken van de onmetelijke grootte van het Web, zal echter gebruik gemaakt moeten worden van online content.

3. Methode

In dit hoofdstuk zal de onderzoeksopzet besproken worden. Als eerste komt het proces van dataverzameling aan de orde, waarbij onder andere de criteria besproken worden waaraan een goede instantie van een polyseem woord moet voldoen om in het corpus opgenomen te worden. Vervolgens zal besproken worden hoe dit corpus opgebouwd is. Daarna worden de kenmerken besproken van de twee corpora die binnen dit onderzoek vergeleken worden, namelijk het kinderboeken-corpus en het Internet-corpus. Tot slot wordt ingegaan op het gebruikte materiaal binnen dit onderzoek, waaronder de zoekmachine Google.nl en de WSD-tagger MBWSD-D.

3.1 Dataverzameling

Een essentieel onderdeel van dit onderzoek was het verzamelen van een redelijke hoeveelheid instanties (voorkomens inclusief hun context) van de 30 onderzochte polyseme woorden. In de volgende subparagrafen zullen de werkwijze voor dit verzamelen besproken worden, alsook de criteria waaraan een goede instantie moet voldoen.

3.1.1 Werkwijze voor het verzamelen van instanties en context

Als zoekmachine is gebruikt www.google.nl (zie voor bespreking van deze zoekmachine paragraaf 3.4.1). Binnen Google werd gebruik gemaakt van de optie waarmee pagina's in het Nederlands werden opgezocht. Als zoekterm werd alleen het polyseme woord gebruikt. Handmatig werden alle teruggegeven pagina's doorgezocht en aan de hand van hier onderstaande criteria kon een keuze gemaakt worden of de gevonden instantie geschikt was om in dit onderzoek gebruikt te worden. Vervolgens werden de instanties inclusief zowel de zin voor als de zin na die zin waarin de instantie voorkwam gekopieerd in een document, van waaruit de verzamelde tekst later omgezet zou worden naar platte tekst.
Het zoeken naar instanties van het polyseme woord werd vergemakkelijkt doordat bij de "in cache" functie, die achter de teruggegeven pagina's in Google staat, het gezochte woord geel geaccentueerd wordt. Dit bespaarde veel tijd tijdens de handmatige dataverzameling.
Wanneer de link in de zoekmachine verwees naar een homepagina, kwam alleen de tekst op deze homepagina in aanmerking om in het corpus te komen. Als de eerste pagina die in beeld kwam alleen bestond uit één link naar de achterliggende eigenlijke site (bijvoorbeeld "Welkom" of "Enter") dan werd er doorgeklikt en werd in de tekst op de eerstvolgende pagina gezocht naar een mogelijke instantie.

3.1.2 Criteria voor het verzamelen van instanties en context

Door middel van een zoekmachine werden de eerste 100 instanties waarin het gezochte ambigue woord voorkomt op het Internet verzameld. Om in aanmerking te komen voor opname in het corpus is het belangrijk dat het gezochte woord:

(1) niet voorkomt in een eigennaam, zowel van een persoon als van een instantie. De aanwezigheid van een hoofdletter kan hierop duiden, tenzij het woord aan het begin van een zin staat of tijdens het handmatig taggen duidelijk bleek dat het woord geen eigennaam was. Als een woord begint met een hoofdletter kan dit bijvoorbeeld ook een tikfout zijn van de auteur, of bewust gebruikt zijn om een metaforische betekenis aan te duiden;

(2) geen Engelstalig of anderstalig woord is dat niet overgenomen is in de Nederlandse taal zoals die omschreven is in Van Dale Groot Woordenboek der Nederlandse Taal;
(3) niet voorkomt binnen een tekst geschreven in een op Nederlands lijkende taal zoals het Afrikaans, Fries of in de tekstuele vorm van een dialect, zoals Limburgs of Twents, wat problemen op zou kunnen leveren in het verdere verloop van het onderzoek;
(4) geen deel is van een langer woord, of verbonden is met een ander woord met behulp van een verbindingsstreepje. Als de instantie grenst aan aanhalingstekens of een openend of sluitende haak wordt dit wel getolereerd zolang aan de andere kant van de haak er maar geen ander woord direct tegen aan staat. Uiteraard wordt een instantie wel geaccepteerd als deze grenst aan de gangbare leestekens zoals punten, komma's etcetera;
(5) geen deel is van een enkele zin die een deel vormt van een opsomming. Hierbij is het vaak onmogelijk om context te vinden bij deze zin. Wanneer het gezochte woord voorin of achterin de zin staat, is er niet genoeg context om hiervan een instantie te maken die het systeem kan gebruiken als trainingsmateriaal;
(6) niet voorkomt in een titel of een andere vorm van een syntactisch onvolledige of afgebroken zin. Een titel is vaak te kort om een instantie te voorzien van voldoende context. Grammaticale zinnen bezitten over het algemeen een regelmatige woordvolgorde en bovendien verwacht het WSD-systeem dat in dit onderzoek gebruikt zal worden dat het hele zinnen als input ontvangt.

Vooral het laatste criterium, namelijk een verplichte verschijning van het woord in een grammaticale zin is streng, maar nodig. Uitzonderingen worden gemaakt wanneer de instantie voorkomt in een stuk songtekst of gedicht, waarin zinnen niet per definitie grammaticaal hoeven te zijn, of wanneer de zin waarin de desbetreffende instantie staat alle kenmerken van een goede zin heeft behalve grammaticaliteit, terwijl de zin ervoor en de zin erna wel grammaticaal zijn. In deze uitzonderingsgevallen kan in de context namelijk vaak nog wel bepaalde woorden gevonden worden die als extra belangrijk kunnen worden geacht in het disambigueringproces, zoals werkwoorden. Hoewel dit onderzoek alleen gebruik maakt van directe context en geen keywords die verderop in de context staan, wordt de voorgaande zin vaak afgesloten met een werkwoord, terwijl de gezochte instantie (in een ongrammaticale zin) vaak aan het begin van de nieuwe zin staat; dit is dan ook de reden waarom deze instanties wel meegenomen zijn in de dataverzameling.
Zonder de aanwezigheid van deze criteria zou het onderzoek beduidend minder gemakkelijk te reproduceren zijn. Zonder het in acht nemen van deze criteria zullen veel meer instanties in een corpus terecht kunnen komen wanneer een even grote hoeveelheid webpagina's geraadpleegd was geweest. De kwaliteit van de gevonden instanties zal echter beduidend minder zijn, omdat het korte en snelle taalgebruik op het Web veel instanties zal opleveren die niet vergelijkbaar zijn met andere vormen van taalgebruik zoals die te vinden zijn in genres als boeken en journalistieke teksten.

Het kan voorkomen dat een woord meerdere malen op dezelfde webpagina voorkomt. Omdat we op zoek zijn naar de eerste 100 instanties van dat woord worden de woorden die meermaals op één pagina voorkomen gewoon meegenomen in het corpus totdat het aantal van 100 instanties inclusief context is bereikt. Soms komt binnen de grammaticale zin of binnen de context hetzelfde gezochte woord meerdere keren voor. Uiteindelijk zal hierdoor bij een aantal van de 30 woorden het aantal instanties hoger uitvallen dan 100.
Een kenmerk van Internet is tevens dat mensen teksten kopiëren. In dit geval zullen nooit meer dan twee keer dezelfde zin inclusief dezelfde context in het corpus worden opgenomen. Dit wordt gedaan omdat een grote hoeveelheid instanties die precies hetzelfde zijn niets toevoegt aan het testmateriaal dat gegeven wordt aan het WSD-systeem, laat staan als potentieel trainingsmateriaal voor zo'n systeem. Bij gebruik van een deel van het corpus als testmateriaal kan een grote hoeveelheid van dezelfde instanties zelfs een vervorming van de classificatieresultaten opleveren.

3.2 Opbouw van het Internet-corpus

Nadat de instanties verzameld werden zijn nog een aantal handelingen hierop verricht voordat deze gebruikt konden worden door het WSD-systeem. Deze handelingen, inclusief een uitweiding over het taggen van betekenissen in het corpus zullen hier besproken worden.

3.2.1 Voorbereidende handelingen verricht op het corpus

De instanties inclusief de context met de voorgaande en de daaropvolgende zin worden als platte tekst in één bestand opgeslagen. Iedere zin waarin het gezochte polyseme woord voorkwam werd op een aparte regel geplaatst met de twee contextzinnen ervoor of erna. Dat wil zeggen dat voor alle 30 woorden 100 regels in dit bestand te vinden zijn, allen gescheiden door een linebreak.
Overigens was het in een redelijk groot aantal gevallen niet mogelijk om twee goede zinnen als context te vinden. Vaak kwam dit omdat de zin waarin de instantie werd gevolgd niet gevolgd werd door een andere zin. Webteksten zijn vaak zo kort mogelijk, dus deze situatie was dan ook te verwachten.
Bij de volgende stap werd alle tekst in het corpus omgezet naar kleine letters. Een aantal instanties had vanwege het feit dat ze het beginwoord waren in een zin een hoofdletter, wat hinderlijk is voor het gebruikte WSD-systeem. Vervolgens zijn alle zinnen getokenizeerd, waarbij alle leestekens los werden gezet van de woorden. Een aantal handmatige correcties was hier noodzakelijk, om er voor te zorgen dat alle sensetags en daaropvolgende woorden gescheiden zouden zijn door een spatie.
Vervolgens werd dit opgebouwde corpus bewerkt met een part-of-speech tagger, namelijk de Memory Based Tagger (Daelemans, Zavrel, Berck & Gillis, 1996) getraind op het Eindhoven corpus. Deze POS-tagger zal in paragraaf 3.4.2 verder besproken worden. De POS-informatie die toegekend werd aan het focuswoord en de contextwoorden in het window zal gebruikt worden in het verdere disambigueringsproces.

Hieronder wordt een voorbeeld gegeven van een deel van de originele zin in het Internet-corpus, dat uiteindelijk verwerkt is tot een instantie in de dataset. Het polyseme woord zelf is uiteindelijk niet opgenomen als feature in de instantie. Voor ieder woord wordt namelijk in het geval van het kinderboeken-corpus een apart lerend systeem getraind, dat alleen dat specifieke woord hoeft te disambigueren. Hierdoor is dat specifieke woord niet informatief en hoeft het niet in de instantie opgenomen te worden.

- Originele zin:
ik vind dit persoonlijk wel jammer omdat als je al de grafische opties
uit moet zetten je aardig//aardig_behoorlijk teleurgesteld wordt .

- Instantie in de dataset:
V zetten Pron je Adj V teleurgesteld V wordt aardig_behoorlijk

De instanties die ingevoerd werden in het WSD-systeem bestaan uiteindelijk uit de 4 contextwoorden inclusief de bijbehorende part of speech tags die er voor staan en de betekenis van het polyseme woord binnen die instantie. Het geheel van al deze instanties van een bepaald polyseem woord, waarmee getraind wordt voor het disambigueren van een specifiek woord, wordt een word expert genoemd. Voor de 30 polyseme woorden in het kinderboeken-corpus is dus een word-expert gecreëerd. Voor iedere word expert is de beste setting gebruikt om deze te trainen, zoals beschreven in (Hendrickx et al., 2002).

3.2.2 Het taggen van betekenissen

Aan alle instanties wordt handmatig één van de betekenissen toegekend zoals die ook aan het kinderboeken-corpus toegekend werden. De verwachting dat een deel van de betekenissen tijdens het handmatig annoteren van de gevonden instanties in het Internet-corpus binnen de betekenissen vallen die al aangewezen zijn in het kinderboeken-corpus kwam uit. Bij een deel van de teruggevonden instanties van bijna alle polyseme woorden moesten echter één of meerdere nieuwe betekenissen toegevoegd worden, aangezien deze nog niet voorkwamen in het kinderboeken-corpus.
Het is niet bezwaarlijk dat deze nieuwe betekeniscategorieën opgesteld worden. De data geeft hier namelijk gelegenheid toe, doordat de verscheidenheid aan teksten op het Internet veel groter is dan in kinderboeken. Daarnaast werd binnen webteksten een deel van de instanties van diverse woorden op een metaforische wijze gebruikt worden – iets dat binnen teksten in kinderboeken minder vaak voorkomt.
De betekenissen werden in vierkante haken direct achter het polyseme focuswoord geplaatst, van dit focuswoord alleen gescheiden door een spatie. De betekenissen werden op dezelfde wijze geformuleerd als in het kinderboeken-corpus; er werd vaak gekozen voor een synoniem van de betekenis of een ander woord met een duidelijk verband met de betekenis. In het geval van collocaties werd deze collocatie voluit geschreven als betekenis, waarbij de diverse woorden binnen de collocatie aan elkaar werden verbonden door liggende strepen (underscores). Op deze manier kunnen later alle mogelijke verschillende betekenissen worden gescoord op hoe vaak ze voorkomen, zodat er een totaaloverzicht ontstaat van de verschillende betekenissen van elk van de dertig polyseme woorden.

3.3 De twee datasets

Om de schaal aan te geven waarop dit onderzoek is uitgevoerd wordt hieronder van beide corpora een overzicht gegeven van een aantal kwantitatieve kenmerken. Daarnaast wordt er dieper ingegaan op het kinderboeken-corpus, dat hiervoor alleen nog maar zijdelings besproken is.

3.3.1 Het kinderboeken-corpus

Dit kinderboeken-corpus is samengesteld voor een studie van Schrooten en Vermeer (Schrooten & Vermeer, 1994) en bestaat uit de tekst van 102 geïllustreerde kinderboeken voor de leeftijdsgroep van 4 tot 12 jaar. De oorspronkelijke bedoeling van de ontwikkeling van dit corpus was het creëren van een realistische woordenlijst van de woorden die via onderwijsmateriaal aangeboden worden op basisscholen. Deze lijst kon verder gebruikt worden in de studie naar testen die de geletterdheid van de kinderen konden meten; een voorbeeld van zo'n test is het stellen van vragen aan kinderen over de hoeveelheid betekenissen die zij konden formuleren van bepaalde polyseme woorden (Hendrickx et al., 2002).
Elk woord in deze tekst is handmatig geannoteerd met de bijbehorende betekenis, door zes personen. Iedere annotator heeft een ander gedeelte van de data geannoteerd. Elk woord kreeg een betekenis die in de meeste gevallen uitgedrukt werd door een gerelateerd concept, of in ieder geval een beschrijving van de specifieke betekenis van het woord. Net als in het Internet-corpus werd geen hiërarchische informatie toegevoegd aan de tags.
De tags die gegeven werden bestonden uit het lemma van het woord en de beschrijving van de betekenis in één of twee woorden <drogen_nat> of een verwijzing naar de grammaticale categorie <fiets_N>, <fietsen_V>. Werkwoorden hebben als tag vaak het lemma en een verwijzing naar hun functie in de zin <is/zijn_kww> (Hendrickx & Van den Bosch, 2001). Wanneer een woord in het corpus maar één betekenis had werd dit aangegeven met een simpele "=". Dat zelfde geldt voor eigennamen en geluidsimitaties. Ook werden in de database betekenissen van uitdrukkingen, bestaande uit meerdere woorden opgenomen, waarbij ieder woord de betekenis van de uitdrukking toegewezen kreeg.

3.3.2 Het Internet-corpus

De opzet van het Internet-corpus is hiervoor al besproken. Daarom zal hier nu vooral ingegaan worden op een aantal kwantitatieve kenmerken van dit corpus. Het corpus bestaat uit 30*100 tekstfragmenten, ieder bestaande uit 2 of 3 zinnen.Het volledige corpus bestaat uit in totaal 3881 voorkomens van de in totaal 30 polyseme woorden.
In tegenstelling tot het kinderboeken-corpus zijn alle betekenissen door één en dezelfde persoon getagd, wat de consistentie van de betekenistoekenning waarschijnlijk groter maakt dan die in het kinderboeken-corpus. In tegenstelling tot het kinderboeken-corpus zijn alleen de instanties van de 30 polyseme woorden handmatig getagd. De rest van de woorden is vooralsnog niet voorzien van een sense tag. Hoe de tags weergegeven zijn is te vinden in paragraaf 3.2.2.

3.4 Materiaal

Als laatste zal in dit hoofdstuk het materiaal dat gebruikt is voor dit onderzoek besproken worden. Als eerste zal de zoekmachine Google besproken worden die gebruikt werd bij de dataverzameling. Vervolgens zal in het kort de gebruikte-POS tagger besproken worden. Daarna zal het WSD-systeem MBWSD-D aan de orde komen dat gebruikt werd bij de WSD-test.

3.4.1 Zoekmachine Google

Voor dit onderzoek wordt gebruik gemaakt van de zoekmachine Google. Google indexeert een groot aantal pagina's van het Internet (huidige opgave: ruim 3 miljard webpagina's). Als startpunt is het adres http://www.google.nl gekozen. De verwachting is dat er altijd wel Engelstalige links teruggegeven zullen worden door de zoekmachine, maar deze instanties zullen, zoals hiervoor al gezegd, niet meegenomen worden tijdens de dataverzameling.
Google werd tijdens de verzameling van instanties altijd geraadpleegd door middel van dezelfde browser, namelijk Internet Explorer 5.0 van Microsoft. Tijdens de dataverzameling werd er altijd gebruik gemaakt van een personal computer met dezelfde configuratie en met aanwezigheid van software om websites die gebruik maken van Macromedia Flash te kunnen oproepen in de browser.

3.4.2 Memory Based Tagger

De Memory Based Tagger (Daelemans et al., 1996) werd gebruikt om de parts of speech aan te geven van de woorden in het Internet-corpus. Deze tagger is getraind op het Eindhoven corpus, dat bestaat uit een grote hoeveelheid essays en literaire teksten van voor de jaren '80, waarvan de part of speech door mensen is getagd. Deze tagger is in staat om in 95 % van de gevallen de goede part of speech toe te kennen, maar zal dit resultaat voornamelijk behalen op nieuwsteksten.

3.4.3 MBWSD-D en classificatie

De automatische toekenning van betekenissen zal uitgevoerd worden door het programma MBWSD-D. MBWSD-D is een WSD-systeem, dat bestaat uit algoritmes voor memory based learning die geïmplementeerd zijn in TIMBL (Daelemans, Zavrel, van der Sloot & van den Bosch, 2001). TIMBL is verkrijgbaar op http://ilk.uvt.nl).
Memory based learning is een vorm van lazy learning, waarbij alle instanties die als trainingsmateriaal aangeboden worden, worden opgeslagen in het geheugen. Wanneer het systeem een betekenis moet toekennen gebeurt dat door te kiezen voor de betekenis die in het geheugen de grootste overeenkomst vertoont met het huidige focuswoord en diens locale context (Hoste et al., 2002).
MBWSD-D is gebouwd vanuit de gedachte dat WSD een classificatietaak is. Gegeven een ambigu of polyseem woord en de context die hierbij hoort, kan de classifier die getraind is op een dataset de juiste klasse, oftewel de juiste betekenis, toekennen gegeven de context (Hoste et al., 2002). Deze aanpak volgt de lijn van andere memory based systemen (zie onder andere Veenstra et al., 2000).
Het MBWSD-D systeem kan naast de locale context en de part of speech tag ook keywords of een combinatie van twee van deze drie vormen van informatie gebruiken als input voor het disambigueringsproces, maar deze andere vormen van learners zullen niet in dit onderzoek toegepast worden, omdat ze in een eerdere studie niet lijken bij te dragen aan een betere accuraatheid (Hendrickx et al., 2002).
Op het moment van dit onderzoek is het WSD-systeem getraind op leesmateriaal dat aangeboden wordt op de basisschool. Dit is niet hetzelfde als het niveau van taalbeheersing van een basisschoolleerling, aangezien kinderen meer gelezen en gehoord hebben aan taaluitingen via hun ouders, vrienden en televisie en leesmateriaal buiten de klas. Hierdoor hebben kinderen dus ook meer kans gehad om met bepaalde betekenissen in aanraking te komen.

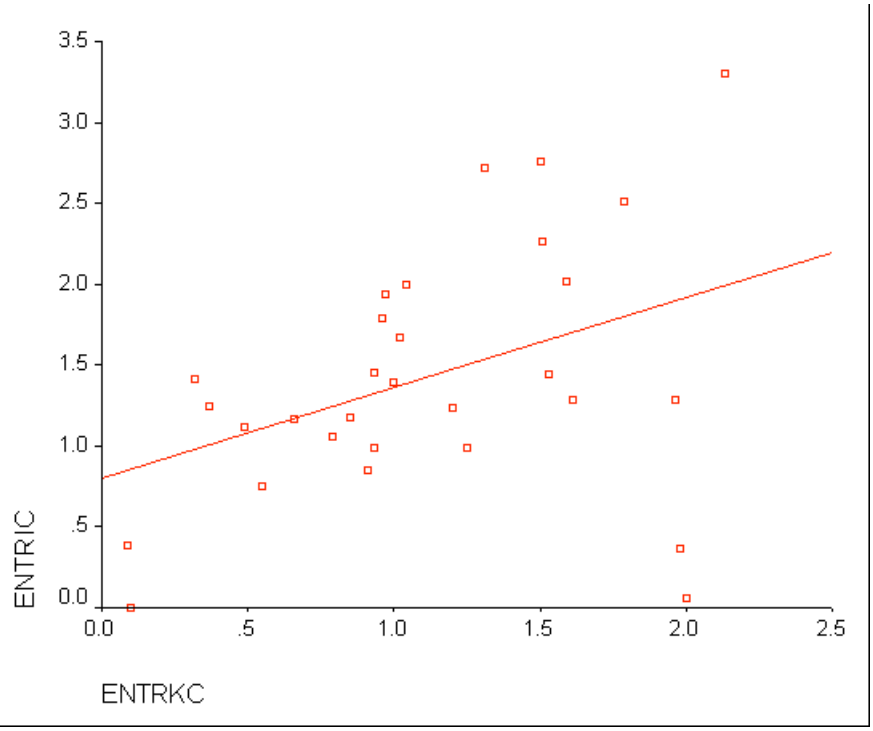
4. Resultaten

In dit hoofdstuk zullen de resultaten besproken worden die de uitvoering van dit onderzoek heeft opgeleverd. In de eerste paragraaf zullen de entropiecijfers van beide corpora besproken worden. Vervolgens zullen in de tweede paragraaf de resultaten op de accuraatheid van het WSD-systeem op beide corpora besproken worden. In beide paragrafen zal stil worden gestaan bij opvallende zaken die naar voren kwamen uit de resultaten.
Vervolgens zal gekeken worden naar een mogelijk verband tussen entropie en accuraatheid voor zowel het kinderboeken-corpus als het Internet-corpus.

4.1 Entropie: resultaten

In tabel 1 zijn de entropiecijfers te vinden voor alle 30 polyseme woorden die opgenomen zijn in dit onderzoek, voor zowel het kinderboeken-corpus als het Internet-corpus. Daarnaast is vermeld hoeveel verschillende betekenissen voor elk woord gevonden werden in beide corpora.
Voor een groot aantal woorden heeft het Internet-corpus meer instanties aangedragen dan het kinderboeken-corpus, waardoor de kans groter was dat er meer verschillende betekenissen gevonden werden. De 30 willekeurig gekozen woorden hadden gemiddeld 3.17 betekenissen per woord in het kinderboeken-corpus en 6.23 betekenissen per woord in het Internet-corpus. Het Internet-corpus bevat bijna twee keer zo veel verschillende betekenissen als het kinderboeken-corpus.
In tabel 1 zijn een aantal bijzondere zaken te ontdekken. Zo is de entropie voor het woord “bos” in het Internet-corpus minimaal. Er was geen andere betekenis binnen de 100 gevonden instanties te ontdekken dan de betekenis <bos_bomen>. Opvallend is dat ook het kinderboeken-corpus zelden een andere betekenis voor dit woord kon verstrekken, terwijl een betekenis als een gebundelde entiteit zoals een bos bloemen best te verwachten zou zijn.
Aan de andere kant konden voor het woord “passen” maar liefst 14 verschillende betekenissen onderscheiden worden. Doordat deze betekenissen daarnaast in aantal voorkomens nog in redelijk gelijke mate verdeeld waren in het Internet-corpus heeft dit geleid tot een zeer hoge entropie, namelijk 3.30. Het valt te verwachten dat het WSD-systeem bijzonder veel moeite zal hebben met dit polyseme woord. Hier zullen we straks verder op ingaan. Dit woord was overigens in het kinderboeken-corpus ook al een uitschieter met een entropie van 2.13.

In Figuur 2 wordt het verband in beeld gebracht tussen de entropie van de 30 polyseme woorden in het kinderboeken-corpus en de entropie van dezelfde woorden in het Internet-corpus. Er valt duidelijk te zien dat een overgrote meerderheid van de woorden binnen één groep valt die een stijgend verband laten zien tussen de entropie binnen het kinderboeken-corpus en de entropie binnen het Internet-corpus. Oftewel: woorden die in het Internet-corpus een hoge entropie bezaten, hadden vaak ook al een hoge entropie in het kinderboeken-corpus. De outliers die te zien zijn, zijn voornamelijk woorden met een opvallend lage entropie binnen beide corpora, of alleen binnen het Internet-corpus, zoals de woorden ‘bos’ en ‘stoppen’, die beiden in het Internet-corpus een entropie van onder 0.10 bezitten.
De lijn die te zien is in de grafiek is de best passende rechte lijn waarvoor geldt dat de som van de gekwadrateerde verticale afstanden van alle punten tot die lijn minimaal is. De correlatie van de entropie cijfers is .403 en significant (p= .027). De regressiecoëfficiënt is .292: de richting van het verband is positief.



Figuur 2: Spreidingsdiagram van de entropie op de 30 polyseme woorden in het kinderboeken-corpus (x-as) en in het Internet-corpus (y-as).

4.2 Accuraatheid: resultaten

Accuraatheid kan omschreven worden als het percentage correct geclassificeerde instanties. Een instantie is correct geclassificeerd als de betekenis teruggegeven door het WSD-systeem dezelfde is als de tag die toegekend werd aan de instantie in het corpus (Towell & Voorhees, 1998). Met accuraatheid is te meten hoe goed een WSD-systeem presteert op een bepaalde taak.
De score voor de accuraatheid is af te zetten tegen een baseline score. Een veelgebruikte baseline score is de score op de accuraatheid die het systeem behaalt wanneer aan alle polyseme woorden de meest voorkomende betekenis gegeven wordt. Het verschil tussen de baseline en de accuraatheid van het WSD-systeem is dus in feite de verbetering die het systeem heeft behaald dankzij het trainen van het systeem en ook het toepassen van een bepaalde configuratie.
Tabel 2 geeft voor alle 30 polyseme woorden de scores op de accuraatheid weer voor de disambiguering van twee testsets bestaande uit instanties uit het kinderboeken-corpus en instanties uit het Internet-corpus, net als de baseline scores voor beide corpora. Als configuratie is in dit onderzoek steeds de beste setting op de training set gebruikt zoals deze gevonden is in Hendrickx (Hendrickx et al., 2002) waarbij het context window op 2 werd gehouden. Van de in totaal 30 polyseme woorden kwamen er acht minder dan 10 x voor in de training set. Deze gevallen hebben als score op de accuraatheid de baseline score gekregen en worden in tabel 2 aangeduid met een asterisk (*). Voor een deel van deze woorden was de baseline score overigens 0.0.

De baseline score is gemiddeld over alle 30 polyseme woorden, met in totaal 3881 instanties 33.3 %. Dat betekent dat 33.3 % van de instanties afkomstig uit het Internet-corpus in de testset correct de meest voorkomende betekenis uit het kinderboeken-corpus toegewezen kregen. Bij het toekennen van de meest gebruikte betekenis uit het kinderboeken-corpus op een testset van instanties uit ditzelfde corpus was deze baseline score gemiddeld 81.2 %.
Wanneer het systeem de betekenissen toekent bij een testset, wordt een accuraatheid van 36.7 % bereikt bij het testen op instanties uit het Internet-corpus. Bij het toekennen van

betekenissen op een testset van instanties uit het kinderboeken-corpus behaalt het op het kinderboeken-corpus getrainde WSD-systeem een accuraatheid van 86 %. Er kan grofweg gesteld worden dat op beide corpora dezelfde verbetering wordt behaald tussen baseline score en feitelijke testscore.

Woord	Frequentie KC	Aantal senses KC	Entropie KC	Frequentie IC	Aantal senses IC	Entropie IC
bos	79	2	0.10	144	1	0.00
stoppen	4	4	2.00	131	2	0.06
hoop	9	4	1.98	156	4	0.37
keer	84	2	0.09	126	3	0.39
weer	289	5	0.55	120	3	0.75
stof	19	3	0.91	156	5	0.85
leven	52	3	0.93	125	5	0.99
spelen	60	3	1.25	136	5	0.99
meer	250	3	0.79	115	4	1.06
fijn	28	2	0.49	139	4	1.12
beter	42	2	0.66	112	5	1.17
buurt	18	2	0.85	130	4	1.18
land	37	3	1.20	125	5	1.24
staart	28	2	0.37	149	10	1.25
kopje	17	4	1.61	140	5	1.29
licht	26	5	1.96	151	7	1.29
geknipt	2	2	1.00	115	5	1.39
aardig	17	2	0.32	126	3	1.41
slot	9	3	1.53	117	8	1.44
hoor	53	3	0.93	135	4	1.45
werk	39	4	1.02	117	8	1.67
arm	18	2	0.96	121	8	1.79
hoog	38	4	0.97	115	7	1.94
hoofd	73	4	1.04	121	7	2.00
scheppen	3	3	1.59	127	12	2.02
pad	31	3	1.51	164	7	2.26
keren	6	4	1.79	107	12	2.51
best	30	4	1.31	109	9	2.72
klap	4	3	1.50	126	11	2.76
passen	7	5	2.13	126	14	3.30

Tabel 1: Entropiecijfers voor het kinderboeken-corpus (KC) en het Internet-corpus (IC) per polyseem woord.

Opvallend is dat in het Internet-corpus een groot deel van de verbetering ten opzichte van de baseline score behaald wordt op één woord, namelijk “hoor” (25.9 als baseline vergeleken met 76.3 als testscore). Ook bij “stof” en “weer” is er sprake van een hele verbetering (respectievelijk van 1.3 naar 19.9 en van 18.3 naar 44.2). Bij een groot aantal woorden (13 van de overige 23) is er geen sprake van een verbetering ten opzichte van de baseline score, waarvan bij 4 woorden de accuraatheid hetzelfde blijft als de baseline score en bij 9 woorden de accuraatheid zelfs lager is dan de baselinescore, zoals bijvoorbeeld bij de woorden “licht” en “kopje”, die beide een daling van zo’n 10 % ten opzichte van de baseline laten zien.
Tot slot kan een vergelijking gemaakt worden tussen de accuraatheid scores op de testset afkomstig uit het Internet-corpus en de baseline scores van het kinderboeken-corpus. Bij 6 woorden had het kinderboeken-corpus een lagere baseline score dan de accuraatheid van het Internet-corpus. In deze laatste gevallen zou training op de data die geleverd wordt door het Internet-corpus mogelijk een toekomstige verbetering brengen voor het WSD-systeem.

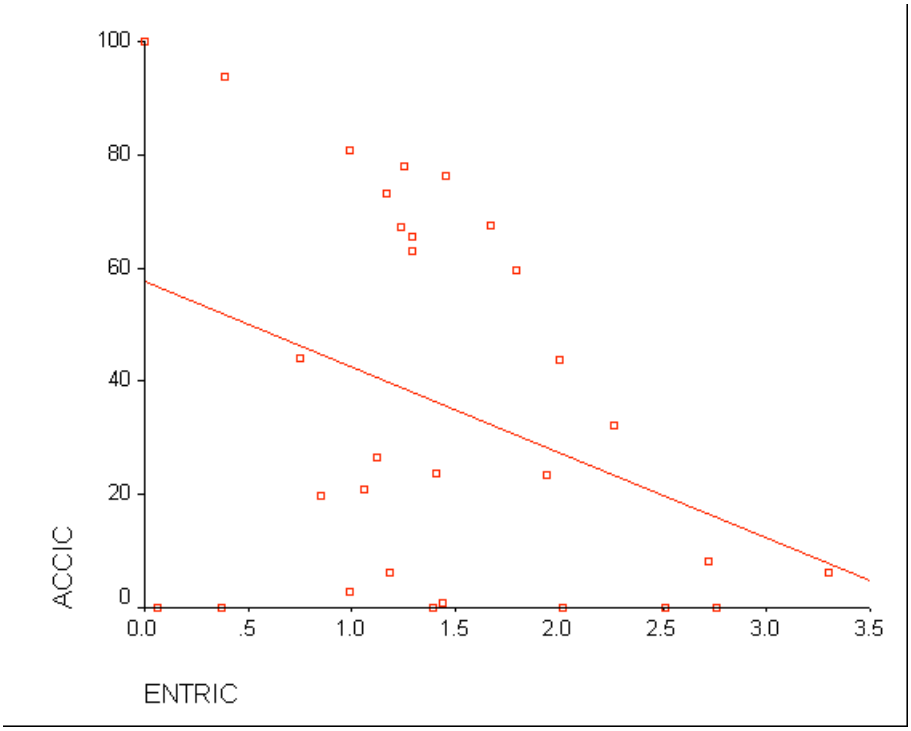
Woord	Aantal senses IC	Baseline score KC	Accuraatheid op testset KC	Baseline score IC	Accuraatheid op teststet IC
bos	1	100	100	100	100
keer	3	93.1	93.1	93.7	93.7
leven	5	74.2	90.3	73.6	80.8
staart	10	33.3	33.3	79.4	77.8
hoor	4	69.2	92.3	25.9	76.3
beter	5	72.7	72.7	74.1	73.2
werk	8	100	100	67.5	67.5
land	5	80	100	68.8	67.2
licht	7	53.3	53.3	74.2	65.6
kopje	5	62.5	75	72.1	62.9
arm	8	100	100	57.9	59.5
weer	3	88.9	96.0	18.3	44.2
hoofd	7	90	90	46.3	43.8
pad	7	25	100	18.9	32.3
fijn	4	100	100	26.6	26.6
aardig	3	100	100	24.6	23.8
hoog	7	75	75	21.7	23.5
meer	4	84.1	87.5	15.7	20.9
stof	5	80	100	1.3	19.9
best	9	55.6	22.2	9.2	8.3
passen	14	0	0 *	6.4	6.4 *
buurt	4	85.7	85.7	6.2	6.1
spelen	5	66.7	55.6	3.7	2.9
slot	8	100	100 *	0.9	0.9 *
geknipt	5	0	0 *	0	0 *
hoop	4	50	50 *	0	0 *
keren	12	50	50 *	0	0 *
klap	11	0	0 *	0	0 *
scheppen	12	0	0 *	0	0 *
stoppen	2	0	0 *	0	0 *

Tabel 2: Scores op de accuraatheid op het kinderboeken-corpus (KC) en het Internet-corpus (IC) en baseline scores op het kinderboeken-corpus en het Internet-corpus in percentages correct geclassificeerde instanties.

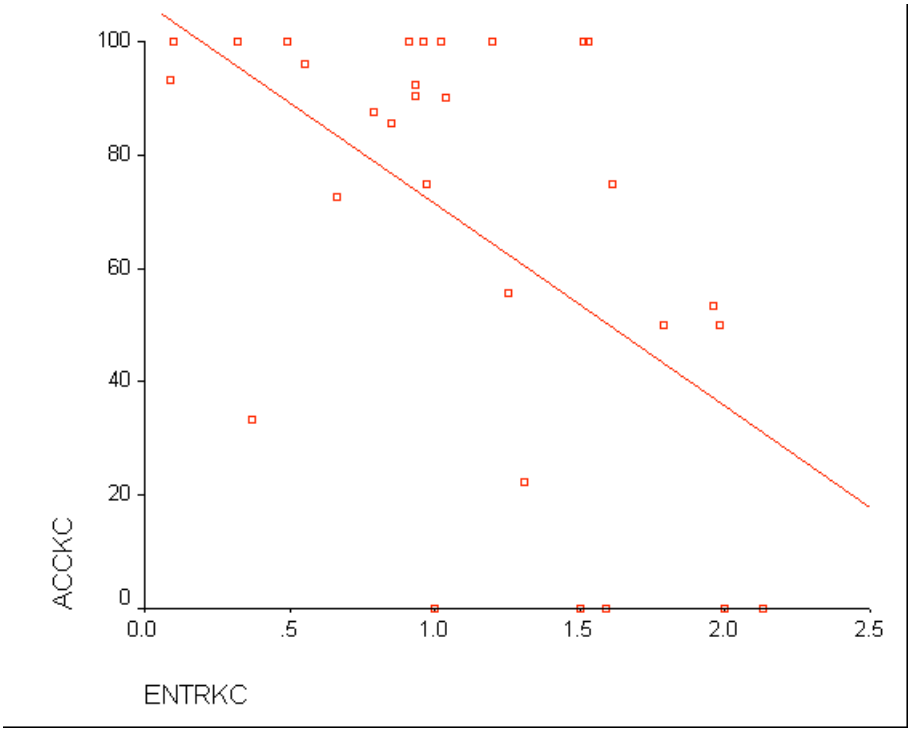
4.3 Verbanden tussen entropie en accuraatheid

In deze afsluitende paragraaf van hoofdstuk 4 wordt gezocht naar een verband tussen entropie en accuraatheid. Het valt te verwachten dat bij een hoge score op de accuraatheid er sprake is van een lage entropie binnen het corpus voor het desbetreffende polyseme woord. Aan de hand van spreidingsdiagrammen worden eventuele verbanden duidelijk gemaakt. De data waarop deze spreidingsdiagrammen zijn gebaseerd zijn terug te vinden in tabel 1 en tabel 2.

Als eerste wordt voor het Internet-corpus in Figuur 3 een grafische weergave gegeven van een mogelijk verband tussen de entropie binnen het Internet-corpus en de accuraatheid die behaald wordt op de testset. Het blijkt dat er amper een verband is te zien tussen een dalende accuraatheid en een stijgende entropie. De mogelijkheid om een verband te ontdekken wordt zwaar verstoord door het feit dat onderin het spreidingsdiagram de zes woorden terug te vinden zijn waarvan de accuraatheid feitelijk 0.0 was. Dit is dan ook de reden dat er een lage en negatieve correlatie van -.354 wordt behaald.

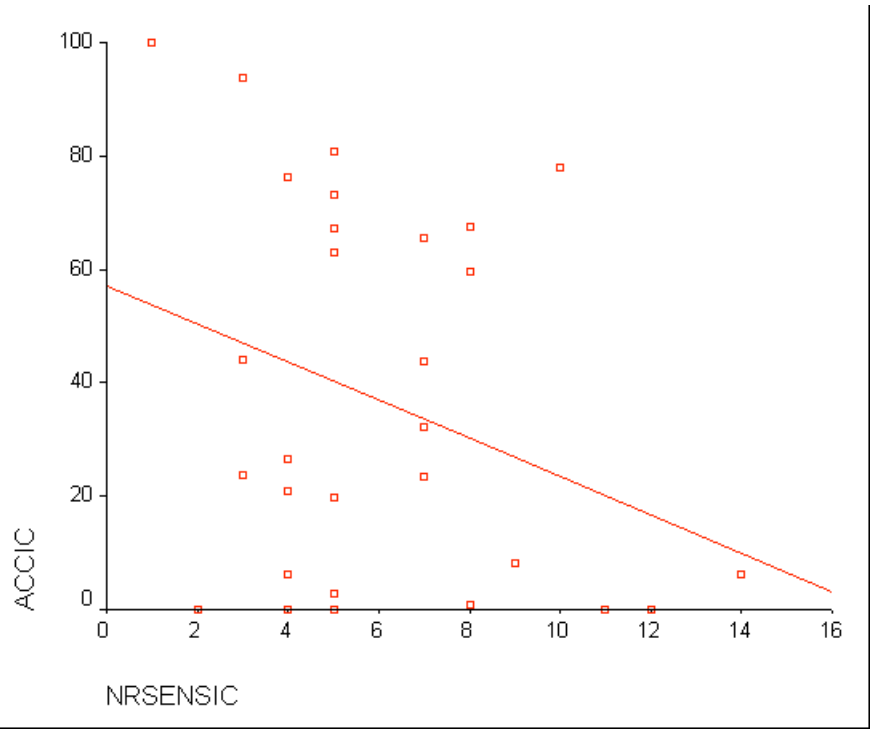


Figuur 3: De score op de accuraatheid in procenten op de testset van de instanties uit het Internet-corpus ten opzichte van de entropie van de instanties uit het Internet-corpus.



Figuur 4: De score op de accuraatheid in procenten op de testset van de instanties uit het kinderboeken-corpus ten opzichte van de entropie van de instanties uit het kinderboeken-corpus.

Figuur 4 toont hetzelfde als figuur 3, maar dan de scores op de accuraatheid voor de 30 polyseme woorden in het kinderboeken-corpus ten opzichte van de entropie die deze woorden bezaten binnen het kinderboeken-corpus. Het blijkt dat hier wel een redelijk sterk verband te ontdekken valt. De correlatie tussen accuraatheid en entropie binnen dit corpus is -.540 en deze correlatie is significant (p= .002).



Figuur 5: De score op de accuraatheid in procenten op de testset van de instanties in het Internet-corpus (y-as) ten opzichte van de hoeveelheid verschillende betekenissen in het Internet-corpus per polyseem woord.

Figuur 5 toont het verband tussen het aantal betekenissen van elk polyseem woord in het Internet-corpus en de accuraatheid die gescoord werd op dat woord. De lijn in de grafiek toont de lineaire regressie, waarvan de regressiecoëfficiënt -.312 is. De correlatie tussen de accuraatheid en het aantal betekenissen is -.324 en deze is niet significant (p= 0.08). Er valt enigszins te zien dat bij een toename van het aantal betekenissen van een bepaald woord de accuraatheid op dat woord ook lager wordt.
Hoewel in de eerste plaats de entropie als een betere voorspeller gezien wordt voor de afname van de accuraatheid geldt dit in mindere mate ook voor het aantal betekenissen van een polyseem woord. De drie woorden met de meeste betekenissen behalen een bijzonder lage tot zelfs minimale score in termen van accuraatheid.

5. Conclusies

In dit hoofdstuk zullen conclusies worden getrokken uit de in het voorgaande hoofdstuk gepresenteerde onderzoeksdata. Daarbij wordt teruggekeken op de deelvragen en de verwachtingen van dit onderzoek, om uiteindelijk tot een antwoord op de algemene onderzoeksvraag te komen. Als eerste zal de betekenisverdeling van de polyseme woorden besproken worden. Daarna zal gekeken worden naar de resultaten op de accuraatheid van het WSD-systeem. Vervolgens zal gezocht worden naar een mogelijk verband tussen entropie en accuraatheid. Tenslotte wordt getracht een antwoord te geven op de algemene onderzoeksvraag.

5.1 De betekenisverdeling

De betekenisverdeling van de polyseme woorden binnen het Internet-corpus wijkt af van de betekenisverdeling binnen het kinderboeken-corpus. Het blijkt dat het Internet-corpus gemiddeld een veel hogere entropie bezit dan het kinderboeken-corpus, wat te zien is aan de grote verschillen tussen de entropie cijfers van de woorden in beide corpora. Eén van de redenen hiervoor is te vinden in het feit dat het Internet-corpus veel meer verschillende betekenissen aanlevert, wat om te beginnen te verklaren is uit het feit dat het Internet-corpus voor de meeste woorden meer instanties aanbiedt dan het kinderboeken-corpus. Het Internet-corpus biedt gemiddeld per woord twee keer zoveel betekenissen aan als het kinderboeken-corpus. Daarnaast biedt het Internet-corpus teksten aan uit meer verschillende domeinen dan het kinderboeken-corpus en bevat mede daarom meer betekenissen die voorkomen in het taalgebruik van volwassenen, in tegenstelling tot het taalgebruik van kinderen.
Opvallend is tevens dat er - een aantal outliers uitgezonderd - een verband te ontdekken is tussen de entropiewaarden voor woorden in het kinderboeken-corpus en woorden in het Internet-corpus. Als een woord een hoge of lage entropie heeft in het ene corpus, is de kans groot dat hetzelfde het geval is in het andere corpus. De entropiewaarden stijgen ten opzichte van elkaar in een min of meer vast patroon.
De hogere entropie van het Internet-corpus kan er toe leiden dat een WSD-systeem meer problemen zal ondervinden wanneer het op instanties uit het Internet-corpus getest zou worden dan wanneer het op het kinderboeken-corpus getest zou worden. Omdat de sterke stijging van de entropie voor praktisch alle woorden geldt, kan deze verwachting gesteld worden. Dat deze verwachting bewaarheid zal worden wordt in de volgende paragraaf duidelijk gemaakt.
Overigens moet bij de aanwezigheid van de grote verschillen in betekenisverdeling nog een kanttekening gemaakt worden. De verdeling van betekenissen is beïnvloed door de criteria die gesteld werden tijdens de verzameling van instanties van de desbetreffende polyseme woorden. Eén van de vele voorbeelden hiervan is het woord “slot”, waarbij tijdens de verzamelingsfase vaak werd gestuit op de betekenis <slot_tot_slot>, die echter zelden is meegenomen in het corpus doordat deze zeer vaak voorkwam in titels van teksten, maar niet in de lopende tekst zelf. Over de invloed van de onderzoeksopzet in de uiteindelijk behaalde resultaten is meer te vinden in paragraaf 6.1.

5.2 De betekenistoekenning

De resultaten voor de toekenning van betekenissen laten zien dat het WSD-systeem veel slechter presteert wanneer het getest wordt op de instanties uit het Internet-corpus. Verder kan geconcludeerd worden dat zowel de baseline score als de accuraatheid van het systeem bij het Internet-corpus laag is, namelijk respectievelijk 33.3 % en 36.7 %. Hoewel hier sprake is van een WSD-systeem dat getest wordt op teksten uit een verscheidenheid aan domeinen kan deze score als laag ingeschat worden ten opzichte van de ingestuurde resultaten van systemen die meedoen aan de SENSEVAL competitie, die meestal rond de 85 à 90 % scoren.
Het systeem doet het wel beter dan de baseline score, zij het niet veel beter (3.4 %). Er is echter wel degelijk winst geboekt. Opvallend is dat het systeem min of meer een zelfde winst boekt wanneer het getraind is op het kinderboeken-corpus en getest wordt op het kinderboeken-corpus als wanneer het getest wordt op het Internet-corpus.
Het feit dat het testen op instanties afkomstig van het Web voor verbetering zorgt is hoopvol voor de toekomst. Dit kan betekenen dat een systeem dat in de toekomst intensief getraind is op instanties afkomstig van het Web een goede kans heeft om nog beter te scoren dan de hier behaalde 36.7 % in de vorm van een WSD-systeem dat volledige teksten kan disambigueren.
Zoals eerder al is opgemerkt moet rekening gehouden worden met het feit dat het Internet-corpus een grotere verscheidenheid aan betekenissen biedt dan het kinderboeken-corpus. In dit onderzoek bood het Internet-corpus in feite te veel betekenissen, aangezien voor een aantal woorden een nulscore werd behaald wegens een gebrek aan trainingsinstanties voor een deel van de meer specifiekere betekenissen.
Ook een vergelijking met andere systemen zal enigszins vertekend zijn, gezien het feit dat getest is op teksten uit allerlei domeinen, terwijl de meeste systemen getraind en getest worden op een corpus dat bijvoorbeeld bestaat uit krantenartikelen. Het systeem moet in dit onderzoek een meer ingewikkelde keuze maken om de goede betekenis toe te kennen.

5.3 Entropie als voorspeller van accuraatheid

Naar aanleiding van de resultaten die behaald zijn in dit onderzoek kan geconcludeerd worden dat er een redelijk verband is tussen entropie en accuraatheid. Voor het kinderboeken-corpus bleek dit verband echter sterker dan voor het Internet-corpus. Dit komt mede door de outliers in de vorm van woorden waar geen accuraatheid voor berekend kon worden, wat op zijn beurt veroorzaakt werd door het feit dat het WSD-systeem op andere data met een andere betekenisverdeling en deels andere betekenissen is getraind.
Wanneer een aantal woorden als voorbeelden uit de testset gelicht worden, blijkt het verwachte verband tussen entropie of het aantal senses en accuraatheid ondersteund, maar ook volledig ontkracht te kunnen worden. Woorden als “keer” en “leven” voldoen in het Internet-corpus precies aan de verwachtingen. Beide woorden hebben een lage entropie (respectievelijk 0.39 en 0.99 met drie en vijf betekenissen) en een hoge accuraatheid (respectievelijk 93.7 en 80.8).
Aan de andere kant laat het woord “stoppen” zien dat een lage entropie niet garant staat voor een hoge accuraatheid. In het Internet-corpus heeft dit woord maar 2 verschillende betekenissen, terwijl het in het kinderboeken-corpus nog 4 betekenissen toegekend kreeg. Dankzij de grote hoeveelheid websites die handelen over het stoppen met roken, had een zeer grote meerderheid van de instanties de betekenis <stoppen_met>. Deze betekenis kwam echter niet voor in het kinderboeken-corpus. Dit is dan ook de reden dat voor dit woord geen score op de accuraatheid berekend kon worden en dat de baselinescore van 0.0 is toegekend.
Hetzelfde geldt voor het woord “hoop”, dat met evenveel betekenissen in beide corpora toch een veel lagere entropie behaald in het Internet-corpus. In de betekenisverdeling van het Internet-corpus wordt één nieuwe betekenis geïntroduceerd ten opzichte van het

kinderboeken-corpus, namelijk <hoop_meta> die maar twee maal voorkomt. In het Internet-corpus hebben veruit de meeste instanties de betekenis <hoop_wens>. Omdat de andere betekenissen zeer weinig voorkomen kon ook hier geen accuraatheid voor berekend worden en is voor dit woord ook een score op de accuraatheid van 0.0 toegekend.
In het algemeen blijkt een deel van de willekeurig gekozen woorden een radikaal andere betekenisset te bezitten in het Internet-corpus vergeleken met het kinderboeken-corpus. Het kinderboeken-corpus blijkt hier onvoldoende of zelfs misleidende instanties aan te leveren als trainingsmateriaal en het systeem lijkt getraind te moeten worden op een nieuwe hoeveelheid instanties.

5.4 Internet geschikt als corpus?

Op de vraag of teksten al dan niet instanties afkomstig van het Internet geschikt zijn als inhoud van een corpus voor het verrichten van woordbetekenisonderzoek kan een tweeledig antwoord gegeven worden.
Aan de ene kant bieden de hiervoor besproken onderzoeksresultaten op de testset afkomstig uit het Internet-corpus geen uitermate positief beeld. De scores op de accuraatheid zijn bij veel woorden matig. Door de hoge entropie en een redelijk grote hoeveelheid specifieke betekenissen die de onderzochte woorden bezitten zal het veel moeite vergen om een hoge accuraatheid te behalen met instanties afkomstig van het Web. Er zal veel trainingsmateriaal afkomstig van het Web in plaats van uit het kinderboeken-corpus nodig zijn om een WSD-systeem de betekenissen die nu niet of onvoldoende gedekt werden te laten herkennen.
Aan de andere kant is juist die verscheidenheid aan domeinen het sterke punt van het gebruik als trainingsmateriaal van instanties afkomstig van het Internet. Corpora die bestaan uit krantenartikelen of boeken zijn niet zo divers als de teksten die gevonden worden op het Internet. Juist die diversiteit kan er toe leiden dat er genoeg instanties te vinden zijn om op termijn een uitstekend algemeen toepasbaar WSD-systeem te kunnen trainen, dat een groter aantal betekenissen van een polyseem woord kan onderscheiden dan andere systemen. Wel moet rekening gehouden worden met het feit dat dit systeem ook weer niet te gedetailleerde betekenissen hoeft te onderscheiden. De applicaties waarin een WSD-systeem wordt gebruikt eisen dit namelijk in de meeste gevallen niet.
Daarnaast ontstaan door vernieuwend taalgebruik constant nieuwe betekenissen. Van WSD-systemen gebaseerd op de betekenissen van woordenboeken kun je verwachten dat deze meer statisch zijn als het gaat om toepassen van nieuwe betekenissen, omdat deze misschien minder gemakkelijk uit te breiden zijn. Een goed werkend WSD-systeem wordt echter gedwongen om regelmatig een update toe te passen. Wat dat betreft heeft een systeem dat regelmatig nieuwe instanties afkomstig van het Web tot zich neemt een duidelijk voordeel. Teksten afkomstig van het Web vormen namelijk een goede bron voor voorbeelden van vernieuwend taalgebruik, aangezien juist op het Web vernieuwend taalgebruik vroegtijdig de kop opsteekt.

6. Discussie

In dit laatste hoofdstuk worden een aantal aspecten uit dit onderzoek verder toegelicht, voornamelijk met het oog op toekomstige onderzoeken binnen het onderzoeksgebied van word sense disambiguation. In de eerste paragraaf wordt een bespreking gegeven over de onderzoeksopzet en dan met name wat hieraan verbeterd kan worden.
In de tweede paragraaf wordt vervolgens een overzicht gegeven van potentiële problemen die aan het licht kunnen komen bij het gebruik van instanties afkomstig van het Internet. Deze paragraaf heeft voornamelijk een praktische inslag. In de derde paragraaf van dit hoofdstuk wordt in het kort verder ingegaan op de mogelijkheden van hiërarchieën bij het verbeteren van het WSD-onderzoek. Deze laatste paragraaf is min of meer het verlengde van paragraaf 2.2.

6.1 Discussie over onderzoeksopzet

De vraag is of de onderzoeksopzet van invloed is geweest op de tegenvallende resultaten met name op de betekenistoekening. Een aantal keuzes die gemaakt zijn in de onderzoeksopzet kunnen beschouwd worden als invloeden op de tegenvallende resultaten op de WSD-taak. De twee voornaamste zullen in de volgende twee paragrafen besproken worden. Daarna worden verbeteringen aangegeven voor een toekomstige onderzoeksopzet.

6.1.1 Menselijke annotator als bottleneck

De eerste keuze waar kritiek op geleverd kan worden is de keuze voor één menselijke annotator, in plaats van meerdere annotatoren. Verder wordt wel eens gesteld dat mensen minder goed zijn in WSD tagging dan in een taak als POS tagging. 'Als mensen, degenen die het trainingsmateriaal zouden moeten leveren, WSD al niet zeer goed zouden beheersen, verspillen we dan niet onze tijd met WSD (op deze wijze) te automatiseren?' vraagt Wilks (Wilks, 2000) zich af. Het antwoord op deze de vraag is negatief; het heeft zeker zin om mensen supervised corpora te laten annoteren, aangezien in de meeste gevallen niet eens zo'n hoog niveau verwacht wordt van een WSD-systeem.
Net als in veel andere onderzoeken ligt de moeilijkheid voor menselijke annotatoren ook in dit onderzoek voornamelijk bij het categoriseren van betekenissen die dicht bij elkaar liggen; deze gekozen tags komen bij menselijke taggers onderling niet snel overeen (Kilgarriff & Palmer, 2000). Véronis (Véronis, 2000) meldt zelfs dat 6 annotatoren het maar in minder dan de helft van de gevallen eens kunnen worden.
Doordat menselijke taggers niet in staat zijn om bijna perfect betekenissen toe te kennen tijdens het taggen zou dit betekenen dat er met het gebruik maken van door mensen getagd trainingsmateriaal een bovengrens ontstaat, als het gaat om het percentage correct geclassificeerde betekenissen. Want juist de tekst zelf kan een heel nieuwe betekenis aanbieden, die door menselijke taggers moeilijk te benoemen is, zoals bijvoorbeeld in dit onderzoek te zien was bij Internetteksten die handelden over spirituele abstracte concepten.
In het geval van dit onderzoek was er sprake van maar één menselijke annotator voor alle instanties. De keuzes voor het creëren van een bepaalde betekenis zijn door deze persoon niet op basis van een woordenboek gemaakt, maar arbitrair. Hoewel alle keuzes van deze annotator te motiveren zijn, kan worden verondersteld dat bepaalde keuzes geleid hebben tot minder goede resultaten op de WSD-taak. Een voorbeeld hiervan is de keuze om een onderscheid te maken tussen diverse vormen van "scheppen", namelijk <scheppen_metschep>, <scheppen_metnet> en <scheppen_metkookgerei>, wat een explosie aan betekenissen oplevert.

Wanneer meerdere professionele en goed getrainde annotatoren de instanties hadden geannoteerd en na goed overleg senses hadden vastgesteld, zou het mogelijk zijn geweest dat het resultaat van deze samenwerking een corpus had opgeleverd dat meer geschikt was geweest om te dienen als trainingsmateriaal. Als testmateriaal was de invloed op de hier behaalde resultaten minimaal geweest. Het grote verschil tussen de accuraatheid voor beide corpora was ook in deze situatie grotendeels blijven bestaan vanwege het grote verschil tussen het trainingsmateriaal (het kinderboeken-corpus) en het testmateriaal (het Internet-corpus).

6.1.2 De schaal van het onderzoek

Voor elk van de 30 polyseme woorden zijn 100 instanties verzameld afkomstig van het Web, om gebruikt te kunnen worden als testmateriaal. Voor veel woorden bestond het trainingsmateriaal uit nog minder instanties. De vraag rijst nu of een hoeveelheid van zo'n 100 instanties wel genoeg is voor een WSD-systeem om genoeg voorbeelden tot zich te kunnen nemen van minder vaak gebruikte betekenissen wanneer dit als trainingsmateriaal gebruikt zou worden. Zo noemt Ng (Ng, 1997) (zie het slotwoord) een duizendtal instanties aan trainingsmateriaal pas voldoende om een woord goed te kunnen disambigueren.
In het geval van dit onderzoek kon voor 6 van de 30 woorden slechts een minimale accuraatheid van 0.0 toegekend worden doordat van de betekenissen van deze woorden niet genoeg exemplaren in het trainingscorpus gevonden konden worden, of omdat de betekenis helemaal niet voorkwam in het trainingsmateriaal. De aanwezigheid van deze woorden had dan ook geen positieve invloed op de score op de totale accuraatheid die behaald werd op de WSD-taak. Tevens hielpen deze instanties niet om bepaalde verbanden aan te tonen tussen entropie en accuraatheid.
Het valt te verwachten dat wanneer er 200 of meer testinstanties van deze woorden waren verzameld de onderzoeksresultaten niet beter waren geweest. Het trainingsmateriaal waarop het WSD-systeem was getraind was tenslotte hetzelfde geweest. Wel zou een grotere hoeveelheid testmateriaal eventueel duidelijkere onderzoeksresultaten hebben opgeleverd, doordat de hoge scores op de accuraatheid iets lager zouden zijn en sommige lage scores op de accuraatheid iets hoger uitgevallen zouden zijn.
Wanneer het systeem getraind was op een grotere hoeveelheid instanties afkomstig uit kinderboeken, waren de resultaten die behaald werden tijdens de WSD-taak met het testmateriaal afkomstig van het Internet-corpus waarschijnlijk beter geweest. De nulscores die op een aantal woorden werden behaald waren waarschijnlijk in dit geval hoger uitgevallen. Doordat de nulscores beperkt zouden zijn geweest zou er een duidelijker verband aan te wijzen zijn tussen entropie en accuraatheid.

6.1.3 Toekomstige verbeteringen

Wanneer dit onderzoek in de toekomst opnieuw uitgevoerd zal worden, zijn een aantal verbeteringen ten opzichte van de huidige onderzoeksopzet aan te raden.
Ten eerste zal er gebruik gemaakt moeten worden van meerdere menselijke taggers van een professioneel niveau. Een voorwaarde voor goed trainingsmateriaal is volgens Wilks (Wilks, 1997) dat er training nodig is aan menselijke taggers om teksten te "sense-taggen op een redelijk hoog niveau en met redelijke consistentie tussen annotatoren". Training en geletterdheid zijn volgens hem vereist, omdat sommige betekenissen niet bij iedere spreker van een taal bekend zijn.
Hier komt dan gelijk een conflict met de tweede verbetering. Zoals al in hoofdstuk 1 gesteld werd is het menselijk taggen van betekenissen een kostbare zaak. Toch zal ook de scope van het onderzoek vergroot moeten worden, om niet zoals in dit onderzoek uiteindelijk voor een aantal woorden geen scores op de accuraatheid te kunnen behalen. Dit betekent dat het WSD-systeem veel meer door mensen getagde trainingsinstanties nodig heeft en liefst afkomstig van een andere bron dan kinderboeken. Getrainde

annotatoren zullen een hoeveelheid instanties van polyseme woorden moeten annoteren die velen malen groter is dan de hoeveelheid waarin die woorden in het kinderboeken-corpus voorkomen.

Vaak werden binnen een bepaalde tekst op het Web veel instanties van eenzelfde polyseem woord aangetroffen. Al deze instanties zijn vervolgens in het corpus verwerkt. Dit heeft gezorgd voor een bepaalde bias wat betreft aanwezigheid van bepaalde betekenissen ten opzichte van andere betekenissen. Het meest opvallende voorbeeld hiervan was het feit dat alleen de betekenis <bos_bomen> werd gevonden voor alle instanties van het woord "bos".
Gale, Church en Yarowski (Gale, Church & Yarowski, 1992) claimen dat in een enkele tekst een polyseem woord in 94 % van de gevallen in maar één enkele betekenis gebruikt wordt (de one-sense-per-discourse stelregel). Als echter van elke webpagina maar één instantie verzameld wordt, dan zou dit betekenen dat er veel meer teksten op het Web geraadpleegd moeten worden om alle, ook onvoorspelbare, betekenissen te voorzien van voldoende trainingsinstanties. Op de wijze waarop nu instanties zijn verzameld zijn in ieder geval ongetwijfeld een aantal betekenissen onvoldoende of zelfs helemaal niet in het corpus terecht gekomen.
Om in de toekomst een grote trainingsset beschikbaar te hebben is het misschien aan te raden om diverse corpora te combineren. Het Internet-corpus kan bijvoorbeeld na het consistent maken van betekenistags gecombineerd worden met het kinderboeken-corpus en eventueel een kranten-corpus. Op deze manier blijft een verscheidenheid van betekenissen behouden, maar is er wel meer trainingsmateriaal voorhanden om uiteindelijk een WSD-systeem te ontwikkelen met een hogere accuraatheid. Voorbeelden van betekenissen die in het ene corpus minder vaak voorkomen, kunnen bijvoorbeeld wel in overvloed geleverd worden door een ander corpus. Op deze manier is de kans groter dat voor minder vaak voorkomende betekenissen genoeg trainingsinstanties aangeleverd kunnen worden.

6.2 Geschiktheid van teksten op Internet

Zijn teksten afkomstig van het Web nu wel of niet te gebruiken als data voor WSD-onderzoek? In paragraaf 2.5 werd alvast een groot overzicht gegeven van algemene voordelen en nadelen van het Web als bron voor het opbouwen van een corpus. Daarna werd in paragraaf 5.3 een bondige conclusie getrokken, mede naar aanleiding van de resultaten die behaald werden in dit onderzoek. In deze sectie zullen verdere bevindingen over het gebruiken van instanties afkomstig van het Web voor WSD-onderzoek besproken worden, die opgedaan zijn tijdens de praktijk van het uitvoeren van dit onderzoek. Hierbij ligt het accent vooral op de problemen, zowel grote als kleine, die hierbij aan het daglicht kwamen.
In de eerste paragraaf worden eventuele problemen besproken bij het zoeken van instanties of goede tekst vanaf het Web. In de tweede paragraaf worden vervolgens gevolgen voor de betekenisverdeling besproken naar aanleiding van de instantieverzameling. In de derde paragraaf wordt daarnaast nog een kwaliteitsoordeel gegeven van de teksten op het Web.

6.2.1 Het zoeken van tekst op Internet

Het gebruik van zoekmachines, zoals in dit onderzoek, levert de nodige problemen op. Zo worden vaak verwijzingen gegeven naar pagina's die op dat moment niet toegankelijk zijn. Veruit de meeste pagina's zijn overigens wel toegankelijk als ze gegeven worden als resultaat van een zoekactie.
Naast teksten in het Nederlands zorgt de zoekmachine er ook voor dat pagina's in het Fries en in het Afrikaans, aan het Nederlands verwante talen, terugkomen in de resultaten van een zoekactie. Hoewel bij een automatische verzameling van teksten op Internet anderstalige teksten wel bij een automatische verwerking in een corpus met hoge precisie eruit gefilterd kunnen worden, worden ze in principe door de zoekmachine als Nederlands beschouwd.
Veel informatie op Internet wordt aangeboden op een manier waarop het moeilijker is om met de hand tekst te kopiëren. Zo geven zoekmachines steeds meer PDF-files terug en zorgen auteurs ervoor dat de functies van de rechtermuisknop niet werken, zodat het de gebruiker onmogelijk wordt gemaakt tekst te kopiëren. Uiteraard kan bij het automatisch verzamelen van teksten gebruik gemaakt worden van de broncode, waarna vervolgens de HTML-tags uit de broncode verwijderd worden. Op deze manier kan de tekst dan alsnog gebruikt worden om een corpus op te bouwen. Voor PDF-files geldt dat de tekstuele inhoud ook te kopiëren is, maar dan vanuit de HTML-versie die geleverd wordt door de zoekmachine of vanuit een PDF-reader. Uiteraard is dit wel een omweg vergeleken met het kopiëren uit een broncode.
Verder verschijnen er steeds meer paysites op Internet. De content die vroeger vrij toegankelijk was en nu nog steeds teruggegeven wordt in zoekmachines staat in werkelijkheid verborgen achter een password, waarvoor de gebruiker abonnementsgeld moet betalen. Hetzelfde geldt voor steeds meer digitale krantenuitgaven waarvoor de gebruiker zich op zijn minst aan moet melden. De verwachting is dat dit problemen op kan gaan leveren voor het eventueel automatisch verzamelen van Internetteksten om daarmee een groter corpus op te bouwen. Aan de andere kant komen hiervoor weer andere, nieuw gepubliceerde teksten, in de plaats. Het kan wel zo zijn dat juist Internetteksten met een bepaald topic niet meer vrij toegankelijk zijn en dat hierdoor in de toekomst bijvoorbeeld de verdeling van betekenissen van ambigue woorden kan veranderen.

6.2.2 Specifieke Internet-aspecten van verdeling van betekenissen

Een belangrijk kenmerk van Internet is dat het de auteurs van content erg gemakkelijk wordt gemaakt om elkaar te citeren of om grote stukken tekst van iemand anders zelf ook te publiceren. Om deze redenen is het niet verstandig om alle geschikte instanties van een bepaald woord ook te gebruiken in een corpus, omdat er regelmatig dezelfde stukken tekst in voor kunnen komen. Dit zou een verkeerd beeld geven van de mogelijke context doordat bepaalde contextwoorden vaker voorkomen in de omgeving van een instantie. Daarom is in het Internet-corpus dat voor dit onderzoek werd gecreëerd dezelfde zin hoogstens twee maal in het corpus geplaatst.
Nieuwsberichten, vooral wanneer ze afkomstig zijn van het ANP, zijn een typisch voorbeeld van teksten die vaak verbatim op diverse pagina's te vinden zijn, doordat ze bijvoorbeeld overgenomen worden in de digitale edities van vele kranten. Daarnaast wordt op forums bijvoorbeeld soms op elkaars postings gereageerd door de voorgaande posting te citeren.
Ook hebben auteurs van Internetteksten de neiging om binnen hun eigen teksten vaak zinnen, of soms zelfs hele alinea's, te hergebruiken. Zo wordt bij het geven van productinformatie voor elk apart product vaak dezelfde tekst gegeven en is alleen de productnaam of het serienummer vervangen. Wanneer al de instanties van het gezochte

woord, dat voorkomt binnen deze hergebruikte alinea's, opgenomen zouden worden in het corpus zal dit een zeer vertekend beeld geven van de context die hoort bij een bepaald polyseem woord.

Er werden door de zoekmachine veel teksten teruggegeven die handelen over esoterische onderwerpen. Dit zijn bovengemiddeld lange teksten, waarbij de kans groter is dat de gezochte instantie meerdere malen voorkomt. Dit gaf in dit geval problemen bij de woorden 'licht' en 'pad', die vaak in een figuurlijke, spirituele betekenis gebruikt werden. Hierbij is het zo dat een mens die handmatig betekenis toe moet kennen al problemen heeft, laat staan dat het door de computer met grote precisie gedaan kan worden.
Het Internet is ook een vehikel geworden voor webpagina's over zeer specifieke onderwerpen. Bepaalde onderwerpen komen daardoor meer voor dan andere onderwerpen. Het was opvallend bij de dataverzameling dat het Internet vooral gebruikt werd als nieuwsmedium, een plaats om te schrijven over hobby's, discussieforum en medium om mensen voor te lichten.
Vaak zijn het specifieke onderwerpen waarover schrijvers persoonlijk enthousiast zijn, zoals dierenliefhebbers. Dit enthousiasme is bepalend voor de hoeveelheid teksten die over deze specifieke topics gaan en dus uiteindelijk ook bepalend voor de verdeling van betekenissen bij bepaalde polyseme woorden. De tekst die uiteindelijk verzameld is in het corpus kan zeker beschouwd worden als een dwarsdoorsnede van het Internet, maar juist de grote aanwezigheid van voornamelijk nieuws, hobby's, discussie en voorlichting op het Internet kan de reden zijn dat de betekenisverdeling tussen het Internet-corpus en het kinderboeken-corpus zo veel verschilt, en dat deze betekenisverdeling ook veel verschillen zou kunnen bevatten ten opzichte van bijvoorbeeld een kranten-corpus.
De verdeling van woordbetekenissen in deze steekproef kan tot slot door een extra handeling zijn beïnvloed, die als doel heeft om 'gelezen te worden' op het Web. Zo zal bij ambigue woorden die zowel een betekenis hebben als Noun-vorm als Verb-vorm de nadruk binnen de teruggegeven instanties gemakkelijker kunnen liggen op de Noun-vorm, doordat de Noun-vorm meegenomen zou kunnen zijn in de metatags of titel van de desbetreffende webpagina, waardoor deze hoger scoort in de zoekmachines.

6.2.3 Kwaliteit van de teksten

De kwaliteit van de teksten op het Internet is vaak lager te noemen als je deze vergelijkt met de kwaliteit van teksten in boeken en professionele kranten. Wanneer volledige teksten afkomstig van het Web gebruikt worden om een corpus op te bouwen, moet gelet worden op de volgende zaken.

Een belangrijk kenmerk van teksten op Internet is het feit dat er gemiddeld meer spelfouten in kunnen staan dan teksten die overgenomen zijn uit kranten en boeken. Aan de grammaticaliteit van zinnen wordt door de niet-professionele auteurs van Internetteksten niet al te veel aandacht besteed. Ook worden er bij Internetteksten veel fouten in de interpunctie gemaakt, doordat er auteurs zijn die hun zinnen bijvoorbeeld niet afsluiten met een punt. Dit alles maakt dat Internetteksten in principe minder geschikt kunnen zijn om middels een corpus taal te leren aan een computer. Zo zijn gedichten en songteksten op Internet veel voorkomende tekstvormen, met kenmerken zoals een korte regellengte en vaak een onorthodoxe woordvolgorde.
Iedereen kan een bijdrage leveren aan een open domein als het Internet. Hierdoor kunnen ook kinderen en Nederlanders die Nederlands als tweede taal hebben teksten plaatsen op het Internet, die in hun mate van grammaticaliteit vaak afwijken van teksten die terug te vinden zijn in boeken en tijdschriften. Hoewel er expliciet op zoek werd gegaan naar grammaticale zinnen, zijn bijvoorbeeld delen uit teksten geschreven door basisschoolleerlingen wel degelijk opgenomen in het corpus.
Gesteld kan worden dat het samen nemen van al deze teksten een beter beeld geeft als dwarsdoorsnede van het taalgebruik zoals dat in Nederland gebezigd wordt in de huidige tijd dan andere corpora. Wanneer er de beschikking is over een zeer groot corpus met teksten

afkomstig van het Internet zal dit corpus, gezien ook de hoeveelheid vernieuwende taal, eerder de potentie hebben om als WSD-trainingsmateriaal te dienen dat ook in de toekomst nog gebruikt kan worden voor de Nederlandse taal.

Sommige alinea's die uiteindelijk in het corpus terecht zijn gekomen bestaan uit een mengsel van woorden uit de Engelse en Nederlandse taal. Dit is vooral het geval als er vaktermen uit een onderzoeks- of hobbygebied besproken worden. Alleen als deze Engelse woorden voor problemen zouden kunnen zorgen zijn deze instanties niet meegenomen in het corpus. Deze regel hoefde echter niet vaak toegepast te worden.
Wel is duidelijk dat het Nederlands meer en meer onder invloed staat van andere talen, met als gevolg ook dat sommige anderstalige woorden in de context zeer geschikt zijn om de betekenis van een polyseem woord te kunnen herleiden. Verder is eigenlijk te verwachten dat het grootste deel van de aanwezige anderstalige woorden in de context geen kwaad kunnen doen wanneer deze gebruikt worden binnen trainingsinstanties, maar aan de andere kant ook geen frequente contextwoorden zijn waarmee goede resultaten behaald kunnen worden tijdens een WSD-test.
De teksten die behoren bij het ene onderwerp verschillen veel in kenmerken die kwantitatief te benoemen zijn van teksten die behoren bij een ander onderwerp. Zo hebben teksten over hobby's en op forums een duidelijk kortere zinslengte dan wetenschappelijke en voorlichtende teksten. Dit zelfde kan ook gelden voor de hoeveelheid tekst op zich op een webpagina. Een grotere tekstlengte kan er voor zorgen dat het gezochte polyseme woord een groot aantal keren in een bepaalde betekenis opgenomen wordt in het corpus, zoals hiervoor al werd vermeld.

6.3 Discussie mogelijkheden hiërarchieën

Hoewel de accuraatheid op de WSD-test met de instanties afkomstig van het Internet veel lager is dan op de test met de instanties afkomstig uit het kinderboeken-corpus is dus wel eenzelfde verbetering ten opzichte van de baseline gehaald, namelijk zo'n 3,4 % ten opzichte van 4,8 % bij een test op instanties uit het kinderboeken-corpus. En dat terwijl het Internet-corpus een veel grotere hoeveelheid betekenissen aanbiedt en bijna alle woorden een hogere entropie bezitten, wat zou kunnen betekenen dat de WSD-taak moeilijker is. De vraag is vervolgens of een toevoeging van hiërarchieën aan het systeem voor een grotere verbetering ten opzichte van de baseline kan zorgen.
Voor het probleem dat betekenissen die dicht bij elkaar liggen bij veel polyseme woorden moeilijk te categoriseren zijn, kan verwacht worden dat een hiërarchische indeling van betekenissen een goede oplossing zal zijn. Mits er een overvloed aan trainingsmateriaal is voor elke betekenis zou theoretisch gezien op deze manier een WSD-systeem als bovengrens de menselijke kunde in het taggen kunnen bereiken, iets dat goed genoeg is binnen de meeste NLP-applicaties.
Hoewel deze hiervoor genoemde bovengrens ook behaald kan worden met de manier van WSD bedrijven zoals deze in dit onderzoek is uitgevoerd, heeft een WSD-systeem met hiërarchieën meer kans om deze bovengrens te bereiken. Het wordt voor zo'n systeem namelijk simpeler gemaakt om een keuze te maken voor een bepaalde betekenis. Als er geen betekenis gekozen kan worden uit een lager niveau in de hiërarchie, kan altijd teruggevallen worden op een betekenis uit een hogere categorie binnen de boomstructuur.

De vraag die vervolgens rijst is of een keuze voor een betekenis uit een hoger gelegen tak in de boomstructuur pertinent fout is wanneer de accuraatheid van een dergelijk systeem berekend moet worden. Resnik en Yarowski (Resnik & Yarowksi, 1997) menen dat een binaire evaluatie (goed/fout) niet voldoende is voor de beoordeling van een WSD-systeem en wezen dat per definitie de accuraatheid per instantie berekend moest worden aan de hand van een matrix die de afstand tussen de verschillende betekenissen aantoont, gebaseerd op een betekenishiërarchie. Dit zou volgens hen voor de meeste systemen moeten gelden. In het geval van een systeem dat gebruik zou maken van hiërarchieën zal dit dus zeker van toepassing moeten zijn.

Net als bij de verdeling van betekenissen op zich kan er door diverse annotatoren en onderzoekers lang gediscussieerd worden of een bepaalde hiërarchie nu wel of niet correct ontworpen is. Ook de vraag in hoeverre ze gebaseerd zouden moeten zijn op een instituut als WordNet (Leacock et al., 1998) is daarbij relevant. Het lijkt belangrijk om vooral nu, op het moment dat WSD-systemen die gebruik maken van hiërarchieën nog in opkomst zijn, criteria vast te stellen die een leidraad vormen waarlangs hiërarchieën van afzonderlijke polyseme woorden vastgesteld worden. Verder zullen deze hiërarchieën voor algemeen gebruik openbaar gemaakt moeten worden, zodat niet iedere onderzoeker afzonderlijk hiërarchieën hoeft te ontwerpen. Op deze manier wordt het onderzoeksgebied van WSD ook veel tijd en moeite bespaard.

Slotwoord

Wanneer onderzoek gedaan wordt naar word sense disambiguation, lijkt het tegenwoordig of WSD meer en meer een taak op zichzelf is geworden. Er wordt gestreefd naar systemen die een zo hoog mogelijke accuraatheid behalen. Het is echter nog steeds niet zeker of een goed WSD-systeem ook in dezelfde mate bij zal dragen aan de verbetering van de NLP-applicaties waarin deze gebruikt wordt, zoals bijvoorbeeld applicaties voor Machine Translation. Maar een klein deel van het WSD-onderzoek vindt plaats binnen een omgeving waarin deze geïmplementeerd is. Toch zal een goed WSD-systeem op zich ook al een goed stuk gereedschap zijn, claimt Wilks in (Wilks, 2000) en (Wilks, 1997).
Een goede basis voor een goed werkend WSD-systeem blijft voorlopig nog wel de aanwezigheid van goed trainingsmateriaal. Juist aan goed trainingsmateriaal zal in de toekomst gewerkt moeten worden. Ng (Ng, 1997) maakt een schatting dat voor de Engelse taal een handgetagd corpus van 3200 polyseme woorden met elk (in het slechtste geval, zonder selectiecriteria) 1000 instanties voldoende is om een goed werkend, breed toepasbaar WSD-systeem op te bouwen. Hij schat tevens dat de opbouw hiervan in het slechtste geval 16 mensjaren in beslag gaat nemen, een moeite die volgens hem gerechtvaardigd is.
Ondanks dat het Nederlands steeds meer onder invloed komt te staan van het Engels binnen een steeds internationaler gerichte samenleving, zal deze taal nog voor een lange tijd gebruikt worden. Hierdoor is het voor het Nederlands gerechtvaardigd en rendabel om een dermate grote investering te doen.

Literatuur:

Armstrong-Warwick, S. (1993). Preface. Computational Linguistics 19(1) iii-iv.
Brill, E. & Mooney, R.J. (1997). An Overview of Empirical Natural Language Processing. AI magazine 18:4 13-24
Chodorow, M., Leacock C. & Miller, G.A. (2000), A Topical/Local Classifier for Word Sense Identification. Computers and the Humanities 34 115-120
Church, K. & Mercer, R.L. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora. Computational Linguistics 19(1) 1-24.
Cruse, D.A. (1986) Lexical Semantics. Cambridge, England: CUP.
Daelemans, W. Zavrel, J., Berck, P. & Gillis, S. (1996). Mbt: A memory-based part of speech tagger-generator. In E. Ejerhed and I. Dagan (Eds.), Fourth Workshop on Very Large Corpora (pp. 14-27).
Daelemans, W. Zavrel, J., van der Sloot, K. & van den Bosch, A. (2001). TiMBL: Tilburg memory based learner, version 4.0, reference guide. ILK Technical Report 01-04, Tilburg University, te vinden op http://ilk.uvt.nl
Edmonds, P. (2002). SENSEVAL: The evaluation of word sense disambiguation systems. ELRA Newsletter, Vol. 7 No. 3.
Ellman, J., Klincke, I. & Tait, J. (2000). Word Sense Disambiguation by Information Filtering and Extraction. Computers and the Humanities 34 127-134
Gale, W., Church, K. and Yarowsky D. (1992) One sense per discourse, Proceedings of the 4th DARPA Speech and Natural Language Workshop.
Hendrickx, I. & van den Bosch, A. (2001). Dutch Word Sense Disambiguation: Data and Preliminary Results. Proceedings of Senseval-2, Toulouse. Te vinden op http://ilk.uvt.nl
Hendrickx, I. van den Bosch, A., Hoste, V. & Daelemans, W. (2002). Dutch Word Sense Disambiguation: Optimizing the Localness of Context. Proceedings of the workshop Senseval: recent successes and future directions. Te vinden op http://ilk.uvt.nl
Hoste, V., Daelemans, W., Hendrickx, I. & van den Bosch, A. (2002). Evaluating the Results of a Memory-Based Word-Expert Approach to Unrestricted Word Sense Disambiguation. Proceedings of the workshop Senseval: recent successes and future directions, Philadelphia. Te vinden op http://ilk.uvt.nl
Ide, N. & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. Computational Linguistics, 24(1), 1-40.
Karov, Y. & Edelman, S. Similarity-based Word Sense Disambiguation. Computational Linguistics (24(1) 41-60.
Kilgarriff, A. (1993). Dictionary word-sense distinctions: an enquiry into their nature. Computers and the Humanities 26.
Kilgarriff, A. (1997). “I don’t believe in word senses”. Computers and the Humanities 31 91-113
Kilgarriff, A. (2001). Web as corpus. Te vinden op http://www.itri.bton.ac.uk/~Adam.Kilgarriff/PAPERS/corpling.txt
Kilgarriff, A. & Palmer, M. (2000). Introduction to the Special Issue on SENSEVAL. Computers and the Humanities 34, 1-13
Kilgarriff, A. & Rosenzweig, J. (2000). Framework and results for english senseval. Computers and the Humanities. Special Issue on SENSEVAL, 34(1-2) 15-48
Leacock, C., Chodorow, M. & Miller, G.A. (1998). Using Corpus Statistics and WordNet relations for Sense Identification. Computational Linguistics 24 147-165
Madhu, S. & Lytle, D.W. (1965). A figure of merit technique for the resolution of non-grammatical ambiguity. Mechanical translation, 8(2) 9-13
Melamed, I.D. (1997). Measuring Semantic Entropy. Te vinden op http://acl.ldc.upenn.edu/W/W97/W97-0207.pdf

Ng, H.T. (1997). Getting Serious about Word Sense Disambiguation. In: Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How? Washington DC, USA
Resnik, P. & Yarowksi, D. (1997). A perspective on word sense disambiguation methods and their evaluation. In Proceedings of the ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What and How?" (pp. 76-86) Washington DC, USA.
Schrooten, W. & Vermeer, A. (1994). Woorden in het basisonderwijs. 15.000 woorden aangeboden aan leerlingen. TUP (Studies in meertaligheid 6)
Towell, G. & Voorhees, E.M. (1998). Disambiguating Highly Ambiguous Words. Computational Linguistics 24 125-145
Veenstra, J., van den Bosch, A., Buchholz, S. Daelemans, W. & Zavrel, J. (2000). Memory-based word sense disambiguation. Computers and the Humanities, 34(1-2) 171-177.
Véronis, J. (2000). Sense Tagging: Don't Look For The Meaning But For The Use? Te vinden op http://www.up.univ-mrs.fr/~veronis/pdf/2000comlex.pdf
Weaver, W. (1955). Translation. In W.N. Locke & D.A. Booth (Eds), Machine Translation of Languages (pp. 15-23). John Wiley & Sons, New York.
Wilks, Y. Senses and Texts. (1997). Computers and the Humanities. Te vinden op http://www.dcs.shef.ac.uk/~yorick/papers/cs-95/23.ps
Wilks, Y. (2000). Is Word Sense Disambiguation Just One More NLP Task? Computers and the Humantities 34 235-243.
Wilks, Y. & Stevenson, M. (1996). The grammar of sense: Is word sense tagging much more than part-of-speech tagging? Technical Report CS-96-05, University of Sheffield, Sheffield UK
Wilks, Y. & Stevenson, M. (1997a) Combining Independent Knowledge Sources for Word Sense Disambiguation. Proceedings of the Conference Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria, 1-7
Wilks, Y. & Stevenson, M. (1997b). The Grammar of Sense: Using part-of-speech tags as a first step in semantic disambiguation. Te vinden op ftp://ftp.dcs.shef.ac.uk/home/yorick/grammar.ps.
Wilks, Y. & Stevenson, M. (1998). Optimising Combinations of Knowledge Sources for Word Sense Disambiguation. Proceedings of the 36th Meeting of the Association for Computational Linguistics (COLING-ACL-98). Montreal, Canada.
Wittgenstein, L. (1953). Philosophische Untersuchungen (Philosophical Invensigations, translated by G.E.M. Anscombe). New York, Macmillan.
Yarowsky, D. (2000). Hierarchical Decision Lists for Word Sense Disambiguation. Computers and the Humanities 34 179-186
Zipf, G.K. (1935). The psycho-biology of language: an introduction to dynamic phililogy. Cambridge, MA: MIT Press.