Handleiding T-Scan

Versie 10 december 2020

Auteurs:

*Henk Pander Maat**
*Rogier Kraf**
*Nick Dekker**

Andere leden van het team dat T-Scan bouwde:

*Ko van der Sloot***
*Martijn van der Klis**
*Antal van den Bosch****
*Maarten van Gompel****
*Suzanne Kleijn**

* UiL-OTS, Universiteit Utrecht
** CIW, Tilburg University
*** CLS, Radboud Universiteit

Inhoudsopgave

1. Wat is T-Scan?

T-Scan is een softwaretool voor de analyse van Nederlandse teksten. De tool is vooral bedoeld om kenmerken in kaart te brengen die de complexiteit van de tekst beïnvloeden: SCAN kan ook gelezen worden een afkorting voor Software voor Complexiteits-Analyse van het Nederlands. De tool leent zich echter ook voor onderzoek naar andere stijlkwesties.

Er wordt sinds 2008 aan T-Scan gewerkt. De eerste versie van de tool is ontwikkeld door Rogier Kraf en Henk Pander Maat, waarbij de code is geschreven door Rogier Kraf in Python. In deze fase van het project hebben we geprofiteerd van financiering van de Stichting Taaltechnologie Utrecht, van het Utrecht Institute of Linguistics-OTS en van derde-geldstroominkomsten van de afdeling Taalbeheersing van de Universiteit Utrecht. Met de eerste T-Scanversie is bijvoorbeeld een heranalyse van de CLIB-leesbaarheidsdata uitgevoerd (Kraf en Pander Maat, 2009). In de periode 2008-2012 is T-Scan onderhouden door Rogier Kraf.

Deels ondersteund door een NWO-subsidie voor het project 'Naar een leesbaarheidsindex voor het Nederlands' is de tool overgezet naar C++, en sterk uitgebreid met nieuwe kenmerken. Daarbij heeft eerst Ko van der Sloot de code geschreven, van oktober 2012 tot 1 februari 2013 in overleg met Rogier Kraf, en daarna tot 1 juli 2014 in overleg met Henk Pander Maat. Sindsdien wordt de code geschreven door Martijn van der Klis, in overleg met Henk Pander Maat. Er wordt tot op heden aan T-Scan doorontwikkeld.

Ook andere mensen zijn belangrijk geweest voor T-Scan: Antal van den Bosch (veel software onder de motorkap van T-Scan is onder zijn leiding ontwikkeld), Maarten van Gompel (hij ontwierp en onderhoudt de CLAM-interface waarop T-Scan draait), Nick Dekker (testen van kenmerken; samenstellingsinformatie in woordenlijsten; handleiding), Suzanne Kleijn (testen van kenmerken), Lucas van Hoeij Schilthouwer Pompe, Nicole van Houten en Anniek Scholten (uitbreiden van woordenlijsten en samenstellingsinformatie).

T-Scan baseert zijn tekstkenmerken op de volgende tools en resources:

- Frog[1] (Van den Bosch et al., 2007): tokenisatie, lemmatisering, PoS-tagging en named entity recognition;
- Alpino[2] (Bouma, Van Noord, and Malouf, 2001): dependency parsing;
- SoNaR[3] (Oostdijk et al. 2013) and SUBTLEX-NL[4] (Keuleers et al. 2010): frequentielijsten;
- Referentie Bestand Nederlands[5] (Martin & Maks 2005): deze resource stond aan de basis van de semantisch geannoteerde woordenlijsten voor nomina, adjectieven en werkwoorden. De annotatielabels in deze lijsten zijn echter sterk aangepast en aangevuld door Henk Pander Maat. Alle woorden in de lijsten zijn gecorrigeerd door Henk Pander Maat Nick Dekker en Nicole van Houten. Er zijn ook veel woorden aan de lijsten toegevoegd. De nominalijst is van 35000 woorden gegroeid naar 83300 woorden, de adjectievenlijst van 8900 naar 13600 woorden, en de werkwoordenlijst van 6600 naar 7200.
- Nieuwe woordenlijsten samengesteld door Henk Pander Maat en Nick Dekker: nominale samenstellingen, ruimtewoorden, plaatswoorden, causale woorden, emotiewoorden, algemene nomina en algemene werkwoorden;

[1] http://ilk.uvt.nl/frog
[2] http://www.let.rug.nl/vannoord/alp/Alpino
[3] http://tst-centrale.org/nl/producten/corpora/sonar-corpus/6-85
[4] http://crr.ugent.be/programs-data/subtitle-frequencies/subtlex-nl
[5] http://tst-centrale.org/nl/producten/lexica/referentiebestand-nederlands/7-20

- Wopr[6] (Berck & Van den Bosch, 2009): maten voor trigramprobabiliteit, entropie en perplexiteit.

Deze handleiding is grotendeels geschreven door Henk Pander Maat. Rogier Kraf is de oorspronkelijke auteur van de tekstdelen in hoofdstuk 3 over D-level, nominalisaties, lijdende vormen en ontkenningen. Nick Dekker heeft materiaal aangeleverd voor hoofdstuk 2.

T-Scan is nog niet helemaal af; er wordt met name nog gewerkt aan nieuwe kenmerken op basis van een semantische ruimte voor het Nederlands (SoNaR Spaces).

T-Scan is op dit moment voor onderzoeksdoeleinden toegankelijk via de CLAM-interface (http://webservices-lst.science.ru.nl/). Een account aanvragen is daar noodzakelijk. Binnenkort komt T-Scan te staan op het Clarin-portal van het Instituut voor Nederlandse Taal. Als je T-Scan gebruikt in publicaties, verzoeken wij je te verwijzen naar Pander Maat et al. (2014) als bron.

[6] http://ilk.uvt.nl/wopr

2. Werken met T-Scan

2.1 Teksten geschikt maken voor T-Scan

T-Scan levert de beste resultaten wanneer teksten worden gebruikt die op de juiste manier zijn opgemaakt.

Algemene format-eisen

- Het bestand is in txt-formaat, en gecodeerd in UTF8.
- De bestandsnaam bevat geen spaties (je kunt wel lage streepjes gebruiken: bestand_nieuwsbericht)
- Het bestand bevat bij voorkeur geen typ- of spelfouten. Die fouten bemoeilijken de herkenning van woorden en de ontleding van zinnen.
- De tekst bevat geen vierkante haken ('[' en ']'). Verander ze in ronde haken.
- Speciale karakters worden juist weergegeven (aanhalingstekens, accentstreepjes, etc.). Zorg zo nodig dat aanhalingstekens en apostroffen zijn gecodeerd met de gestandaardiseerde symbolen (U0022, U0027).
- Als je wilt dat losstaande elementen zoals titels, kopjes en links worden overgeslagen, plaats er dan drie hekjes voor (###). Waarom zou je dat doen? Welnu, kopjes en titels leveren problemen op bij het splitsen van zinnen, omdat er geen punt achter staat. En als je er een punt achter zet, levert T-Scan een vertekende zinslengte doordat plotseling allerlei heel korte zinnen meetellen. Let op: als het document begint met overgeslagen element, raakt de software soms in de war. Zet in dat geval een witregel boven in het document.

Faciliteer het correct splitsen van zinnen

De volgende tips zorgen ervoor dat T-Scan zo weinig mogelijk fouten maakt bij het splitsen van zinnen.

- Sluit elke zin af met een punt of puntkomma.
- Controleer of je tekst op 'onverwachte' momenten punten bevat. Afkortingen (*o.a.*) kunnen soms problemen geven; dat geldt ook voor URL's.
- Puntkomma's worden steevast als zinsgrens gezien. Controleer zo nodig of je tekst op 'onverwachte' momenten puntkomma's bevat.
- T-Scan laat de zin doorlopen na een dubbele punt. Hij ziet de relatie tussen het deel voor en na de punt als een soort nevenschikking. Wil je dat niet, dan moet je de dubbele punt vervangen door een punt of puntkomma.
- Zorg dat je tekst geen opsommingstekens met een punt er meteen achter bevat: die worden abusievelijk als zinnen van één woord gezien.
- Zet de aanhalingstekens aan het eind van de zin vóór de punt, zeker wanneer er daarna meteen een harde return volgt. T-Scan is namelijk niet betrouwbaar bij het splitsen van zinnen als er na de punt nog een aanhalingsteken (enkel of dubbel) staat. Een andere optie is om het aanhalingsteken niet voor de punt te zetten, maar het te laten volgen door een witregel.
- T-Scan rekent delen tussen haakjes of gedachtestreepjes tot de zin. De ontleding van die gedeelten is soms wel problematisch. Verder ontstaan er problemen bij het scheiden van zinnen wanneer er punten of puntkomma's staan tussen haakjes of gedachtestreepjes.
- Zorg dat er geen harde returns voorkomen midden in een zin; die kunnen tot onterechte splitsingen leiden.
- Geef een nieuwe alinea aan door een witregel. Eén harde return is dus niet voldoende als alineagrens.

Opsommingen zijn soms lastig in te delen in zinnen. We hebben drie regels gehanteerd bij het redigeren van opsommingen.

1. Zijn de inleiding en de opsommingsleden volledige zinnen met een vervoegd werkwoord, dan is de tekst intact gelaten. Eventuele dubbele punten worden vervangen door punten, zodat de tool de zinnen apart houdt.

2. Is de inleiding van de opsomming een volledige zin en de leden niet, dan is de dubbele punt vervangen door een punt en zijn de opsommingsleden overgeslagen.
   Vergelijk voorbeeld (a), waarin het eerste onderdeel op zichzelf kan staan. Het stuk na de dubbele punt wordt overgeslagen; doen we dat niet, dan ontstaat een verkeerd beeld van de zinsbouw van de tekst.

   (a) Stuur de volgende documenten naar het Centraal Testamentenregister:
   - uw verzoek of het ingevulde aanvraagformulier;
   - de overlijdensverklaring van uw ouder.

3. Is de inleiding geen volledige zin, dan wordt één van de opsommingsleden daaraan toegevoegd. Veelal wordt daartoe een dubbele punt geschrapt. Als toevoeging kiezen we voor het langste lid dat grammaticaal past bij de inleiding. De andere leden worden overgeslagen.
   Waarom doen we dit? Koppelen we die andere leden ook aan de inleiding, dan wordt de zin vaak bijzonder lang. Zo kan de indruk ontstaan dat hij moeilijk is; maar dat is niet terecht, want door de opsommingsvorm wordt hij juist overzichtelijker. Vergelijk voorbeeld (b).

   (b) Deze volmacht eindigt in ieder geval:
   - in geval van faillissement, surseance van betaling, bij onder beschermingsbewind- of curatelestelling van volmachtgever;
   - bij overlijden van volmachtgever;
   - door herroeping van de volmacht door volmachtgever of door opzegging door volmachtnemer.

   In geredigeerde vorm wordt dit:

   Deze volmacht eindigt in ieder geval in geval van faillissement, surseance van betaling, bij onder beschermingsbewind- of curatelestelling van volmachtgever.

Aanhalingstekens

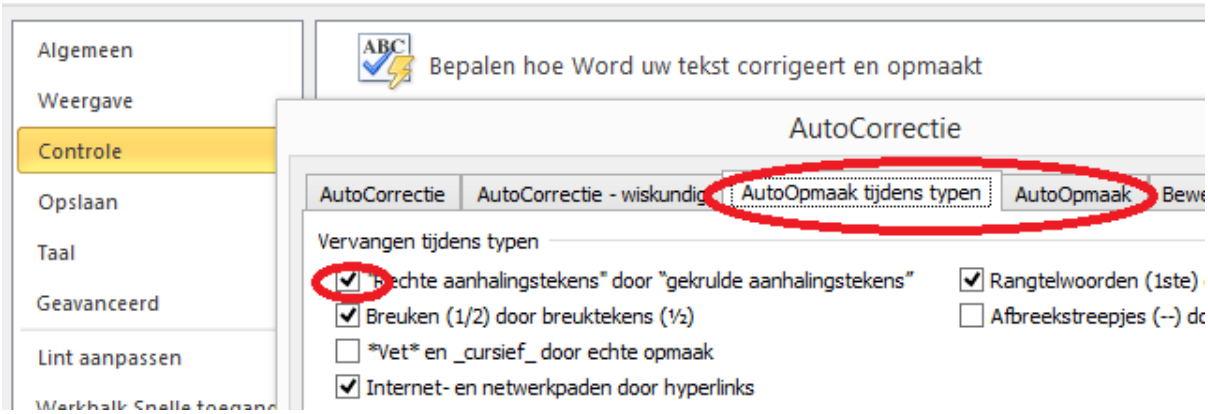
Speciale aandacht vergt de vorm van de aanhalingstekens. Daarmee ontstaan vaak problemen, omdat Word deze automatisch verandert. Voor T-Scan moeten de apostrof (bijv. in *zo'n* of *'s nachts*) en het aanhalingsteken (bijv. 'T-Scan') gelijk aan elkaar zijn. Wanneer een apostrof als aanhalingsteken gebruikt wordt, maakt Windows er echter automatisch een 'gekromd aanhalingsteken' van:

Gekruld: Zo’n ‘T-Scan’
Recht: Zo'n 'T-Scan'

Oplossing 1: speciale karakters vervangen in Word

Wanneer je je getypte tekst in Word direct wilt gebruiken in T-Scan, is het handig Word daarop in te stellen. Wanneer je bestaande teksten hebt (in Word of een andere tekstverwerker) die je T-Scanklaar wilt maken, kun je het beste de volgende oplossing (oplossing 2) gebruiken, zie daarvoor pag. 6.

1. Zet alle teksten in Word.
2. Ga naar *bestand* (in oudere versies van Word moet je op de ronde 'Windows-knop' klikken).
3. Kies voor *opties* en vervolgens voor *controle*, en voor *Autocorrectie opties*.
4. Nu moet je in twee tabbladen (Zie afbeelding) een vinkje weghalen (In de tabbladen *Opmaak tijdens typen* & *AutoOpmaak*).



Kies nu voor zoeken en vervangen (control + F). In het bovenste veld kies je voor de kromme aanhalingstekens en lange streepjes (die kun je kopiëren en plakken vanuit je tekst). In het onderste veld voer je een 'gewone' apostrof (door gewoon een apostrof in te toetsen) of afbreekstreepje in .
[soms wordt er nóg een type apostrof gebruikt in woorden als *zo'n* of foto's. Die moet je dan ook even vervangen door dat type apostrof in het zoeken-en-vervangen-veld te plakken en te vervangen voor een rechte apostrof.]

5. Plak de teksten in Notepad / Kladblok, en kies bij Encoding voor UTF-8.

Tip: kies in Word voor het lettertype Consolas, dat wordt ook in Kladblok gebruikt. De aanhalingstekens zijn daarin goed te onderscheiden.

Oplossing 2: speciale karakters vervangen in Notepad++

Zoals hierboven besproken heeft T-Scan moeite met bepaalde tekens, zoals gekromde aanhalingstekens en apostroffen. Het handmatig aanpassen van deze tekens kost veel tijd, en is behoorlijk foutgevoelig. Gelukkig is er een snelle manier om de tekst 'T-Scanklaar' te maken. Daarvoor is het wel nodig het (gratis) programma Notepad++ te downloaden (http://notepad-plus-plus.org/download/v6.6.7.html).

STAP 1

1. Maak alvast een mapje (bijvoorbeeld op het bureaublad) waarin je de teksten zet die je T-Scanklaar wilt maken. Deze moeten wel als .txt opgeslagen zijn, in UTF-8-bestandsindeling.
2. Open Notepad ++.
3. Zet de tekens die je wilt veranderen alvast klaar (hoeft niet, maar het is wel gemakkelijker). Het gaat dus om:
   - " *(kromme dubbele aanhalingstekens openen)*
   - " *(kromme dubbele aanhalingstekens sluiten)*
   - ' *(kromme enkele aanhalingstekens openen)*
   - ' *(kromme enkele aanhalingstekens sluiten)*
   - " *(rechte dubbele aanhalingstekens)*
   - ' *(rechte enkele aanhalingstekens)*

Je kunt de aanhalingstekens niet handmatig invoeren in Notepad++, omdat Notepad++ rechte aanhalingstekens niet automatisch vervangt door kromme. Je kunt de kromme aanhalingstekens dus het beste kopiëren vanuit Word.
4. Ga naar zoeken en vervangen (search > replace). Kies voor 'vervangen in files'.
5. Voer in 'directory' de naam van de map in waarin je bestanden staan.

STAP 2

1. Ga naar macro > start recording.
2. Voer nu de kromme aanhalingstekens in het bovenste vak in. De 'rechte aanhalingstekens' voer je onderin in (in zoeken en vervangen in files). Die kun je – als je ze hebt klaargezet – vinden in het bestand (zie punt 2 van stap 1). Doe dat voor de 2 'dubbele aanhalingstekens' (links/rechts) én voor de 2 'enkele aanhalingstekens (links/rechts).
3. Ga weer naar macro's en kies voor 'stop recording'.
4. Ga naar macro's en kies voor 'opslaan'. Geef je macro een handige naam.

Nu hoef je voortaan alleen nog maar de bestanden in de map te zetten en de macro opnieuw te draaien om de bestanden T-Scanklaar te maken.

Woordversie en zinsversie

Het is vaak verstandig om een tekst in twee versies aan te leveren aan T-Scan.

- De '*zinsversie'* gebruik je om de zinsbouw van je tekst te analyseren. Nu bevatten teksten allerlei elementen die geen zin vormen: kopjes, adressen, formulier- of tabelelementen in telegramstijl, opsommingsleden, enzovoort. In de zinsversie sla je die elementen over, of je plakt ze aan de elementen waarmee ze grammaticaal verbonden zijn. Dat laatste is vooral van toepassing op opsommingen. Wat betreft kopjes die wel een volledige zin zijn: veelal is het verstandig ook die over te slaan, omdat ze geen representatief beeld geven van de zinsbouw in de eigenlijke tekst.
- De zinsversie geeft geen goed beeld van kenmerken als de lexicale complexiteit, de concreetheid en de persoonlijkheid van je tekst. Om de niet-syntactische kenmerken te analyseren, gebruik je daarom een '*woordversie'* van de tekst. De woordversie bevat dus wél de kopjes, formulier- en

tabelelementen, opsommingsleden e.d. De woordversie slaat alleen de dingen over die T-Scan niet zelf over kan slaan. Namen hoef je dus niet over te slaan omdat T-Scan meestal zelf namen herkent en varianten van maten levert waarbij de namen niet meedoen. Het is wel veilig om mailadressen en huisadressen over te slaan, omdat die niet altijd als naam herkend worden.

Instructies rond speciale typen informatie

1. Afkortingen
   Bekende afkortingen zoals *e.d.* en *t.a.v.* zijn geen probleem voor T-scan. Deze kunnen als afkorting blijven staan.
2. Tijden
   Als je wilt dat T-Scan tijden opvat als een telwoord, moet je deze weergeven met een punt (bijvoorbeeld *12.00*). Bij een dubbele punt (*12:00*) wordt de tijdsaanduiding als een SPEC gezien.
3. Geldbedragen
   Wanneer je geldbedragen opgevat wilt zien als een SPEC naast een telwoord, dan moet je een spatie tussen munteenheid en getal zetten (*€ 50*).
4. Telefoonnummers
   Schrijf deze als één geheel. T-Scan pakt dit als SPEC (bijvoorbeeld *012-3456789* of *0123456789*).
5. Bankrekeningnummers.
   Schrijf deze als één geheel, zonder NLXX erbij. T-Scan pakt dit als SPEC. (bijvoorbeeld *BANK000123456*).
6. Samentrekkingen van woorden
   Als je wilt dat samentrekkingen van woorddelen goed herkend worden als woord, dan kun je deze beter uitschrijven. Zo wordt *week- en maandbedrag* veranderd in *weekbedrag en maandbedrag*.
7. 'Lege elementen'
   In sommige documenten staan met puntjes ontbrekende gegevens aangegeven, zoals plaats, datum, tijden, naam e.d. Wil je dat deze zinnen goed geanalyseerd worden, dan kun je die elementen beter aanvullen met fictieve plaatsen enz.
8. Adresgegevens
   In de zinsversie van je document sla je losstaande adresgegevens over (###).
   Bijvoorbeeld:
   ### Gemeente Amsterdam
   ### Postbus 10.100
   ### 1234 AB Amsterdam
   Adresgegevens in zinnen moeten wel in het geheel bewaard blijven.
   Bijvoorbeeld: *Schrijf dan een brief naar Gemeente Amsterdam, t.a.v. Mevrouw Jansen, Kerkstraat 1, 1234 AB Amsterdam.*
9. Mailadressen
   In de zinsversie sla je losstaande mailgegevens over (###)
   Mailadressen in een zin moeten wel bewaard blijven.T-Scan pakt een mailadres als SPEC.

2.2 T-Scan tekst laten overslaan

In teksten komen dingen voor die niet geanalyseerd moeten worden door T-Scan, zoals kopjes, tabellen en figuren. T-Scan biedt de optie om die dingen als 'commentaar' te labelen. Dat kan op twee manieren.

1. Als je de regel laat beginnen met ### (3 hekjes), gevolgd door een spatie, beschouwt T-Scan wat er op deze regel staat als commentaar. Deze optie is geschikt voor koppen die niet langer zijn dan een regel. *Let op*: als je document begint met een ### , zet daar dan een lege regel boven.

2. Als je een groter element wilt markeren, kan dat als volgt. Je laat de eerste regel van dat stuk tekst beginnen met <<<; je sluit het commentaar af door de laatste regel te laten beginnen met >>>. Ook de rest van die regel wordt nog genegeerd.

Neem het volgende fragment:

```
### Hoofdstuk 1
<<< Hier gaat de auteur in op
de eerste Krimoorlog
>>> en de nasleep daarvan
De Krimoorlog (1853-1856) ging tussen het Keizerrijk Rusland en een
alliantie van het Tweede Franse Keizerrijk, het Britse Rijk, het Ottomaanse
Rijk en Koninkrijk Sardinië.
```

Dat fragment bevat voor T-Scan alleen de tekst:

```
De Krimoorlog (1853-1856) ging tussen het Keizerrijk Rusland en een alliantie
van het Tweede Franse Keizerrijk, het Britse Rijk, het Ottomaanse Rijk en
Koninkrijk Sardinië.
```

We merken op dat grote hoeveelheden <<<-markeringen soms niet betrouwbaar verwerkt worden. Controleer dus goed of T-Scan overslaat wat jij wilt laten overslaan. Is dat niet zo, probeer het dan nog eens met ###-markeringen.

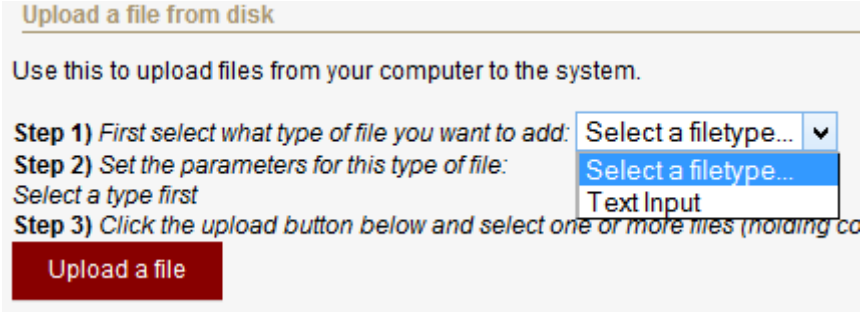
2.3 T-Scan starten; teksten en stoplijsten invoeren

1. Ga naar **http://webservices-lst.science.ru.nl** en kies voor *T-Scan.* Voer nu je gebruikersnaam en wachtwoord in.
   Tip: Ben je het webadres vergeten? Google dan op 'clam Radboud': de juiste link is dan het eerste resultaat in Google.

2. Maak een nieuw project aan. De projectnaam (*Project-ID*) mag geen spaties bevatten.

3. Nu kun je teksten invoeren in T-Scan. Daarvoor zijn twee manieren: je kunt de teksten uploaden (a) of je kunt je teksten inplakken of intikken in het scherm 'input tekst' (b).

a. **Vanaf harde schijf**. Als je teksten van je harde schijf wilt uploaden gebruik je het menu onder '*Upload a file from disk'*. Vergeet niet om je teksten in UTF-8 te coderen (Zie 2.2)!
Klik op het pijltje en kies voor 'Text Input'. Vervolgens klik je op 'Upload a file'. Nu kun je de bestanden kiezen die je wilt gebruiken.



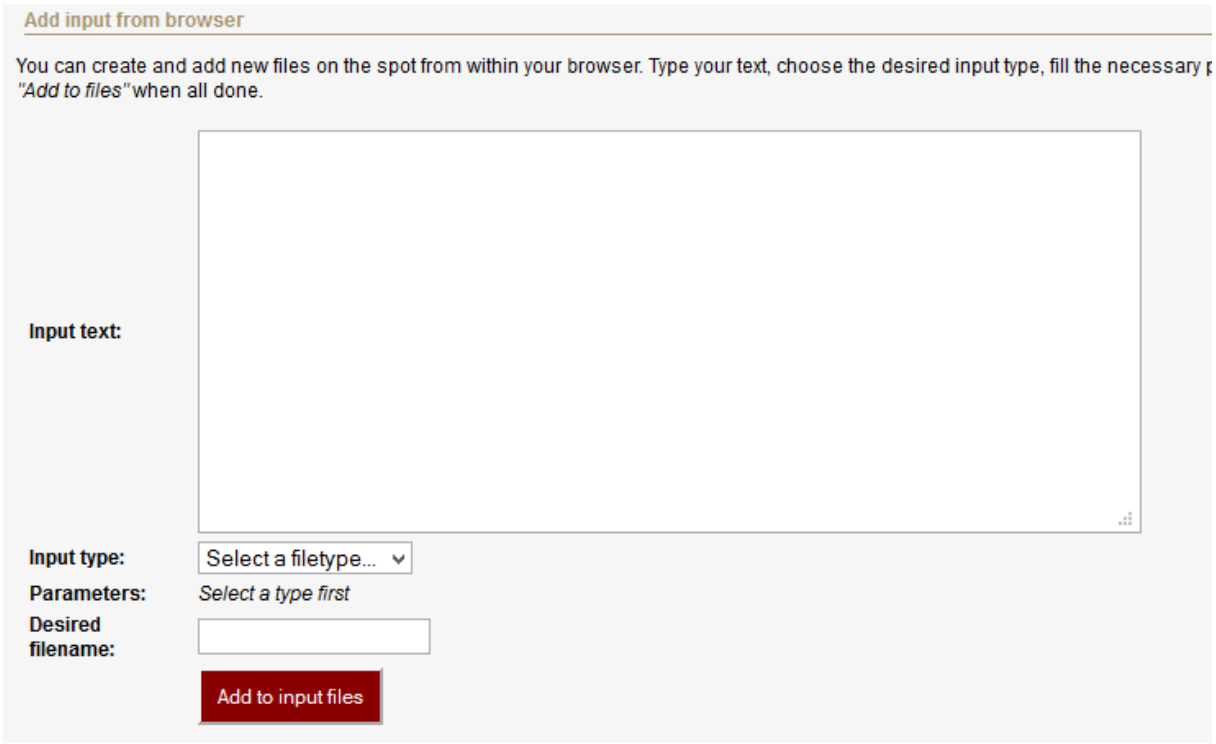
Wanneer je veel bestanden hebt, kun je ook een .zip-bestand uploaden. Dat doe je als volgt.

1. Selecteer de gewenste bestanden (dit moeten .txt bestanden zijn in UTF-8-indeling).
2. Klik met de rechtermuisknop op de selectie en kies voor *kopiëren naar > gecomprimeerde (gezipte) map.* (In oude versies van Windows bestaat deze optie niet. In dat geval moet je een apart programma gebruiken, zoals WinZip).
3. De bestanden staan nu in een gezipte map. Deze map kun je selecteren met 'upload a file' (zie bovenstaande afbeelding). T-Scan pakt ze uit en plaatst ze in de browser.

Je kunt naast een te analyseren tekst ook een stoplijst invoeren, dat wil zeggen een lijst met woorden die je bij je analyses graag buiten beschouwing wilt laten. Stel bijvoorbeeld dat jij teksten schrijft over de 'buurtteams' in een bepaalde gemeente. In die teksten zal vaak het woord *buurtteam* vallen. Je kunt eigenlijk niet om dat woord heen, en daarom wil je niet dat dit woord mee gaat tellen in de kenmerken wat betreft woordlengte, woordfrequentie en concreetheid. Zo'n woord neem je dan op in je stoplijst. Die lijst voer je in door als 'filetype' te kiezen voor *stoplist.* De stoplijst maak je het makkelijkst in Excel; je moet hem wel opslaan in het csv-format kiest (csv=comma separated values). Een stoplijst werkt door in de volgende tekstkenmerken:

- alle woordlengtes;
- alle frequentiekenmerken;
- alle kenmerken gebaseerd op semantische klassen van zelfstandige naamwoorden, bijvoeglijke naamwoorden en werkwoorden.

b. In de browser: Ga naar het gedeelte 'Add input from browser':



Bij 'Input type' kies je voor 'Text input'. In het grote scherm (*Input Text)* kun je teksten typen of plakken. Je kan direct vanuit Microsoft Word kopiëren, T-Scan codeert de teksten automatisch in UTF-8. Geef elke tekst en naam, en kies voor 'Add to input files'.

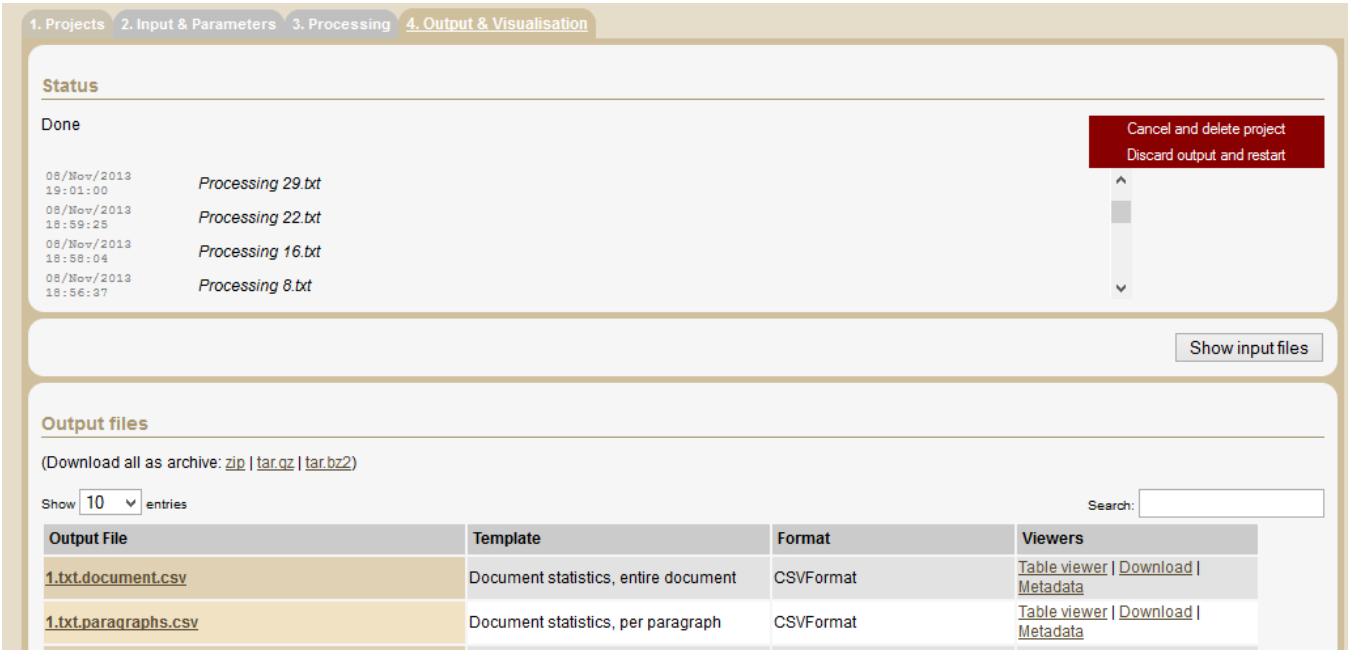
4. Stel nu de parameters (zeg maar: de instellingen) in voor je analyse.
    a. Overlap size; zie 3.5 bij 'bufferoverlap' voor een toelichting.

b. Frequency clipping; het gaat hier om het inkorten van de gebruikte woordfrequentielijsten om tijd te winnen; zie verder 3.3 bij woordfrequenties.
c. MTLD factor size; zie 3.5 bij MTLD voor een toelichting.
d. Use Alpino parser; het gaat hier om het al of niet gebruiken van Alpino. Standaard wordt Alpino gebruikt; we raden aan om dat zo te laten, omdat Alpino bij nogal wat kenmerken betrokken is.
e. Use Wopr; het gaat hier om het al of niet gebruiken van Wopr om probabiliteitskenmerken te berekenen, zie daarover verder 3.10.
f. Word frequency list; er is een keuze tussen verschillende frequentielijsten voor woordfrequenties, zie 3.3.
g. Lemma frequency list; er is een keuze tussen verschillende frequentielijsten voor lemmafrequenties, zie 3.3.
h. Top frequency list; er is een keuze tussen verschillende frequentielijsten om mee te bepalen hoeveel tekstwoorden behoren tot de meest frequente 1000, 2000, 3000, 5000, 10000 en 20000 woorden, zie 3.3.

5. Klik op *Start*.

2.4 T-Scanoutput verwerken

Tijdens het verwerken van de teksten kun je T-Scan gewoon wegklikken (of je computer afsluiten). De verwerkingstijd is afhankelijk van de hoeveelheid teksten, parameters en het aantal gebruikers op dat moment. Als T-Scan klaar is met verwerken, zie je dit scherm:



Onder *Status* zie je of de teksten succesvol verwerkt zijn (en wanneer dat was). Via *Show input files* kun je terugzien welke teksten je hebt ingevoerd voor analyse. Onder *Output files* zie je bestanden met resultaten. Je kunt de output bekijken op vier niveaus: document (de hele tekst), paragraaf (alinea, gescheiden door witregel), zin (gescheiden door punt) en woord (in het laatste geval zie je een andere, kleinere set kenmerken; zie hoofdstuk 4).

Bekijken in de browser

Als je op de naam van het resultatenfile klikt, of op *Table Viewer*, zie je de output in de browser. Die functie is geschikt als je snel wilt zien of de analyse is gelukt, of T-Scan bijvoorbeeld de zinnen correct heeft onderscheiden en of alle kenmerken ook echt waarden laten zien.
Je kunt ook binnen T-Scan een blik op de tekst werpen. Daartoe klik je op de XML-viewer. Je krijgt dan de tekst op het scherm, en kun door de muis over de woorden te bewegen de gedetailleerde POS-tags per woord bekijken (voor toelichting op die afkortingen, zie Van Eynde 2004). In het venster rechts worden een aantal tellingen weergegeven, waaronder het aantal woorden van de tekst. Op dit moment worden bij die tellingen nog oude variabelennamen gebruikt, dus raden we aan er nog geen gebruik van te maken.
Een laatste optie is *Metadata*. Als je daarop klikt, zie je met welke instellingen (parameters) je analyse heeft gewerkt.
Als de resultaten goed wilt gaan bekijken, is het beter om ze in te lezen in Excel of in SPSS. Om dat voor te bereiden, klik je eerst met je rechtermuisknop op 'download', en kies je voor *link opslaan.* Je slaat de resultaten nu op als *csv-file* (*comma separated values*).

Inlezen in Microsoft Excel

a. Ga naar *gegevens (Engels: data)* en kies voor *van tekst.*
b. Selecteer het csv-bestand dat je in T-Scan hebt opgeslagen.
c. Kies in Stap 1 van de Text Import Wizard voor *gescheiden* (separated) en bij File Origin voor UTF-8.
d. Kies in Stap 2 voor *komma* als scheidingsteken.
e. Stap 3 klik je op Advanced en verwijder je daar de punt als scheidingsteken voor duizendtallen. Vergeet je dat, dan krijg je rare getallen in je file, omdat Excel getallen met punten dan foutief gaat interpreteren.
f. Klik op *voltooien.*
g. Kies voor Existing Worksheet om de data direct te kunnen zien.

Inlezen in IBM SPSS

1. Het komt voor dat de getallen uit het csv-bestand niet meekomen naar SPSS. Meestal is dat een gevolg van de 'locale' waarin jouw SPSS gestart is. Open daarom een syntax-scherm en run eerst de volgende syntax om te zorgen dat de 'locale' voor SPSS de goede is: *SET LOCALE = 'en_US.windows-1252'*.
2. Open *File/Import data/csv-data.*
3. Zoek je csv-file en open dit.
4. Je krijgt een venster getiteld 'Read CSV file'.
5. Je laat het vinkje staan bij 'first line contains variable names',
6. Als 'delimiter between values' vul je in 'comma'.
7. Als 'decimal symbol' kies je 'period'.
8. Verder hoef je niets te doen. Je kunt verdergaan naar de Wizard, maar dit voegt niets toe, behalve dat je de kans krijgt om de invoerprocedure weer te geven als een lijst SPSS-commando's.

Controleren

Check of de analyse goed verlopen is, voordat je de resultaten echt gaat bekijken:

- Kijk of alle kenmerken daadwerkelijk waarden geven.
- Kijk of er kolommen zijn met rare namen als V[nr]. In dat geval bevat de datafile meer kolommen dan er namen waren, doordat er bepaalde waarden ten onrechte in meerdere kolommen zijn gesplitst. Zulke foutieve splitsingen kunnen het gevolg zijn van fouten rond komma's binnen kolommen.
- Kijk of T-Scan de zinnen heeft gescheiden op de punten waarop dat echt moet. Check bij rare zinssplitsingen nogmaals de hierboven genoemde punten.

- Kijk of alle variabelen als Numeric gedefinieerd zijn. SPSS maakt van variabelenkolommen waarin letters staan String-variabelen. Dus als ergens NA staat (wat betekent: not applicable) voor een waarde die niet berekend kon worden, wordt de hele variabele tot String gelabeld. Dat geeft problemen bij latere bewerkingen. Om dat te voorkomen kies je Numeric voor alle variabelen.
- Ook kan het voor de leesbaarheid handig zijn om het aantal decimalen aan te passen. Doe dat nadat je alles op Numeric hebt gezet, en nadat je hebt gekeken of er overal voldoende ruimte is voor je decimalen (bij 'Width').

3. T-Scankenmerken op zins-, paragraaf- en tekstniveau

3.1 Kenmerkgroepen, kenmerktypen en tekstregio's

We onderscheiden een algemene (0) en acht specifieke *kenmerkgroepen* (1-8):

0. Algemeen
1. Woordmoeilijkheid
2. Zinscomplexiteit
3. Referentiële coherentie en woordenrijkdom
4. Relationele coherentie
5. Semantische klassen en woordconcreetheid
6. Persoonlijke elementen
7. Andere informatie over woorden en uitdrukkingen
    a. Namen
    b. Werkwoordkenmerken
    c. POS-tags
    d. Afkortingen
    e. Voorzetseluitdrukkingen
    f. Overig
8. Probabiliteitsmaten

Naar hun berekeningswijzen kunnen kenmerken worden onderscheiden in vier *typen*:

- *Aantallen.* Hier gaat het om aantallen, zo nodig gemiddeld over de tekstregio. De getelde eenheid wordt duidelijk uit de naam van het kenmerk. Voorbeelden:
    - Letters per woord
    - Woorden per zin
    - Afhankelijkheidslengtes
- *Aantallen per deelzin (_dz).* Voor sommige aantallen is het nuttig om die niet alleen per zin te hebben, maar ook per deelzin. De naam van deze kenmerken, die zelf weer gemiddeld kunnen worden over de tekstregio, eindigt altijd op '_dz'.
- *Proporties (_p)*. Bij proporties gaat het om een deling, waarbij een aantal wordt gedeeld op een referentiegroep. Voorbeelden:
    - De proportie tegenwoordige tijds-vormen op het totaal aantal persoonsvormen
    - De proportie strikt concrete bijvoeglijke naamwoorden op het totaal aantal bijvoeglijke naamwoorden.
- *Dichtheden (_d)*. Een dichtheid standaardiseert de frequentie van een verschijnsel op 1.000 woorden. Als bijvoorbeeld een tekstje op 10 zelfstandige naamwoorden 5 strikte concrete naamwoorden telt, is de dichtheid daarvan 500.

Deze kenmerken kunnen worden bekeken in vier *tekstregio's*:

- Woordniveau
- Zinsniveau
- Paragraafniveau; paragrafen worden voor T-Scan onderscheiden door witregels (dus door twee harde returns)
- Tekstniveau

Er zijn bijna 400 kenmerken op de hogere tekstniveaus. We behandelen ze in dit hoofdstuk per groep. In hoofdstuk 4 bespreken we het veel kleinere aantal kenmerken dat op woordniveau geleverd wordt.

3.2 Algemene kenmerken

T-Scan kent op tekstniveau de volgende algemene kenmerken:

1a	Inputfile	Naam van de ingevoerde tekstfile
1b	Segment	Nummer van de zin en/of de alinea waarop de resultaten betrekking hebben
2.	Par_per_doc	Het aantal alinea's in de tekst
3.	Zin_per_doc	Het aantal zinnen in de tekst
4.	Word_per_doc	Het aantal woorden in de tekst
5.	Alpino_status	Meestal: het aantal zinnen dat Alpino-parser niet heeft kunnen ontleden

'Inputfile' spreekt voor zich. Bij 'segment' vind je in de output op zinsniveau het nummer van de zin en de alinea waarvoor de waarden gelden, en in de paragraaf-output alleen het nummer van de alinea. In de output op tekstniveau ontbreekt dit kenmerk.

Vervolgens geeft T-Scan het aantal alinea's, zinnen en woorden van het document. Op alineaniveau vind je het aantal zinnen en woorden voor deze specifieke alinea, niet per tekst. Daarbij worden alinea's gedefinieerd via witregels: een regelsprong is niet voldoende voor een alineagrens. Voor het aantal woorden tellen we leestekens natuurlijk niet mee, al zullen we later daar wel een dichtheid van tegenkomen (zie Woordsoorten). Op zinsniveau vind je het aantal woorden pas later, bij de kenmerken voor zinscomplexiteit. Het meest lastig is de definitie van zinnen, dus gaan we daar uitgebreider op in 3.2.1.

Ten slotte is er informatie over de status van Alpino. Alpino is de ontleedmachine die T-Scan basisinformatie levert voor een behoorlijk aantal kenmerken. Alpino kan echter worden uitgeschakeld wanneer de Alpino-kenmerken niet nodig zijn. Zonder Alpino werkt T-Scan wat sneller. De kenmerkwaarden zijn:

'-1' = Alpino was uitgeschakeld door gebruiker, zin is niet ontleed.
'0' = Alpino heeft zin zonder problemen ontleed;
'1 of hoger' = het aantal zinnen dat Alpino niet heeft kunnen ontleden.

De laatste waarde kan het gevolg zijn van symbolen waarmee T-Scan niet kan omgaan, zoals vierkante haken. Soms echter is de zin domweg te complex. Een voorbeeld van een zin die Alpino niet heeft kunnen ontleden is de volgende zin uit een verzekeringspolis:

```
Niet gedekt is schade die is veroorzaakt met opzet van een verzekerde,
tijdens deelneming aan snelheidswedstrijden of -ritten, tijdens deelneming aan
behendigheidswedstrijden of -ritten geheel of gedeeltelijk buiten Nederland, tijdens verhuur
van het motorrijtuig, tijdens het beroepsmatig vervoeren van personen of van zaken, waaronder
gevaarlijke of milieuverontreinigende stoffen, waarvoor een wettelijke vergunning is vereist.
```

Alpino kan ook problemen krijgen bij zinnen waarin genummerde opsommingen voorkomen.

Wanneer Alpino_status de waarde '1' heeft wordt op zinsniveau worden dan geen waarde gegeven voor kenmerken die een beroep doen op Alpino. Vervelender is, dat dit ook geldt voor een aantal maten op paragraaf- en tekstniveau. In zulke gevallen sta je voor de keuze: de zin verwijderen of handmatig de ontbrekende gegevens in je data opnemen. Er is ook nog de optie om de zin subtiel te vereenvoudigen; daarmee verander je je tekst, maar dat kan nuttig zijn om andere kenmerken van de zin voor je data te behouden, of die betrouwbaarder te kunnen analyseren.

Op zinsniveau vind je naast de bovengenoemde algemene kenmerken nog de zin zelf waarover de betreffende dataregel gaat. Op woordniveau treffen we bij de algemene kenmerken weer inputfile en segment (woordnummer), en daarnaast nog het woord, het lemma, en de morfemen. Zie over de woorddata verder hoofdstuk 4.

3.2.1 Wat is een zin?

Een tool voor automatische tekstanalyse heeft een eenvoudige definitie nodig van zinsgrenzen. In T-Scan worden de volgende leestekens beschouwd als zinsscheiding:

- punten;
- vraagtekens;
- uitroeptekens;
- puntkomma's.

Er was een dilemma rond haakjes, dubbele punten, en puntkomma's. Op dit moment wordt de puntkomma als zinsscheiding beschouwd, en worden dubbele punten en haakjes niet. De vraag is, welke vertekening deze keuzes met zich meebrengen. Om dat na te gaan, zijn handmatige analyses gedaan op zinnen met deze leestekens.

Om te beginnen is een steekproef getrokken van ongeveer 1000 dubbele punten. Daartoe moesten bijna 24000 zinnen worden doorzocht; ruim 4% van de zinnen bevat dus een dubbele punt. In die zinnen is gekeken naar de coherentierelatie tussen de zinsdelen voor en na de dubbele punt. De resultaten daarvan zijn samengevat in de tweede en derde kolom van Tabel 1. Daaruit blijkt dat de dubbele punt meestal gevolgd wordt door een voltooiing van een gedachte die wordt aangekondigd voor de dubbele punt. Daarbij gaat het regelmatig om een invulling (zie Pander Maat 2002) van een concreet of algemeen nomen (zie voorbeeld 1 en 2), een uitwerking van een propositie (zie 3), van een kwantiteit (zie 4), van een ontbrekend element (zie 5), van het antwoord op een ingebedde vraag (zie 6), of van een pronomen (zie 7). Verder kan na de dubbele punt de betekenis van een term volgen, of andersom de term die bij een betekenis hoort (zie 8 en 9); verder staat vaak na de dubbele punt de gedachte of uitspraak die ervoor is ingeleid (zie 10 en 11). Ten slotte staat voor de dubbele punt soms een connectief (zie 12), of de doelgroep van de uiting (zie 13).

1. Gelukkig had ik leuke reisgenoten om mee te kletsen tijdens m'n vlucht: een meisje uit Engeland en een Keniaanse man.
2. Langzamerhand ontvouwde zich voor mijn ogen een patroon: Lotte had een zwak voor mannen met aanzien.
3. De dagen erna was dat wel anders: de vonk sloeg over.
4. Die was prijzig: 85 dollar per persoon.
5. Ze gooit zelden iets weg, ze blijft repareren: schoenen, hemden, jurken, lakens, sokken, wekkerradio's.
6. Bij hem bleef een bips wat het uiteindelijk toch ook is: een kont.
7. Maar het klinkt wel leuk: een picknick.
8. Daarom nam hij ambtenaren in dienst: mensen die de farao helpen bij het besturen van de staat.
9. Je moet dus kiezen wat je het belangrijkste vindt: dat heet prioriteiten stellen.
10. Ze denkt: ik word omringd door ezels en idioten.
11. Hij zei tegen mijn cliënt: je moet kunnen aantonen dat je vermogend bent.
12. En bovendien: ze stond er niet alleen voor.
13. Jonger dan zestien jaar: piercing alleen onder begeleiding.

Deze relaties van voltooiing (nummers 1-12 in Tabel 1) zijn aan de orde bij 94% van de dubbele punten. De enige uitzonderingen zijn de voor- en achterwaartse causale relaties: zie (14) en (15).

14. Omdat sneeuw het zonlicht goed reflecteert, zal er dan minder zonne-energie door de aarde worden opgenomen: de temperatuur zal nog verder dalen.
15. Eigenlijk hoef ik me het niet eens af te vragen: ik weet het zo wel.

	Dubbele punten		Puntkomma's	
Coherentierelatie	**Frequentie**	**%**	**Frequentie**	**%**
1. invulling concreet nomen	150	14.9	9	2.7
2. invulling algemeen nomen	89	8.9	2	0.6
3. topicaankondiging	10	1.0	3	0.9
4. uitwerking propositie	74	7.4	32	10.0
5. uitwerking adjectief/kwantiteit	37	3.7	3	0.9
6. invulling open/ontbrekend element	41	4.1		
7. invulling pronomen	49	4.9	5	1.5
8. aangekondigde gedachte/observatie	44	4.4	1	0.3
9. aangekondigde uitspraak	361	36.0		
10. term-betekenis of betekenis-term	38	3.8	4	1.2
11. connectief / meta-commentaar	45	4.5		
12. doelgroep van uiting	2	0.2		
13. oorzaak-reden-argument	58	5.8	41	12.4
14. gevolg-doel-conclusie	6	0.6	17	5.1
15. wat voorafgaat			1	0.3
16. wat volgt			31	9.4
17. voortgezette beschrijving			69	20.8
18. elementen in reeks			67	20.2
19. evaluatie			9	2.7
20. contrast			27	8.2
21. versterking			3	0.9
22. bron ('zie X')			3	0.9
23. overig			3	0.9
Totaal	1004	100%	331	100%

Tabel 1. Coherentierelaties rondom dubbele punten en puntkomma's

Wat betreft puntkomma's, die komen veel minder voor dan dubbele punten: in de 40372 zinnen komen zij 331 maal voor, dus in minder dan 1% van de gevallen. Slechts een klein gedeelte van de relaties waarmee zij gepaard gaan, valt onder de voltooiingsrelaties (18%). Voorbeelden zijn hieronder 16, een invulling van een concreet nomen, en 17, een uitwerking van een propositie. Aanzienlijk frequenter zijn coherentierelaties waarin beide elementen op zichzelf geconceptualiseerd kunnen worden (de nummers 13-23 in Tabel 1: 82%). De meest frequente daarvan zijn achterwaarts-causale relaties (zie 18), voorwaarts-chronologische relaties (zie 19), voortzettingen van beschrijvingen rond een entiteit (zie 20) en voortzettingen van een reeks (zie 21), en zelfs contrasten (zie 22).

16. Het staat in Gérardmer; een helder, middeltuttig Frans toeristenstadje in de Vogezen.
17. Mombasa is veel armer dan ik had verwacht van zo'n grote stad; heel veel kleine hutjes en kraampjes en mensen die langs de weg tussen het afval zitten.
18. Ik heb besloten dat ik per november toch weer op mezelf ga wonen; na drie jaar een eigen plek in Tilburg te hebben gehad merk ik toch wel dat ik dat begin te missen.
19. Deze week worden de contracten opgesteld; per 28 oktober verhuis ik.

20. Ze waren ongeveer zes jaar ouder dan ik, achter in de twintig; het gezicht van de langste man was hoekig en bleek, dat van zijn metgezel rond en donker, Nubisch bijna.
21. Of er bestaat ergens in hun hoofd het vage idee dat ze naar de bioscoop willen; of ze overwegen om maar te blijven en langzaam door te zakken.
22. Het was echt een schatje; al liep hij wel een beetje raar.

Dit alles betekent dat de beslissing om dubbele punten niet als zinsscheiding op te vatten en puntkomma's wel grosso modo een goede is: deze leidt er enerzijds toe dat T-Scan de zinsgrens meestal uitstelt als een gedachte nog onvoltooid is, maar deze meestal wel plaatst tussen elementen die zelfstandig interpreteerbaar zijn. Ook in de 5% zinnen met dubbele punten en puntkomma's blijft onze zinsdefinitie een redelijk valide operationalisatie van 'een enkele voltooide gedachte'.

Ten slotte een opmerking over passages tussen haakjes. Omdat het hier veelal om invoegingen midden in de zin gaat, is een zinsscheiding hier niet mogelijk. Daarom is afgezien van een coherentieanalyse van fragmenten tussen haakjes. Wel is nader gekeken naar een genrespecifiek gebruik van haakjes, dat onze resultaten kan vertekenen: het tussen haakjes verwijzen naar bronnen in onderzoeksartikelen. Een voorbeeld van een omvangrijke verwijzing is 23.

23. Toch blijft de contacthypothese aantrekkelijk voor wetenschappers en wordt zij gebruikt in recent onderzoek naar stereotypering van bijvoorbeeld moslims (Novotny & Polonsky, 2011, Savelkoul, Scheepers, Tolsma & Hagendoorn, 2010).

Omdat T-Scan de ampersand als woord telt, is zin 23 door de bronvermelding 29 woorden lang, in plaats van 19. Omdat bronvermeldingen frequent zijn in onderzoeksartikelen, zou de zinslengte in dit genre systematisch overschat kunnen worden. Daarom is voor 50 van de 100 fragmenten uit onderzoeksartikelen nagegaan hoeveel woorden de zinnen met haakjesinvoegingen zouden tellen zonder deze invoegingen. Van de 653 zinnen in deze fragmenten bevatten er 115 bronvermeldingen tussen haakjes. Wanneer in die zinnen de invoegingen buiten beschouwing blijven, daalt de gemiddelde lengte van die zinnen van 29.0 naar 22.9 woorden. Met deze correctie daalt de gemiddelde lengte over alle zinnen van 23.4 naar 22.4 woorden. Er is dus inderdaad een lichte overschatting van de zinslengte in onderzoeksartikelen, aangenomen dat lezers bij het verwerken van de zin de bronvermeldingen (al of niet tijdelijk) buiten beschouwing laten.

3.3 Woordmoeilijkheid

6.	Let_per_wrd	Letters per woord
7.	Wrd_per_let	Woorden per letter
8.	Let_per_wrd_zn	Letters per woord, zonder namen
9.	Wrd_per_let_zn	Woorden per letter, zonder namen
10.	Morf_per_wrd	Morfemen per woord
11.	Wrd_per_morf	Woorden per morfeem
12.	Morf_per_wrd_zn	Morfemen per woord, zonder namen
13.	Wrd_per_morf_zn	Woorden per morfeem, zonder namen
14.	Namen_p	Proportie van namen op zelfstandige naamwoorden plus namen
15.	Namen_d	Dichtheid van namen
16.	Wrd_prev	De prevalentie (bekendheid) van de tekstwoorden
17.	Wrd_prev_z	De prevalentie (bekendheid) van de tekstwoorden, z-score
18.	Inhwrd_prev	De prevalentie (bekendheid) van de inhoudswoorden in de tekst
19.	Inhwrd_prev_z	De prevalentie (bekendheid) van de inhoudswoorden, z-score
20.	Dekking_inhwrd_prev	De proportie inhoudswoorden met een prevalentiescore
21.	Freq50_Staph	De proportie woorden die in de Staphorsius-frequentielijst 50% van de meest frequente woordtokens uitmaken
22.	Freq65_Staph	Idem maar nu gaat het om 65% van de woordtokens
23.	Freq77_Staph	Idem maar nu gaat het om 77% van de woordtokens
24.	Freq80_Staph	Idem maar nu gaat het om 80% van de woordtokens
25.	Wrd_freq_log	Woordfrequentie, logaritme
26.	Wrd_freq_zn_log	Woordfrequentie zonder namen, logaritme
27.	Lem_freq_log	Lemmafrequentie, logaritme
28.	Lem_freq_zn_log	Lemmafrequentie zonder namen, logaritme
29.	Wrd_freq_log_zonder_abw	Woordfrequentie, logaritme (algemene bijwoorden uitgezonderd)
30.	Wrd_freq_zn_log_zonder_abw	Woordfrequentie zonder namen, logaritme (zonder bijwoorden)
31.	Lem_freq_log_zonder_abw	Lemmafrequentie, logaritme (zonder bijwoorden)
32.	Lem_freq_zn_log_zonder_abw	Lemmafrequentie zonder namen, logaritme (zonder bijwoorden)
33.	Freq1000	De proportie tekstwoorden horend bij de meest frequente 1000 woorden
34.	Freq2000	Idem voor de meest frequente 2000 woorden
35.	Freq3000	Idem voor de meest frequente 3000 woorden
36.	Freq5000	Idem voor de meest frequente 5000 woorden
37.	Freq10000	Idem voor de meest frequente 10000 woorden
38.	Freq20000	Idem voor de meest frequente 20000 woorden
39.	Freq1000_inhwrd	De proportie inhoudswoorden horend bij de meest frequente 1000 woorden
40.	Freq2000_inhwrd	Idem voor de meest frequente 2000 woorden
41.	Freq3000_inhwrd	Idem voor de meest frequente 3000 woorden
42.	Freq5000_inhwrd	Idem voor de meest frequente 5000 woorden
43.	Freq10000_inhwrd	Idem voor de meest frequente 10000 woorden
44.	Freq20000_inhwrd	Idem voor de meest frequente 20000 woorden
45.	Freq1000_inhwrd_zonder_abw	Freq_1000_inhwrd zonder algemene bijwoorden
46.	Freq2000_inhwrd_zonder_abw	Freq_2000_inhwrd zonder algemene bijwoorden
47.	Freq3000_inhwrd_zonder_abw	Freq_3000_inhwrd zonder algemene bijwoorden
48.	Freq5000_inhwrd_zonder_abw	Freq_5000_inhwrd zonder algemene bijwoorden
49.	Freq10000_inhwrd_zonder_abw	Freq_10000_inhwrd zonder algemene bijwoorden
50.	Freq20000_inhwrd_zonder_abw	Freq_20000_inhwrd zonder algemene bijwoorden
51.	Samenst_d	Dichtheid compositionele nominale samenstellingen
52.	Samenst_p	Proportie samenstellingen op de naamwoorden
53.	Samenst3_d	Dichtheid drie- en meerdelige samenstellingen
54.	Samenst3_p	Proportie drie- en meerdelige samenstellingen op naamwoorden
55.	Let_per_wrd_nw	Woordlengte in letters voor de naamwoorden in de tekst
56.	Let_per_wrd_nsam	Woordlengte in letters voor nomina die geen samenstelling zijn
57.	Let_per_wrd_sam	Woordlengte in letters voor de nominale samenstellingen
58.	Let_per_wrd_hfdwrd	Woordlengte in letters voor het hoofdwoord daarvan
59.	Let_per_wrd_satwrd	Woordlengte in letters voor het satellietwoord daarvan
60.	Let_per_wrd_nw_corr	Gecorrigeerde naamwoordlengte (correctie: voor samenstellingen geldt de hoofdwoordlengte in plaats van de woordlengte)
61.	Let_per_wrd_corr	Gecorrigeerde woordlengte (voor samenstellingen geldt de

		hoofdwoordlengte in plaats van de woordlengte)
62.	Wrd_freq_log_nw	Woordfrequentie (logaritme) van de naamwoorden in de tekst
63.	Wrd_freq_log_ong_nw	Woordfrequentie (logaritme) van de niet-samenstellingen
64.	Wrd_freq_log_sam_nw	Woordfrequentie (logaritme) van de nominale samenstellingen
65.	Wrd_freq_log_hfdwrd	Woordfrequentie (logaritme) van de hoofdwoorden in de samenstellingen
66.	Wrd_freq_log_satwrd	Woordfrequentie (logaritme) van de satellietwoorden in de samenstellingen
67.	Wrd_freq_log_(hfd_sat)	Gemiddelde van de logaritmen van de woordfrequentie van hoofdwoorden en satellietwoorden in de samenstellingen
68.	Wrd_freq_log_nw_corr	Gecorrigeerde naamwoordfrequentie (voor samenstellingen geldt de hoofdwoordfrequentie in plaats van de woordfrequentie)
69.	Wrd_freq_log_corr	Gecorrigeerde woordfrequentie (logaritme), waarbij voor samenstellingen de hoofdwoordfrequentie genomen wordt
70.	Wrd_freq_log_zn_corr	Gecorrigeerde woordfrequentie zonder namen (logaritme), waarbij voor samenstellingen de hoofdwoordfrequentie genomen wordt
71.	Wrd_freq_log_corr_zonder_abw	Gecorrigeerde woordfrequentie zonder algemene bijwoorden
72.	Wrd_freq_log_zn_corr_zonder_abw	Gecorrigeerde woordfrequentie zonder namen en zonder algemene bijwoorden
73.	Freq1000_nw	Proportie naamwoorden horend bij de meest frequente 1000 woorden
74.	Freq5000_nw	Idem voor de meest frequente 5000 woorden
75.	Freq20000_nw	Idem voor de meest frequente 20000 woorden
76.	Freq1000_nsam_nw	Proportie van de niet-samenstellingen die hoort bij de meest frequente 1000 woorden
77.	Freq5000_nsam_nw	Idem voor de meest frequente 5000 woorden
78.	Freq20000_nsam_nw	Idem voor de meest frequente 20000 woorden
79.	Freq1000_sam_nw	Proportie nominale samenstellingen horend bij de meest frequente 1000 woorden
80.	Freq5000_sam_nw	Idem voor de meest frequente 5000 woorden
81.	Freq20000_sam_nw	Idem voor de meest frequente 20000 woorden
82.	Freq1000_hfdwrd_nw	Proportie hoofdwoorden van nominale samenstellingen horend bij de meest frequente 1000 woorden
83.	Freq5000_hfdwrd_nw	Idem voor de meest frequente 5000 woorden
84.	Freq20000_hfdwrd_nw	Idem voor de meest frequente 20000 woorden
85.	Freq1000_hfdwrd_nw	Proportie satellietwoorden van nominale samenstellingen horend bij de meest frequente 1000 woorden
86.	Freq5000_hfdwrd_nw	Idem voor de meest frequente 5000 woorden
87.	Freq20000_hfdwrd_nw	Idem voor de meest frequente 20000 woorden
88.	Freq1000_nw_corr	Gecorrigeerde proportie naamwoorden horend bij de meest frequente 1000 woorden (voor samenstellingen geldt de hoofdwoordfrequentie in plaats van de woordfrequentie)
89.	Freq5000_nw_corr	Idem voor de meest frequente 5000 woorden
90.	Freq20000_nw_corr	Idem voor de meest frequente 20000 woorden
91.	Freq1000_corr	Gecorrigeerde proportie woorden horend bij de meest frequente 1000 woorden (voor samenstellingen geldt de hoofdwoordfrequentie)
92.	Freq5000_corr	Idem voor de meest frequente 5000 woorden
93.	Freq20000_corr	Idem voor de meest frequente 20000 woorden

Woordlengtes en namen

De woordlengte in letters (4) spreekt voor zich. De inverse van dit kenmerk is het aantal woorden per letter (5). Het is denkbaar dat dit kenmerk beter correleert met bijvoorbeeld het tekstbegrip, omdat het een ander verloop kent: het aantal letters per woord stijgt monotoon wanneer je letters toevoegt. Het aantal woorden per letter stijgt steeds trager wanneer je dat doet.

T-Scan geeft de woordlengte ook in morfemen, vanuit de gedachte dat dit de eigenlijk betekenisdragende eenheden zijn, en niet de letters.

Verder geeft T-Scan de dichtheid van namen en de proportie van namen op het geheel van namen en naamwoorden weer. Het aantal namen in een tekst is interessant omdat namen anders dan

zelfstandige naamwoorden een beroep doen op vrij specifieke voorkennis. De herkenning van namen in T-Scan leunt vrij sterk op de aanwezigheid van hoofdletters. Daarom is voorzichtigheid nodig in het aanbieden van teksten met een afwijkend hoofdlettergebruik. In sommige juridische contexten is het bijvoorbeeld gebruik om allerlei termen van een hoofdletter te voorzien. Dat kan in T-Scan leiden tot een overschatting van het aantal namen in de tekst. Teksten met afwijkend hoofdlettergebruik kunnen daarom beter vooraf 'genormaliseerd' worden.

Woordprevalenties

Naarmate lezers de woorden van een tekst kennen, stijgt hun begrip (Schmitt et al. 2011). De bekendheid van woorden kan direct en indirect worden geschat. In de directe benadering wordt aan taalgebruikers gevraagd of zij een woord kennen. Een grootschalige studie op dat punt is gedaan door Keuleers et al. 2015. Zij legden in een webexperiment bestaande woorden (ongeveer 53.000) en niet-bestaande woorden (ongeveer 21.000) voor aan Belgische en Nederlandse proefpersonen. De niet-bestaande woorden dienden om de proefpersonen niet te vaak 'ja' te laten zeggen. De proportie mensen die zegt dat een bestaand woord een Nederlands woord is, wordt door de auteurs de 'prevalentie' genoemd.

Op http://crr.ugent.be/archives/1755 is een lijst te vinden met prevalentiescores voor 54.320 woorden. Het aantal niet-bestaande woorden in die lijst is niet helemaal duidelijk. Onder de minst bekende bestaande woorden in de lijst zijn *poujadisme* en *pueriel* met prevalenties van .09 resp. .11. *Hominem* scoort .35, *interetnisch* .50, *regentesk* .60, *hekeldicht* .70, *secularisering* .80, *wensouder* .85, *devotie* .90, *inkleden, applaudisseren* en *operette* . 95, *overmoed, burgervader* en *trendwatcher* .97, *crediteur, redelijkheid* en *doorschieten* .99, *uitrusting, overkapping* en *radiosignaal* .995, en *verbetering, bestelwagen* en *rusten* 1.00 (afgerond). Daarmee is duidelijk dat er heel veel woorden hoog scoren: bijna 30.000 woorden scoren boven de 95%, zo'n 26.000 woorden boven de 97%.

Er zijn prevalentiescores voor Nederlandse en voor Belgische proefpersonen. T-Scan geeft standaard de scores voor Nederlanders. Er is voor gekozen om naast de woordprevalentie ook die van inhoudswoorden te geven, omdat functiewoorden in het algemeen overbekend zijn. Naast de prevalentieproportie zelf wordt ook de bijbehorende z-gegeven, die varieert van -1.32 voor *poujadisme* tot 3.42 voor *rusten*. Ten slotte wordt bij 'dekking' gemeld hoeveel van de inhoudswoorden terug te vinden waren in lijst met prevalentiescores.

Woordfrequenties

Een meer indirecte manier om de kans te schatten dat de lezer een woord kent is via woordfrequenties. Er zijn drie soorten woordfrequentiegegevens.

Het eerste type is gebaseerd op een woordfrequentielijst die Staphorsius (1994) in de jaren '80 van de vorige eeuw samenstelde op basis van lectuur voor kinderen in de basisschoolleeftijd. Net als Staphorsius deed ten behoeve van de CLIB, deelt T-Scan die lijst op verschillende manieren in tweeën. In de eerste verdeling wordt de streep getrokken na 50% van de woordtokens, en gelden de woorden boven de streep als 'frequent' en die eronder als 'minder frequent'. Vervolgens gaat T-Scan na hoeveel van de tekst woorden boven en hoeveel er onder de streep staan. Die proportie wordt 'Freq50' genoemd. Hetzelfde wordt gedaan voor andere grenzen (resp. 65%, 77% en 80% van de woordtokens). Deze kenmerken geven aan in welke mate de tekstwoorden vertrouwd zouden kunnen zijn voor basisschoolkinderen.

We tekenen hierbij aan dat het Staphorsius-corpus aan de oude kant is. Inmiddels is er een beter corpus van dezelfde aard, het Basilex-corpus (voor meer informatie, zie http://tst-centrale.org/nl/producten/lexica/basilex-lexicon/7-159). Dat corpus zou in de toekomst wellicht in T-Scan verwerkt kunnen worden.

Het tweede type gegevens geeft de exacte frequentie aan per woord (zie kenmerk 20-23), waarbij we ons beperken tot inhoudswoorden (naamwoorden, namen, adjectieven, bijwoorden en 'gewone werkwoorden', dat wil zeggen werkwoorden die geen hulpwerkwoord of koppelwerkwoord zijn of

kunnen zijn). Daarbij is de logaritme (grondtal 10) genomen van de frequentie, zodat een woord met een frequentie van een miljoen niet duizend keer zo makkelijk is als een woord met een frequentie van 100, maar drie keer. Namen worden volgens Camblin et al. (2007) anders verwerkt dan 'gewone woorden'. Daarom zijn een aantal varianten gedefinieerd waarbij namen worden overgeslagen.

Op voorstel van Van Heuven et al. (2014) werken we met de logaritme (grondtal 10) van de relatieve frequenties, waarbij de frequentie wordt gestandaardiseerd op een miljard woorden. Dat heeft het voordeel dat ook bij lage frequenties nog onderscheid gemaakt kan worden. Bijvoorbeeld, wanneer een woord eenmaal voorkomt in een corpus van 1 miljoen woorden, bedraagt de gestandaardiseerde frequentie 1000, en de logaritme dus 3. Deze eenheid wordt door Van Heuven et al. de Zipf-schaal genoemd.

Het derde type woordfrequentiegegevens geeft net als het eerste type aan welke proportie tekstwoorden als frequent kan worden gedefinieerd. Maar nu wordt gewerkt met hedendaagse corpora met 'volwassen' taalgebruik, en wordt het al of niet frequent zijn anders bepaald. Bij bijvoorbeeld *Freq1000* wordt simpelweg gekeken hoeveel van de tekstwoorden horen tot de 'top1000' in de frequentielijst gebaseerd op een corpus. Bij *Freq1000_inhwrd* gaat het om dezelfde proportie, maar dan wordt alleen gekeken naar inhoudswoorden, dat wil zeggen naamwoorden, namen, adjectieven, bijwoorden en 'gewone werkwoorden', dat wil zeggen werkwoorden die geen hulpwerkwoord of koppelwerkwoord zijn of kunnen zijn.

Nu is er discussie mogelijk over de vraag of bijwoorden bij de inhoudswoorden gerekend moeten worden. Daarom zijn er voor alle variabelen betreffende inhoudswoorden twee varianten gemaakt. De ongemarkeerde variant includeert bijwoorden; de variant die eindigt op *zonder_abw* laat algemene bijwoorden buiten beschouwing. In 3.7.4 is uitgelegd dat het daarbij gaat om verreweg de meeste van de 900 bijwoorden uit de lijst waar T-Scan mee werkt.

Daarbij zijn er drie keuzes te maken voor het onderliggende corpus:

- SoNaR totaal (Oostdijk et al. 2013; voor onderzoekers is dit corpus toegankelijk op https://portal.clarin.inl.nl/opensonar_whitelab/page/home?lang=nl);
- SoNaR, subcorpus kranten;
- Subtlex (Keuleers et al. 2010;).

Bij SoNaR gaat het vooral om schriftelijk taalgebruik, waarbij informele genres qua omvang in de minderheid zijn. Subtlex daarentegen is een corpus met Nederlandse ondertitels voor films, en bevat met name alledaagse (zij het niet-spontane) conversatie.

Elke frequentielijst kent slordigheden en eigenaardigheden. Wat betreft slordigheden, de Subtlex-lijst bevat woordcombinaties (*de_tijd, niet_alleen*), items met een punt erachter ('al.', 'goed.'), inleesfouten waarin een 'l' in plaats van een 'i' gelezen is, en natuurlijk spelfouten, die soms onherkenbare woorden opleveren. En de SoNaR-lijst kent ook items die geen woord zijn, en die door tokenisatiefouten in de lijst terecht zijn gekomen. Deze slordigheden zijn zoveel mogelijk handmatig verwijderd uit de lijsten die gebruikt worden voor de Freq1000 … 20000. Die woorden zijn immers betrokken bij de Freq-maten. We willen graag dat het bij die 20000 woorden gaat om 'echte' woorden die kunnen staan voor een basaal Nederlands vocabulaire.

Wat betreft eigenaardigheden, de keuze van het corpus bepaalt wat frequent is. Het Subtlex-corpus komt uit ondertitels bij Engelse en Amerikaanse films en series. Daarom bevat het duizenden Engelse namen (vooral persoons-, maar ook geografische namen en woorden zoals Thanksgiving), Engelse en Spaanse aanspreekvormen (mrs., signor) en ook onvertaald gebleven Engelse woorden. Ook deze Engelse elementen zijn niet wenselijk als we het corpus willen gebruiken als representatie van Nederlands (informeel) taalgebruik. Daarom is het corpus handmatig geschoond waar het gaat om de lijst van de 20000 meest frequente woorden. Voor de de uiteindelijk overblijvende top-20000 woorden zijn bijna 26000 Subtlex-woorden gebruikt.

Voor de 20000-woordenlijst gebaseerd op het SoNaR-totaalcorpus is een opschoning in twee stappen uitgevoerd. Eerst zijn niet-bestaande en buitenlandse woorden verwijderd. Vervolgens is gekeken naar de duizenden namen (plaatsnamen, persoonsnamen, organisatienamen). Omdat Sonar voor

bijna 80% bestaat uit Belgische teksten, zijn die namen nogal Zuid-Nederlands gekleurd. We achten het aannemelijk dat die namen geen goede indruk geven van de vertrouwdheid van het tekstvocabulaire voor Noord-Nederlanders.

Maar ook meer principieel is het kwestieus of namen horen bij het basisvocabulaire. Voor persoonsnamen geldt dat sowieso. Voor geografische en organisatienamen zijn wellicht een handvol namen van nationale betekenis (*Nederland, België, Amsterdam, Brussel*). Voor andere namen geldt dat het nogal toevallig is welke ervan in de tekst voorkomen. Daarom zijn ook uit de Sonar-totaallijst van 20000 woorden ongeveer 4800 namen vervangen door 'gewone woorden' verderop in de frequentielijst.

Om een indruk te geven van de woorden waarom het gaat bij een maat als Freq1000, geven we in Bijlage M de eerste 1000 Subtlex-woorden. Als we functiewoorden definiëren als voornaamwoorden, lidwoorden, voorzetsels, voegwoorden, telwoorden, hulpwerkwoorden, koppelwerkwoorden en tussenwerpsels, dan zijn er bij eerste duizend woorden 215 functiewoorden. Het merendeel van de functiewoorden is dus bijzonder frequent. Naarmate een tekst meer functiewoorden bevat, zal hij ook hoger scoren op de Freq-maten. Dat is een van de redenen om de exacte woordfrequentiematen alleen te geven voor inhoudswoorden.

We hebben ook gekeken naar het aantal verschillende lemma's in de woordfrequentielijsten uit SoNaR. De eerste 1000 woorden bevatten 791 lemma's, een percentage van 79% dus. Dat lemmapercentage daalt langzaam naar ongeveer 70%, zodat we bij de eerste 20000 woorden ongeveer 14000 lemma's aantreffen. Het aantal lemma's is informatief omdat de omvang van iemands vocabulaire vaak wordt uitgedrukt in lemma's en niet in woordvormen. Zo circuleert er op internet een woordenlijst van 2000 woordlemma's die zouden corresponderen met het A2-niveau Nederlands in het Europees Referentiekader (zie https://nl.wiktionary.org/wiki/WikiWoordenboek:Woordenschat_ERK-A2).

Ten slotte hebben we de eigenaardigheden van beide corpora verder verkend door de eerste 1000 woorden te vergelijken. Een korte samenvatting van deze vergelijking is als volgt.

- 626 woorden komen in beide lijsten voor bij de top-1000.
- 185 van de 374 woorden die alleen in de SoNaR top-1000 voorkomen gaan over hoeveelheden (*4, procent*) tijden (*april, 2003*), politiek (*premier, burgemeester*), economie (*directeur, euro, winst*), sport (*titel, finale*), of plaatsen (*nationale, buitenlandse*). Daarnaast treffen we informele spellingvarianten aan als *ni, nie* en *'k*, digitaal jargon als *spam* en de Zuid-Nederlandse pronomina *gij* en *ge*.
- 139 van de 374 woorden die alleen in de Subtlex top-1000 voorkomen betreffen personen (aanspreekvormen als *meneer,* nomina als *kerel,* pronomina als *je* en *mezelf*), misdaad (*vermoord, agent, drugs*), intieme of familierelaties (*liefje, broer, seks, trouwen*), alledaagse interjecties (*alsjeblieft, welterusten, ja*) of evaluaties en emoties (*spijt, geweldig, kwaad, klootzak*).

We concluderen dat SoNaR de sporen draagt van het grote aandeel van nieuwsteksten, waarin hoeveelheden, plaatsen, tijden van belang zijn voor precieze informatie, en waarin thema's als politiek, economie en sport veel besproken worden. Subtlex daarentegen is duidelijk conversationeel van aard: in gewone gesprekken gaat het vaak over personen en evaluaties en komen nogal wat interjecties voor. Daarnaast is er veel belangstelling voor misdaad, gegeven de thematiek van veel films en series.

Voorlopig blijkt uit corpusonderzoek dat maten gebaseerd op het Subtlex-corpus meer verschil maken tussen tekstgenres dan SoNaR-maten, waarschijnlijk omdat Subtlex vanwege zijn spreektaligheid gevoeliger is voor informele en 'gewone' taal, terwijl SoNaR meer schrijftaalwoorden bevat. Een illustratie van enkele verschillen tussen frequentieprofielen op basis van SoNaR en Subtlex geven Pander Maat et al. (2014).

Omdat de lijsten met woordfrequenties erg lang zijn vanwege het grote aantal zeldzame woorden, valt er bij het inlezen van die lijsten tijd te besparen door de minst frequente woorden eraf te knippen. Het percentage woorden dat mee wilt nemen, valt in te stellen bij 'Frequency Clipping'. Standaard staat dit percentage op 99, zodat de minst frequente 1% van de woorden buiten beschouwing blijft.

Samenstellingen

Soms kunnen woordlengte en woordfrequentie een misleidend beeld geven van de complexiteit van een woord. We denken dat sommige lange en infrequente woorden eenvoudiger zijn dan ze lijken, omdat het gaat om transparante samenstellingen. Daarom maakt T-Scan maakt gebruik van een lijst waarin voor ongeveer 46.000 nomina is aangegeven of het samenstellingen zijn, en zo ja wat het hoofd ervan is. We hebben ons in de annotatie van de woordenlijst beperkt tot samenstellingen waarvan de onderdelen steun bieden bij de interpretatie van het woord. Het gaat dus om samenstellingen die ook wel 'transparant' genoemd worden, of ook wel 'compositioneel interpreteerbaar'. Die term hebben we gedefinieerd aan de hand van drie eisen die we kort uiteenzetten. Meer informatie is te vinden in Bijlage L.

1 De samenstelling bevat meerdere vrije morfemen

Woorden als *bij-vakker* en *dorst-lesser* zijn geen samenstellingen in onze definitie, omdat het tweede deel van deze woorden niet los kan voorkomen.

2 Het hoofdwoord kan in deze betekenis op zichzelf staan

Voorbeelden van woorden die niet aan deze hoofdwoord-eis voldoen:

- *Bakboord* en *stuurboord* zijn geen boorden.
- Evenmin zijn *voorspoed* en *tegenspoed* vormen van *spoed.*
- Evenzo is een *asbak* een bak, maar een *bullebak* niet.
- Een *aandeelhouder* is een *houder* (vgl. *kaarthouder*), maar een *aanhouder* niet.

3 Het satellietwoord vertoont een consistente betekenis

De strenge eis van los kunnen voorkomen geldt alleen voor het hoofdwoord. Voor het satellietwoord zijn we rekkelijker. Neem het startmorfeem *mis*. In *mis-daad* en *mis-handeling* kan het eerste deel niet los voorkomen in dezelfde betekenis die het heeft in deze combinaties, maar kent het wel semantische consistentie: telkens is de betekenis 'afkeurenswaardig'.

We hebben in onze lijst ook het aantal onderdelen van de samenstelling opgenomen. Bij het tellen van de onderdelen is weer gekeken naar de compositionaliteit. Zo heeft *hypotheekrenteaftrek* drie onderdelen, want er is sprake van *aftrek,* meer in het bijzonder van *renteaftrek*, en nog meer in het bijzonder van *hypotheekrenteaftrek*. Iets dergelijks geldt voor *vrijhandelsakkoord* (*akkoord > handelsakkoord > vrijhandelsakkoord*, waarin verbijzondering als '>' is aangegeven). Maar allerlei woorden zijn niet op deze manier driedelig compositioneel. Zo heeft *mensenrechtenactivist* heeft slechts twee onderdelen, want er is niet zoiets als een *rechtenactivist*. Met andere woorden, in het kader van *activist* functioneert *mensenrechten* als een eenheid.

De kenmerken rond samenstellingen geven allereerst een indruk van het aantal samenstellingen en het aantal extra lange samenstellingen. Vervolgens worden woordlengtes gegeven voor niet-samenstellingen, samenstellingen, en hoofdwoord en satellietwoord van de samenstellingen. Om te zien in hoeverre samenstellingen invloed hebben op ons totaalbeeld van de woordlengte, worden vervolgens 'gecorrigeerde' woordlengtes gegeven voor naamwoorden en voor woorden algemeen. Als correctie wordt in die maten niet de lengte van de samenstellingen gebruikt, maar de lengte van de hoofdwoorden van de samenstellingen. De gecorrigeerde maat voor woorden in het algemeen is zowel met als zonder namen beschikbaar. Dezelfde correctiebenadering is gekozen voor woordfrequenties en voor Freq1000, Freq5000 en Freq20000. Daarmee kunnen we onder andere zien nagaan of de hoofdwoorden van transparante samenstellingen frequenter zijn dan de samenstelling als geheel. Is dat zo, dan is het denkbaar dat de 'voor samenstellingen gecorrigeerde' frequenties een betere voorspeller van moeilijkheid zijn dan de ongecorrigeerde frequenties.

3.4 Zinscomplexiteit

94.	Wrd_per_zin	Woorden per zin
95.	Wrd_per_dz	Woorden per deelzin
96.	Zin_per_wrd	Zinnen per woord
97.	Dzin_per_wrd	Deelzinnen per woord
98.	Wrd_per_nwg	Woorden per naamwoordgroep
99.	Betr_bijzin_per_zin	Het aantal betrekkelijke bijzinnen per zin
100.	Bijw_bijzin_per_zin	Het aantal bijwoordelijke bijzinnen per zin
101.	Compl_bijzin_per_zin	Het aantal complementszinnen per zin
102.	Fin_bijzin_per_zin	Het totaal aantal finiete bijzinnen per zin
103.	Mv_fin_inbed_per_zin	Meervoudige finiete inbeddingen
104.	Infin_compl_per_zin	Het aantal infinitiefcomplementen in de zin
105.	Bijzin_per_zin	Het totaal aantal bijzinnen in de zin (finiete plus infiniete bijzinnen)
106.	Mv_inbed_per_zin	Meervoudige inbeddingen: bijzinnen die zelf vallen onder een bijzin
107.	Betr_bijzin_los	Aantal losgekoppelde betrekkelijke bijzinnen
108.	Bijw_compl_bijzin_los	Aantal losgekoppelde bijwoordelijke en complementsbijzinnen
109.	Pv_hzin_per_zin	Het aantal persoonsvormen in declaratieve hoofdzinnen
110.	Pv_bijzin_per_zin	Het aantal persoonsvormen in bijzinnen
111.	Pv_ww1_per_zin	Het aantal persoonsvormen aan het begin van de zin
112.	Hzin_conj	Het aantal nevengeschikte declaratieve hoofdzinnen
113.	Bijzin_conj	Het aantal nevengeschikte bijzinnen
114.	Ww1_conj	Het aantal nevengeschikte zinnen dat begint met de persoonsvorm
115.	Pv_Alpino_per_zin	Het totaal aantal persoonsvormen volgens Alpino
116.	Pv_Frog_d	Dichtheid van persoonsvormen volgens Frog
117.	Pv_Frog_per_zin	Persoonsvormen per zin volgens Frog
118.	D_level	D-level
119.	D_level_gd4_p	Proportie zinnen met een D-level hoger dan 4
120.	Nom_d	Nominalisatiedichtheid
121.	Lijdv_d	Dichtheid van lijdende vormen
122.	Lijdv_dz	Aantal lijdende vormen per deelzin
123.	Ontk_zin_d	Dichtheid van zinsontkenningen
124.	Ontk_zin_dz	Zinsontkenningen per deelzin
125.	Ontk_morf_d	Dichtheid van morfologische ontkenningen
126.	Ontk_morf_dz	Morfologische ontkenningen per deelzin
127.	Ontk_tot_d	Dichtheid van ontkenningen totaal
128.	Ontk_tot_dz	Ontkenningen totaal per deelzin
129.	Meerv_ontk_d	Dichtheid van meervoudige ontkenningen
130.	Meerv_ontk_dz	Meervoudige ontkenningen per deelzin
131.	AL_sub_ww	Afhankelijkheidslengte (AL) tussen werkwoord en bijbehorend subject, in woorden
132.	AL_ob_ww	AL werkwoord – bijbehorend direct object
133.	AL_indirob_ww	AL werkwoord – bijbehorend indirect object
134.	AL_ww_vzg	AL werkwoord – bijbehorende bijwoordelijke voorzetselgroep
135.	AL_lidw_znw	AL zelfstandig naamwoord – bijbehorend lidwoord
136.	AL_vz_znw	AL voorzetsel – bijbehorend naamwoord
137.	AL_ww_wwvc	AL werkwoord – werkwoorden uit verbaal complement
138.	AL_vg_wwbijzin	AL voegwoord – persoonsvorm van de bijbehorende bijzin
139.	AL_vg_conj	AL voegwoord – hoofd van de bijbehorende conjuncten
140.	AL_vg_wwhoofdzin	AL voegwoord – persoonsvorm van bijbehorende hoofdzin
141.	AL_znw_bijzin	AL naamwoord – hoofd van de bijbehorende betrekkelijke bijzin
142.	AL_ww_schdw	AL werkwoord – scheidbaar deel van dit werkwoord
143.	AL_ww_znwpred	AL koppelwerkwoord – zelfstandig-naamwoordpredicaat
144.	AL_ww_bnwpred	AL koppelwerkwoord – bijvoeglijk-naamwoordpredicaat
145.	AL_ww_bnwbwp	AL werkwoord – bijwoordelijke bepaling met een bijvoeglijk naamwoord
146.	AL_ww_bwbwp	AL werkwoord – bijwoordelijke bepaling met een bijwoord
147.	AL_ww_znwbwp	AL werkwoord – bijwoordelijke bepaling met een zelfstandig naamwoord
148.	AL_gem	Het gemiddelde van alle afhankelijkheidslengtes per zin
149.	AL_max	Maximale AL per zin
150.	Bijw_bep_d	Dichtheid van bijwoordelijke bepalingen

151.	Bijw_bep_dz	Bijwoordelijke bepalingen per deelzin
152.	Bijw_bep_dz_zbijzin	Bijwoordelijke bepalingen per deelzin zonder de bijwoordelijke bijzinnen
153.	Bijw_bep_alg_d	Dichtheid bijw. bepalingen bestaand uit een algemeen bijwoord
154.	Bijw_bep_alg_dz	Bijw. bepalingen bestaand uit een algemeen bijwoord per deelzin
155.	Bijv_bep_d	Dichtheid van bijvoeglijke bepalingen
156.	Bijv_bep_dz	Totaal aantal bijvoeglijke bepalingen per deelzin
157.	Bijv_bep_dz_zbijzin	Bijvoeglijke bepalingen per deelzin zonder de betrekkelijke bijzinnen
158.	Attr_bijv_nw_d	Dichtheid van attributieve bijvoeglijke naamwoorden
159.	Attr_bijv_nw_dz	Aantal attributieve bijvoeglijke naamwoorden per deelzin
160.	Ov_bijv_bep_d	Dichtheid van bijvoeglijke bepalingen zonder adjectieven
161.	Ov_bijv_bep_dz	Aantal van bijvoeglijke bepalingen zonder adjectieven per deelzin
162.	KConj_per_zin	Aantal 'kleine' conjuncten per zin
163.	Extra_KConj_per_zin	Aantal extra elementen in vergelijking met de situatie zonder conjunctie
164.	KConj_dz	Aantal 'kleine' conjuncten per deelzin
165.	Extra_kconj_dz	Aantal extra elementen in vergelijking met de situatie zonder conjunctie
166.	Props_dz_tot	Het totaal aantal proposities per deelzin

3.4.1 De lengte van zinnen en deelzinnen

De syntactische kenmerken beginnen met de zinslengte in woorden. Nu is een zin vaak lang doordat zij bestaat uit meerdere deelzinnen met elk een vervoegd werkwoord. Daarom is het ook nuttig om te weten hoe lang deze deelzinnen zijn. T-Scan onderscheidt deelzinnen op basis van persoonsvormen (vervoegde werkwoorden). Die persoonsvormen kunnen natuurlijk hulpwerkwoorden zijn. Zogenaamde beknopte bijzinnen met infinitieven worden niet als deelzin worden gezien. De eerstvolgende zin bevat dus volgens T-Scan slechts twee deelzinnen, de zin erna slechts één (persoonsvormen zijn cursief):

a. Nadat we afscheid genomen *hadden*, *vertrokken* we.
b. Na afscheid genomen te hebben *vertrokken* we.

Als je niet geïnteresseerd bent in de lengte van deelzinnen maar in hun aantal, kun je kijken bij de kenmerken die direct op het aantal persoonsvormen ingaan.

In elliptische zinnen staat geen persoonsvorm. Zulke zinnen leveren problemen op bij een aantal maten die de frequentie van een verschijnsel per deelzin aangeven: in die maten wordt namelijk gedeeld door het aantal persoonsvormen. Daarom is voor die maten een 'work-around' gemaakt. Zo geldt voor elliptische zinnen het volgende:

- de lengte van de deelzin gelijk gesteld aan die van de zin;
- het aantal betrekkelijke bijzinnen en lijdende vormen per deelzin is op nul gesteld;
- een aantal andere deelzinsmaten zijn gereconstrueerd op basis van dichtheid en zinslengte; als de dichtheid van het aantal ontkenningen bijvoorbeeld 200 is, en de zinslengte is 5, dan is het aantal ontkenningen per deelzin 1.

3.4.2 Bijzinnen en persoonsvormen

Een aantal andere kenmerken gaat over bijzinnen, een categorie die specifieker is dan deelzinnen. Daarbij wordt onderscheid gemaakt tussen vier soorten bijzinnen.

1. Betrekkelijke bijzinnen; grofweg worden deze in Alpinotermen gedefinieerd als knopen van het type *mod-rel* en *mod-whrel* (telkens wordt eerst het dependentielabel en dan het categorielabel weergegeven; zie verder Bouma et al. 2001).
2. Bijwoordelijke bijzinnen, grofweg gedefinieerd als knopen het type *mod-cp*.
3. Complementszinnen, dat wil zeggen bijzinnen met zinsdeelfunctie. Deze worden gedefinieerd als knopen de categorieën *whsub, whrel* en *cp* die géen dependentielabel hebben van het type *mod*. Een deel van deze bijzinnen is overigens in strikte zin bijvoeglijk, want hangt aan een nomen (bv. *het idee dat ik weg moet*)

4. Infinitiefcomplementen, dat wil zeggen bijzinnen rond infinitieven (Alpinocategorie *ti*). Een deel van deze complementen is wederom bijvoeglijk (*het idee weg te moeten*).

Verder wordt de som berekend van de finiete bijzinnen (1-3) en alle bijzinnen (1-4) per zin. Ten slotte wordt gekeken naar meervoudige inbeddingen. Dat zijn bijzinnen die zelf weer deel uitmaken van een bijzin. Daarbij wordt weer onderscheid gemaakt tussen alleen de finiete bijzinnen en het totaal inclusief de infinitiefcomplementen. Bijlage N bevat voorbeelden van de verschillende soorten bijzinnen en de manier waarop T-Scan ze (via Alpino) categoriseert, evenals de precieze definities waarmee wordt gewerkt.

Bijzinnen kunnen ook 'los' voorkomen, met andere woorden beginnen na een punt. Ook deze gevallen worden geteld (vergelijk de betrekkelijke bijzin *Waardoor alles misliep*). We maken onderscheid tussen losse betrekkelijke bijzinnen enerzijds en losse bijwoordelijke ofwel complementsbijzinnen anderzijds.

We hebben op verschillende manieren gekeken naar onnauwkeurigheden in de automatische analyse van bijzinnen. Om te beginnen zijn september 2016 1003 zinnen geanalyseerd waarin T-Scan een nevenschikking van twee deelzinnen ziet. In negen daarvan bleek slechts één persoonsvorm te staan, zodat er 994 samengestelde zinnen overblijven. In 98 daarvan bleken bijzinnen over het hoofd te zijn gezien. Na analyse van de fouten en aanpassing van de software hebben we naderhand 49 van deze bijzinnen alsnog zichtbaar kunnen maken, vooral door T-Scan op een slimmere manier gebruik te laten maken van de output van zinsontleder Alpino (Bouma et al. 2001). De andere helft van de fouten is vooral te wijten aan incorrecte Alpino-analyses van lastige zinnen; vaak gaat het om lange opsommingen en onderbrekingen van constructies door delen tussen komma's, haakjes of gedachtestreepjes. De conclusie is dat we in analyses van professioneel geredigeerde taal rekening moeten houden met een foutmarge van rond de 5% wat betreft de herkenning van bijzinnen.

De volgende vraag is of het onderscheid tussen betrekkelijke bijzinnen, bijwoordelijke bijzinnen en complementszinnen correct wordt gemaakt. Daarom zijn voor elk type 100 gevallen handmatig geanalyseerd. De resultaten van die controle staan hieronder.

Correct label	**Betrekkelijk**	**Bijwoordelijk**	**Complement**	**Nevenschikking**
T-Scanlabel				
Betrekkelijke bijzin	**94**		5	1
Bijwoordelijke bijzin		**97**	3	
Complementszin	14	4	**82**	

De correctheid van de T-Scananalyse van typen bijzinnen (correcte analyses: vet)

De validiteit van de bijzinsanalyse lijkt behoorlijk. De resultaten voor betrekkelijke en bijwoordelijke bijzinnen zijn geruststellend; de analyse van complementszinnen laat enigszins te wensen over. Het komt regelmatig voor dat T-Scan een betrekkelijke bijzin aanziet voor een complementszin, zoals in de volgende twee voorbeelden.

- De wegen lopen dwars door de gebieden *waar alle wilde dieren leven*.
- Zelfs de manager en de baas van het restaurant zijn echt supergrappig, *wat ik totaal niet had verwacht*.

Naast bijzinnen tellen we het aantal persoonsvormen in de zin, op verschillende manieren. Op basis van Alpino-analyses onderscheiden we drie soorten persoonsvormen:

1. Pv_hzin_per_zin: het aantal smain-knopen in de zin, met andere woorden het aantal declaratieve onafhankelijke deelzinnen;
2. Pv_bijzin_per_zin: het aantal ssub-knopen in de zin, met andere woorden het aantal bijzinnen maar dan grover gemeten dan in termen van de drie bijzinstypen hierboven (betrekkelijk,

bijwoordelijk, complement); de correlatie tussen pv_bijzin_per_zin en fin_bijzin_per_zin is hoog, maar niet perfect. Alpino ziet soms ssubknopen die niet voldoen aan de beschrijvingen van een van de drie bijzinstypen. We gaan daar nog nader naar kijken.

3. Pv_ww1_per_zin: het aantal sv1-knopen, met andere woorden het aantal deelzinnen dat begint met de persoonsvorm.

Het totaal aantal persoonsvormen geven we als Pv_Alpino_per_zin.

Zowel voor hoofd- als bijzinnen is het mogelijk dat zij worden nevengeschikt, in die zin dat zij het Alpino-dependentielabel 'CNJ' dragen: het gaat hier om opsommingen en coördinaties met behulp van nevenschikkende voegwoorden. We tellen dus het aantal CNJ-labels gekoppeld aan declaratieve hoofdzinnen, bijzinnen en werkwoords-initiële zinnen (resp. hzin_conj, bijzin_conj en ww1_conj).

Naast Alpino telt ook Frog het aantal persoonsvormen. We nemen die maat voorlopig op om de verschillen met de Alpinomaat nader te kunnen bestuderen. Handmatige controles leren dat Frog meer persoonsvormen mist dan Alpino. Daarom is in de berekening van deelszinslengte gewerkt met de Alpinomaat. Een indruk van de validiteit van de maat is verkregen door in 100 zinnen (50 uit reisblogs en 50 uit romans) handmatig het aantal persoonsvormen te tellen. De zinnen hadden gemiddeld 2.54 persoonsvormen (SD 2.47). De correlatie tussen de Alpinomaat en het handmatig getelde aantal persoonsvormen was .996.

3.4.3 D-level

Twee kenmerken behandelen het D-level van de zin. D-level is een afkorting van 'Development Level'; het gaat om een classificatie en rangordening van zinstypen naar moeilijkheid, oorspronkelijk afkomstig uit Rosenberg & Abbeduto (1987). Zinsconstructies worden geordend op de D-levelschaal op basis van de volgorde waarin kinderen het gebruik van deze constructies aanleren. Het lijkt een redelijke aanname dat de structuren die als eerste worden beheerst 'gemakkelijk' mogen worden genoemd, en de latere structuren 'moeilijker' van aard zijn.

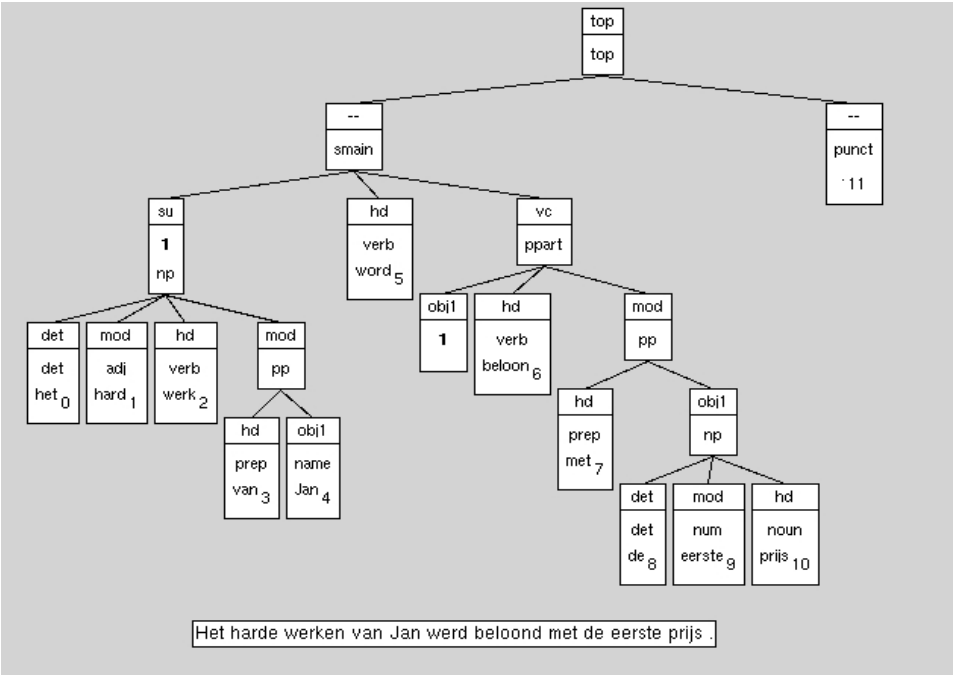
Onze implementatie, die in grote lijnen Covington et al. volgt, kent 8 niveaus; voor details, zie Bijlage A. Iedere zin wordt op D-level geanalyseerd door te kijken of de zin tot het hoogste complexiteitsniveau behoort. Indien dit niet zo is wordt een niveau lager gekeken. Als niveau 1 niet wordt toegewezen, wordt automatisch niveau 0 aan de zin toegekend.

T-Scan geeft per zin het D-level en het gemiddelde daarvan voor hogere tekstniveaus. Daarnaast wordt de proportie zinnen met een D-level hoger dan 4 gegeven.

3.4.4 Nominalisaties

T-Scan besteedt ook aandacht aan nominalisaties; daarvan wordt de dichtheid gegeven. Nominalisaties drukken op een compacte manier situaties uit waaraan de auteur ook een bijzin met werkwoord had kunnen besteden. Nominalisaties worden herkend op basis van een lijst suffixen. Daarbij is wel een selectie gedaan: we hebben alleen de nominalisatie-suffixen gekozen die naar onze indruk tekst abstracter maken, en welke niet. Het suffix *–er* bijvoorbeeld dient om werkwoorden om te zetten in zelfstandige naamwoorden die naar personen verwijzen (*bakker, denker, doorzetter*). Maar omdat daarmee het werkwoord niet veel abstracter wordt van betekenis, is *–er* niet in onze lijst opgenomen. Wel bijvoorbeeld *–atie*,*– ing* en *–ie.* Zie voor onze lijst Bijlage B.

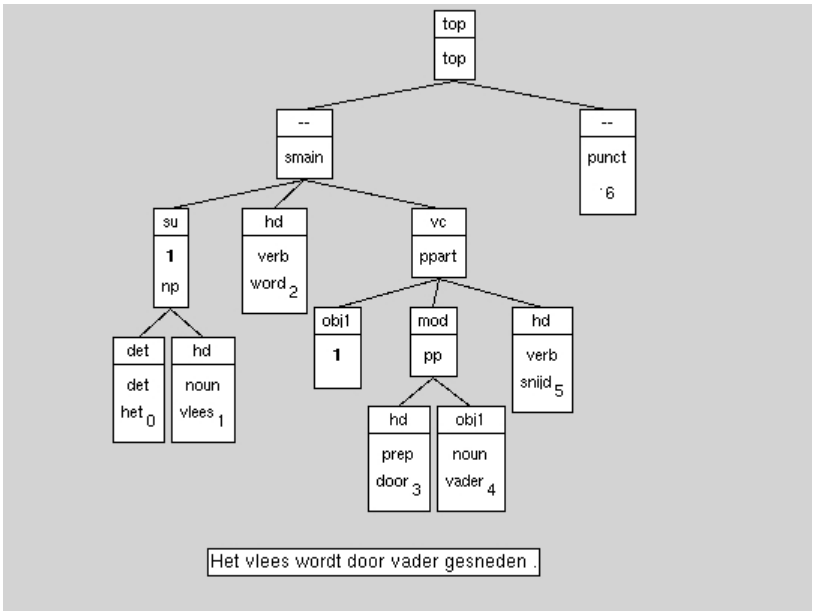
Daarnaast kennen we in het Nederlands nog genominaliseerde infinitieven ('Het harde werken van Jan werd beloond met de eerste prijs.'). Deze zijn redelijk makkelijk te herkennen aan de hand van de Alpino ontleding: het gaat in deze gevallen altijd om een werkwoordsknoop die onder een NP valt. Zie figuur 1.1. Een overzicht van de Alpino-afkortingen is te vinden op https://www.let.rug.nl/vannoord/alp/Alpino/adt.html.



Figuur 1. Alpino-analyse van een zin met een genominaliseerde infinitief en een lijdende vorm.

3.4.5 Lijdende vormen

Ook lijdende vormen worden geïdentificeerd op basis van Alpino. In de Alpino-output wordt daartoe gekeken of er een hulpwerkwoord van passieve vorm te vinden is: een vorm van 'worden' of 'zijn' gekoppeld aan een werkwoordelijk complement (verbaal complement, VC) met daarin een ‘leeg’ direct object (OBJ1) dat via index verwijst naar het subject van de zin. Zo’n configuratie is zowel in Figuur 1 als in Figuur 2 hieronder te zien. Alpino toont ons dat het subject in de passieve zin feite het object is van het hoofdwerkwoord.



Figuur 2. De Alpino-analyse van een zin met een lijdende vorm

3.4.6 Ontkenningen

Er wordt wel gezegd dat ontkenningen een tekst moeilijker kunnen maken. T-Scan onderscheidt ontkenningen op zinsniveau en op woordniveau.

Zinsontkenningen worden gevonden met behulp van een lijst ontkennende of 'negatieve' woorden: *allerminst, allesbehalve, amper, behalve, contra, evenmin, geen, geeneens, geenszins, generlei, kwijt, nauwelijks, nergens, niemand, niemendal, niet, niets, nihil, niks, nimmer, nimmermeer, noch, nooit, ongeacht, slechts, tenzij, ternauwernood, uitgezonderd, weinig, zelden, zeldzaam, zonder*.

Er wordt een woordbuffer bijgehouden om te kunnen controleren op de volgende woordcombinaties die samen ook een ontkenning kunnen markeren: *afgezien van, zomin als, met uitzondering van*. De woorden *moeilijk* en *weg* worden alleen geteld als ze als bijwoord in de zin voorkomen; hiervoor benutten we de part-of-speech-labels van Frog.

Ontkenningen op woordniveau worden gevonden op twee manieren. Allereerst wordt via patroonherkenning op stringniveau gekeken of de volgende morfemen plus verbindingsstreepje in het woord voorkomen: *mis-, non-, niet-, anti-, ex-, on-, oud-.* Vervolgens wordt aan de hand van de morfologische ontleding door Frog gekeken naar woorden met als eerste morfeem mis, de, non, on. De totale morfeemlengte moet wel groter zijn dan één (om bijvoorbeeld het woord 'de' niet mee te tellen). Ook mag voor woorden met twee morfemen het tweede morfeem niet en zijn (om 'nonnen' niet mee te laten tellen).

De aantallen en dichtheden voor zins- en woordontkenningen worden opgeteld tot een totaal. Daarnaast wordt gezocht naar meervoudige ontkenningen; die worden gedefinieerd als het voorkomen van meerdere negaties in dezelfde zin. Een zin met dubbele negatie kan overigens de verschillende negaties in verschillende deelzinnen hebben.

3.4.7 Afhankelijkheidslengtes

Een afhankelijkheidslengte is de afstand tussen twee zinsdelen die bij elkaar horen, zoals bijvoorbeeld het werkwoord en het subject van de zin, of tussen een lidwoord en het bijbehorende naamwoord. Het gaat telkens om afstanden tussen het 'hoofd' van een constructie en de 'dependent', dat wil zeggen het afhankelijke element. Op zinsniveau wordt het hoofd gevormd door de werkwoordsgroep (die soms uit meerdere elementen bestaat, namelijk hulpwerkwoord en hoofdwerkwoord), met als dependenten het subject, direct en indirect object, en de bijwoordelijke bepalingen die direct aan het werkwoord hangen. Op lager niveau zijn er dependentierelaties tussen bijvoorbeeld bijvoeglijke bepalingen en het hoofd van de naamwoordgroep waar deze bij horen.

Hoe groter de afstanden tussen hoofden en hun dependenten, hoe lastiger de zin te verwerken is voor lezers (Gibson 2000); in de Nederlandse taaladviesliteratuur wordt dan gesproken van een tangconstructie. T-Scan drukt de afstanden uit in het aantal woorden dat moet worden overbrugd van het ene naar het andere zinsdeel. Als twee zinsdelen direct naast elkaar staan, is de afstand 0.

Afhankelijkheidslengtes worden bepaald met behulp van Alpinobomen, waarin de verschillende dependentierelaties relatief expliciet zijn weergegeven. De lengte van een dependentierelatie wordt bepaald door de positie van zowel het hoofd als de dependent op te vragen en die van elkaar af te trekken. Er zijn een aantal speciale gevallen waar wel rekening mee gehouden dient te worden. Zo kan een afstand tussen een hoofd en dependent alleen worden berekend indien het om twee woorden gaat, en komt het regelmatig voor dat de dependent van een bepaald woord een hele woordgroep is. In dat geval wordt vanaf deze knoop (bijv. een NP-constituent) recursief de boom naar beneden afgezocht tot een hoofd is gevonden van deze groep dat uit slechts één woord bestaat. De positie van dit ene woord wordt vervolgens gebruikt in de berekening van de afhankelijkheidslengte. Ook kan het zo zijn dat een van de twee knopen 'leeg' is, met een indexering die gedeeld wordt met een woord elders in de zin (zie bijv. afbeelding 2 hierboven). In dat geval nemen we de positie van het geïndexeerde woord mee in de berekening van de afhankelijkheidslengte.

De namen van de meeste afhankelijkheidslengtes worden toegelicht in de overzichtstabel aan het begin van deze paragraaf. In Tabel 2 geven we per type een voorbeeld. Daarbij wordt ook duidelijk dat

nogal wat soorten afhankelijkheidslengtes vaker dan eens voorkomen in dezelfde zin: zo bestaat de werkwoordelijke groep vaak uit een hulpwerkwoord en een hoofdwerkwoord, die allebei een afstand tot bijvoorbeeld het subject hebben. De verschillende lengtes van zo'n vaker voorkomend type worden door T-Scan op zinsniveau gemiddeld. Als de lengte dan bijvoorbeeld 2 is, kan het zijn dat een lengte van 0 en een lengte van 4 zijn gemiddeld. Dat betekent dat de lengtes op zinsniveau niet altijd rechtstreeks te herleiden zijn tot de constructie waarom het gaat. Voor de maximale afhankelijkheidslengte (AL_max) wordt dan weer wel de langste afzonderlijke afhankelijkheid genomen.

De laatste twee kenmerken rondom afhankelijkheidslengtes bieden een samenvatting. Allereerst wordt het gemiddelde gegeven van alle soorten lengtes voor een bepaalde zin. Die soorten lengtes kunnen zelf ook weer een gemiddelde kunnen zijn, zoals we gezien hebben. Op hogere tekstniveaus worden voor AL_gem de gemiddelden per zin op hun beurt weer gemiddeld.

Daarnaast geeft AL_max de hoogste afzonderlijke AL-score die is aangetroffen. Een hoge score bij AL_max, zeker in verhouding tot de zinslengte, is een handige indicatie dat er iets aan de hand is met een zin.

Soort lengte	**Voorbeelden**
AL werkwoord- subject	**- Peter**, mijn neef uit Canada, **schrijft** regelmatig. Het gaat bij een complex subject zoals hierboven altijd om de lengte tussen het hoofd van het subject en het werkwoord. **- Ik ben** gisteren naar de bioscoop **gegaan**. N.B. Bij een complexere werkwoordgroep, zoals in de tweede zin hierboven, wordt het gemiddelde genomen van de afstanden tussen *ik* en *ben* (0) en die tussen *ik* en *gegaan* (4). Dit 'middelen' bij meerdelige werkwoordsgroepen gebeurt alleen bij de AL werkwoord-subject, niet bij de lengtes tussen AL en (indirect) object. **- Ik zag** hem op straat lopen. N.B. In samengestelde zinnen worden meerdere subject-ww-combinaties bekeken, zowel die in hoofdzinnen als bijzinnen, zelfs in non-finiete bijzinnen. In de derde zin gaat het bijvoorbeeld om de afstanden tussen *ik* en *zag* en tussen *hem* en *lopen*. **- Het is** niet vreemd *dat **ik** hem **zag***. N.B. Omdat er hier sprake is van een onderwerpszin, worden in deze zin drie lengtes gemiddeld: die tussen *het* en *is* (0), die tussen *ik* en *zag* (1), en die tussen het hoofd van de onderwerpszin *dat ik hem zag* en het werkwoord van de hoofdzin *is* (5). Het gemiddelde is dus 2. N.B. Bij deze afstand wordt het predicaatsnomen buiten beschouwing gelaten (ook al kan het deel uitmaken van het verbale complement).
AL werkwoord- direct object	Karel **gaf** mij een geweldige **roman**.
AL werkwoord- indirect object	Karel **gaf mij** een geweldige roman.
AL werkwoord- voorzetselgroep	Thea **woonde** al jaren **bij** haar moeder.
AL zelfstandig naamwoord- lidwoord	Ik heb **de** lange **man** niet gezien.
AL voorzetsel- naamwoord	**In** dat kleine **café** zijn er veel bieren op tap.
AL werkwoord- werkwoord uit verbaal complement	Ik **dacht** gisteren dat ik te laat **was**. Ik **nodig** hem uit om te **komen eten**. Ik **beloofde** gisteren te **komen**. Ik **heb** hem al jaren niet meer **gezien**. Het gaat hier om afstanden tussen het hoofdwerkwoord en het werkwoord uit het verbale complement daarvan. Als verbaal complement rekent Alpino niet alleen finiete en infiniete bijzinnen, maar ook voltooid-deelwoordgroepen en passief-deelwoordgroepen ('ppart'-knopen). N.B. Als er twee afstanden van dit type voorkomen, worden die gemiddeld. Dat leidt in de volgende zin tot een gemiddelde van 2.5, omdat de eerste afstand (wilde-prikken) 5 bedraagt en de tweede (zou-zijn) 0: Ze **wilde** zo snel mogelijk een datum **prikken** voordat haar buik te groot **zou zijn** . N.B. Bestaat de werkwoordelijke groep in de objectzin uit meerdere werkwoorden, dan wordt de afstand tussen het hoofdzins-werkwoord en het objectzin-werkwoord verdeeld over de verschillende werkwoorden; dat leidt tot lagere scores. en gemiddeld. Dat leidt soms tot lagere lengtes: zo wordt in 'Ik **beloofde** gisteren te **zullen komen'** de afstand van twee woorden verdeeld over twee werkwoorden, wat een score van 1 oplevert.
AL voegwoord - werkwoord bijzin	Ik ging naar huis **omdat** ik moe **geworden was**. Hij zei **dat** zijn moeder gek **was**. Ik ga naar huis **om** dat te **doen.** Het gaat om alle werkwoorden, dus zowel persoonsvormen als werkwoordelijke complementen.
AL voegwoord - hoofden van de bijbehorende conjuncten	Ik heb hem twee **boeken** gegeven **en** drie hele kleine **plantjes**. Ik **gaf** hem een boek **en** hij **was** niet eens blij. N.B. De aard van het 'hoofd' hangt af van de aard van de door het voegwoord verbonden conjuncten. Het gemiddelde van de conjuncten wordt genomen. Bovendien worden alle voegwoorden in de zin meegenomen. Dit is soms verwarrend, omdat een zin als 'Het **was** echt een super chill hostel **en** er **was** een super chille **woonkamer en dakterras**' twee voegwoorden (tweemaal 'en') met elk twee conjuncten bevat. Er worden dan vier afstanden gemiddeld (5-1-0-0), wat een waarde van 1.5 oplevert.
AL voegwoord- werkwoord hoofdzin	**Omdat** ik moe geworden was, **ben** ik naar huis **gegaan**. Het gaat om alle werkwoorden, dus zowel persoonsvormen als werkwoordelijke complementen. Een omissie op dit moment is dat deze afstand niet wordt berekend voor niet-

Soort lengte	Voorbeelden
	finiete bijzinnen (er is dus geen afstand voor ***Om*** *dat te doen,* ***ga*** *ik naar huis.*)
AL naamwoord- hoofd betrekkelijke bijzin	Hij heeft die **man** gezien die jij gisteren **sprak**. Hij heeft het **idee** dat zijn moeder gek **is**.
AL werkwoord- scheidbaar deel werkwoord	Ik **nodig** hem nooit meer voor zoiets **uit**.
AL koppelwerkwoord- zelfstandig-naamwoordpredicaat	Hij **is** al jaren de beste **skiër** van Nederland.
AL koppelwerkwoord- bijvoeglijk-naamwoordpredicaat	Hij **is** als onderzoeker erg **goed**.
AL werkwoord- bijw. bep. met een bijvoeglijk naamwoord	Hij **liep** de marathon erg **snel**.
AL werkwoord- bijw. bep. met een bijwoord	Hij **liep** de marathon in twee uur **gisteren**.
AL werkwoord- bijw. bep. met een zelfstandig naamwoord	Hij **tennist** al **jaren** niet meer.

Tabel 2. De verschillende soorten afhankelijkheidslengtes in T-Scan geïllustreerd

We hebben de afhankelijkheidslengtes tweemaal op validiteit getest in handmatige analyses. De eerste keer werden 100 zinnen gebruikt met een gemiddelde lengte van 16.3 woorden; de gemiddelde maximale afhankelijkheidslengte was volgens T-Scan 6.93, en handmatig bepaald 7.09. De correlatie tussen de automatische en de handmatig berekende maximale lengte was .84.

De tweede keer ging het om 360 zinnen met een gemiddelde lengte van 15.4 woorden; de gemiddelde maximale afhankelijkheidslengte was volgens T-Scan 7.26, en handmatig bepaald 7.62. De correlatie tussen de automatische en de handmatig berekende maximale lengte was .88.

Toch laten deze niet-perfecte correlaties ook zien dat er fouten gemaakt worden. Zo is er in de eerste zin hieronder een lange link tussen het hoofd van de hoofdzin (*zijn*) en het tweede *omdat*; die link wordt gemist door T-Scan, wellicht omdat er hier sprake is van nevengeschikte bijzinnen. In dit geval is de onderliggende Alpino-analyse correct, en kan de fout waarschijnlijk hersteld worden door T-Scan daarvan beter gebruik te laten maken. En in de tweede zin beschouwt Alpino de betrekkelijke bijzin met *waarvan* foutief als subjectzin. Daarom suggereert T-Scan als langste link die tussen de hoofden van de subjectzin en hoofdzin (resp. *volgde* en *was*). In feite is de langste link een kortere, namelijk die tussen *waarvan* en *volgde* in de betrekkelijke zin.

- Ze **zijn** bijvoorbeeld angstig om hun mening te verkondigen, omdat anderen deze mening kunnen tegenspreken of **omdat** ze bang zijn voor sociale uitsluiting.
- En nog steeds was er geen enkel teken van de groep id-wezens, **waarvan** hij nu al bijna honderd dagen het spoor **volgde**.

We *hebben* de afhankelijkheidslengtes van T-Scan ook op een andere manier op validiteit *ge*test*, namelijk* door te kijken naar de zinnen waarvoor T-Scan een maximale lengte geeft die problematisch lijkt voor de verwerker. Daartoe hebben we 50 zinnen bekeken waarvoor de tool een maximale lengte van 15 geeft: hoeveel van die zinnen hebben daadwerkelijk een tangconstructie van die omvang? De gemiddelde handmatig vastgestelde lengte voor de 50 zinnen bedroeg 14.5 woorden (sd 2.6). Voor 41 van de zinnen bleek de handmatig vastgestelde lengte inderdaad 15 woorden te bedragen. Voor drie zinnen bleek de lengte te zijn onderschat, voor zes zinnen overschat, waarbij vier zinnen feitelijk onder de 10 woorden scoorden. Valse alarmen *wat betreft afhankelijkheidslengtes* komen dus voor, maar de kans erop is niet veel groter dan 10%. Wanneer we naar alle lengtes kijken, lijkt T-Scan die eerder iets te onderschatten dan te overschatten, gezien de hierboven gerapporteerde gemiddelde lengtes.

3.4.8 Bijwoordelijke bepalingen

De gedachte achter het tellen van bepalingen is dat zinnen via bepalingen op efficiënte wijze kunnen worden gevuld met extra informatie, waarmee in principe de informatiedichtheid toeneemt.

Bijwoordelijke bepalingen worden door T-Scan geteld per deelzin en per 1000 woorden. Als bijwoordelijke bepaling tellen we Alpino-knopen met de volgende kenmerken:

- dependentielabel is *mod* (bepaling), of *predm* (bepaling van gesteldheid tijdens de handeling);
- de knoop hangt direct onder een categorie waarin een werkwoord centraal staat, dus onder de volgende categorieën: *smain, ssub, sv1, inf, ti, ppart* of *ppresent*.

Wanneer een bijwoordelijke bepaling wordt samengetrokken, wordt hij twee maal geteld. Dus in *hij denkt en handelt snel* telt *snel* voor twee bijwoordelijke bepalingen.

We moeten wel bedenken dat een bijwoordelijke bepaling zowel kan bestaan uit grote woordgroepen (… *op een feestje in Boston, waar ik toen woonde*) als uit een enkel bijwoord (*eens; gisteren*). Dat betekent dat teksten met veel bijwoordelijke bepalingen ook teksten kunnen zijn met veel bijwoorden. En veel bijwoorden hebben een algemene strekking, zoals *eens* of *daardoor*. Niet alle bijwoordelijke bepalingen voegen dus specifieke informatie toe aan de zin (dit in tegenstelling tot bijvoeglijke bepalingen). Daarom vermeldt T-Scan het aantal bijwoordelijke bepalingen op dat slechts bestaat uit een enkel algemeen bijwoord. Trekt men dat aantal af van het totaal aantal bijwoordelijke bepalingen, dan blijven de bepalingen met specifieke extra inhoud over; dat zal de gebruiker zelf moeten doen, want T-Scan geeft alleen het totaal en het aantal bepalingen gevormd door algemene bijwoorden. In 3.7.4 (Algemene en specifieke bijwoorden) wordt toegelicht wat T-Scan verstaat onder algemene bijwoorden.

Een andere kanttekening is dat ook bijwoordelijke bijzinnen ingeleid door voegwoorden vallen onder onze definitie van bijwoordelijke bepaling. Wil alleen kijken naar 'kleinere' bepalingen, dan moet je kijken naar het aantal bijwoordelijke bepalingen met buiten beschouwing laten van de bijwoordelijke bijzinnen (*bijw_bep_dz_zbijzin*).

Hoe valide worden de bepalingen geïdentificeerd? Er is een kleinschalige handmatige controle op validiteit gedaan met 50 zinnen, die samen 129 bijwoordelijke en 66 bijvoeglijke bepalingen tellen. Daarbij is gekeken naar de correlatie tussen het aantal bepalingen per deelzin op basis van handmatige analyse en dat op basis van T-Scan. Voor bijwoordelijke bepalingen bedroeg die .96, wat geruststelt.

3.4.9 Bijvoeglijke bepalingen

Ook voor bijvoeglijke bepalingen worden globale maten gegeven, en daarnaast aparte maten voor attributieve adjectieven en andersoortige ('overige') bijvoeglijke bepalingen. Onder die laatste groep vinden we zowel voor- als nabepalingen. Bij de voorbepalingen gaat het om:

- Telwoorden (*twee en een halve liter; een tweede huis*)
- Genitiefconstructies (*Peters honden*)
- Deelwoorden en infinitieven (*blaffende honden; verstoorde relaties; te nemen maatregelen*)
- Substantieven (*de stad Antwerpen; een glas rode wijn*)

Bij de nabepalingen spreken we over:

- Bijstellingen (*de schipper, een voorzichtig man, bleef thuis; de wedstrijd Nederland-België*)
- Naamwoorden of naamwoordgroepen (*de wedstrijd vorige week*)
- Bijwoorden (*de kamer boven*)
- Genitiefconstructies (*de plek des onheils*)
- Voorzetselgroepen (*de jeugd van tegenwoordig*)
- Groepen na een voegwoord (*alle kinderen behalve de oudste*)
- Beknopte bijzinnen (*een kind om te zoenen*)
- Bijvoeglijke bijzinnen (bijvoeglijke complementen zoals *de kans dat hij weer opknapt;* betrekkelijke bijzinnen zoals *de groep waartoe de herten behoren*).

Het laatste type nabepaling vergt extra aandacht: een bijzin kan meetellen als bepaling. Vergelijk bijvoorbeeld de volgende zinnen:

a. De man die daar loopt mijn vriend.
b. De lange man die daar loopt is mijn vriend.

c. De man met hoed die daar loopt is mijn vriend.

In alle zinnen vormt de betrekkelijke bijzin *die daar loopt* een 'overige bijvoeglijke bepaling'. In de tweede zin wordt daaraan een attributief adjectief toegevoegd (*lange*), in de derde een nabepaling in de vorm van een voorzetselgroep (*met hoed*). De tweede en derde zin tellen dus meerdere bijvoeglijke bepalingen, waarvan er een de vorm heeft van een bijzin.

Dus wil je alleen kijken naar minder omvangrijke bijvoeglijke bepalingen, dan moet je afgaan op het aantal bijvoeglijke bepalingen zonder de bijvoeglijke bijzinnen: *bijv_bep_dz_zbijzin*.

Er zijn ook bijvoeglijke complementszinnen, als het gaat om nomina zoals *verwachting* (dat een finiet complement kan hebben zoals *dat hij zou komen*) of *wens* (dat een infiniet complement kan hebben zoals *om te komen*). Voor het aantal bijvoeglijke complementszinnen is niet te corrigeren: deze finiete complementen zijn niet apart gehouden van de andere finiete complementen, en de infiniete evenmin van de andere infinitiefcomplementen (zie hierboven de paragraaf over bijzinnen). Hier vindt dus een lichte vorm van dubbeltelling plaats.

Ten slotte melden we dat net als bij bijwoordelijke bepalingen de samengetrokken bijvoeglijke bepalingen dubbel worden geteld.

Hoe valide worden de bijvoeglijke bepalingen geïdentificeerd? In 50 zinnen, die 66 bijvoeglijke bepalingen tellen is gekeken naar de correlatie tussen het aantal bepalingen per deelzin op basis van handmatige analyse en dat op basis van T-Scan. De was .93; de validiteit van de analyse lijkt dus goed.

3.4.10 Nevenschikkingen binnen deelzinnen

Een zin is 'zwaarder' wanneer hij nevenschikkingen en opsommingen bevat. Daarom telt T-Scan wat hieronder 'kleine conjuncten' genoemd wordt. Dat zijn leden van nevenschikkingen waarin geen werkwoorden voorkomen. Meer specifiek gaat het om de volgende kenmerken.

- *KConj_per_zin (aantal kleine conjuncten per zin)*

Het gaat om het aantal knopen in de zin met dependentielabel cnj, uitgezonderd de cnj-gevallen met categorielabel smain, sv1, ssub,rel, whrel, cp, oti, ti en whsub; want dat zijn 'grote' conjuncten.

- *Extra_KConj_per_zin (aantal extra elementen per zin in vergelijking met de situatie zonder conjuncties)*

Van KConj_per_zin wordt afgetrokken het aantal knopen in de zin met categorielabel Conj, wederom uitgezonderd de Conj-knopen waaronder knopen vallen met categorielabel smain, sv1, ssub,rel, whrel, cp, oti, ti en whsub.

- *KConj_dz (=aantal kleine conjuncten per deelzin)*

Deel KConj_per_zin door het aantal persoonsvormen volgens Alpino, waarbij eerst de deelzinnen zonder persoonsvorm op 1 gezet zijn.

- *Extra_kconj_dz (aantal extra elementen per deelzin in vergelijking met de situatie zonder conjuncties)*

Deel Extra_KConj_per_zin door het aantal persoonsvormen volgens Alpino, waarbij eerst de deelzinnen zonder persoonsvorm op 1 gezet zijn.

Een voorbeeld. De volgende zin (en deelzin) bevat eerst een nevenschikking met drie conjuncten (*informatietechnologie, integratie, internationalisering*) en daarna, binnen het derde conjunct, weer een nevenschikking met twee conjuncten (*economie, leefwereld*), wat per saldo vier extra conjuncten oplevert.

Informatietechnologie, de Europese integratie en internationalisering van de economie en sociale leefwereld stuwen de ontwikkelingen op.

Hoewel T-Scan soms fouten maakt bij het tellen van nevengeschikte bijzinnen en conjuncten in conjuncten, correleert de T-Scanscore voor extra conjuncten hoog (.96) met het handmatig vastgestelde

aantal (n=100); we hebben dus een behoorlijk valide benadering van het aantal nevenschikkingen binnen deelzinnen.

Om tot een schatting te komen van de informatie per deelzin, is het totaal aantal proposities per deelzin bepaald (*props_dz_tot*), en wel als volgt. De kernmededeling van de deelzin is de eerste propositie, en daarbij zijn per deelzin opgeteld de bijwoordelijke bepalingen zonder bijzinnen, de bijvoeglijke bepalingen zonder bijzinnen en het aantal extra conjuncten per deelzin.

3.5 Referentiële coherentie en lexicale diversiteit

167.	TTR_wrd	Type-token-ratio voor woorden
168.	MTLD_wrd	Measure of textual lexical diversity voor woorden
169.	TTR_lem	Type-token-ratio voor lemma's
170.	MTLD_lem	Measure of textual lexical diversity voor lemma's
171.	TTR_namen	Type-token-ratio voor namen
172.	MTLD_namen	Measure of textual lexical diversity voor namen
173.	TTR_inhwrd	Type-token-ratio voor inhoudswoorden
174.	MTLD_inhwrd	Measure of textual lexical diversity voor inhoudswoorden
175.	TTR_inhwrd_zonder_abw	Type-token-ratio voor inhoudswoorden, zonder algemene bijwoorden
176.	MTLD_inhwrd_zonder_abw	Measure of textual lexical diversity voor inhoudswoorden, zonder algemene bijwoorden
177.	Inhwrd_d	Dichtheid van inhoudswoorden
178.	Inhwrd_dz	Aantal inhoudswoorden per deelzin
179.	Inhwrd_d_zonder_abw	Dichtheid van inhoudswoorden, zonder algemene bijwoorden
180.	Inhwrd_dz_zonder_abw	Aantal inhoudswoorden per deelzin, zonder bijwoorden
181.	Vnw_ref_d	Dichtheid van terugverwijzende voornaamwoorden
182.	Vnw_ref_dz	Terugverwijzende voornaamwoorden per deelzin
183.	Arg_over_vzin_d	Dichtheid van argumenten die voorkomen in de vorige zin
184.	Arg_over_vzin_dz	Aantal argumenten die voorkomen in de vorige zin per deelzin
185.	Lem_over_vzin_d	Dichtheid van lemma's die voorkomen in met de vorige zin
186.	Lem_over_vzin_dz	Aantal lemma's die voorkomen in de vorige zin per deelzin
187.	Arg_over_buf_d	Dichtheid van argumenten die voorkomen in de voorgaande X woorden; X is de bufferomvang die in te stellen is. Standaard is die 50.
188.	Arg_over_buf_dz	Aantal argumenten die voorkomen in de voorgaande X woorden per deelzin
189.	Lem_over_buf_d	Dichtheid van lemma's die voorkomen in de voorgaande X woorden
190.	Lem_over_buf_dz	Aantal lemma's die voorkomen in de voorgaande X woorden per deelzin
191.	Onbep_nwg_p	Proportie onbepaalde naamwoordgroepen op naamwoordgroepen
192.	Onbep_nwg_dz	Aantal onbepaalde naamwoordgroepen per deelzin

Over de naam van deze kenmerkgroep

Veel kenmerken uit deze groep gaan over de vraag in hoeverre de tekst woorden herhaalt. Woordherhaling kan worden gezien als een indicatie van 'informatiedichtheid', een term met een informatiekundige achtergrond waarin informatie wordt gedefinieerd in termen van de waarschijnlijkheid van woorden op basis van eerdere woorden.

Vanuit meer praktisch tekstanalytisch perspectief is belangrijk om te weten dat het gebruiken van telkens nieuwe woorden kan duiden op twee zaken:

- De tekst snijdt telkens nieuwe onderwerpen aan (er is dus weinig referentiële coherentie);
- De auteur gebruikt verschillende woorden voor min of meer dezelfde verschijnselen (er is veel lexicale diversiteit).

Vandaar in de titel van hoofdstuk niet wordt gesproken over informatiedichtheid maar over referentiële coherentie en lexicale diversiteit.

Type-token-ratio en zeldzaamheidsindex

De klassieke maat voor informatiedichtheid is de type-token-ratio (TTR), waarbij het aantal verschillende woorden (types) wordt gedeeld op het totaal aantal woorden (tokens). Deze maat kan zowel voor woorden als voor lemma's worden bekeken.

Van belang is hier het onderscheid tussen functiewoorden en inhoudswoorden. In T-Scan worden die categorieën als volgt opgevat:

- functiewoorden zijn voornaamwoorden, lidwoorden, voorzetsels, voegwoorden, telwoorden, hulpwerkwoorden, koppelwerkwoorden en tussenwerpsels;

- inhoudswoorden zijn dan dus naamwoorden, namen, adjectieven, bijwoorden en 'gewone werkwoorden', dat wil zeggen werkwoorden die geen hulpwerkwoord of koppelwerkwoord zijn of kunnen zijn.

Omdat functiewoorden veel herhaald worden en dus de TTR drukken, maar ook weinig onderscheid maken tussen teksten, is het ook informatief om de TTR alleen op inhoudswoorden uit te rekenen.

Ten slotte is er een TTR voor namen toegevoegd. Een tekst met veel namen kan een groot beroep doen op de voorkennis van de lezer, maar alleen als het gaat om veel verschillende namen.

Measure of Lexical Diversity in Text

Een nadeel van de TTR is dat hij gevoelig is voor tekstlengte. Naarmate een tekst langer wordt, worden steeds minder nieuwe woorden toegevoegd. Daarom is het lastig om teksten van verschillende lengtes te vergelijken met de TTR. Daarom hebben McCarthy & Jarvis (2010) een alternatief ontwikkeld, the Measure of Textual Lexical Diversity (MTLD). Ook die wordt berekend voor woorden, voor lemma's, voor inhoudswoorden en voor namen.

De basis voor MTLD vormt de observatie dat naarmate een tekst vordert, de TTR daalt. De eerste 10 woorden zijn vaak nog verschillend (TTR=1), in de volgende 10 woorden zitten meer herhalingen, zodat de TTR lager wordt dan 1. Bij MTLD wordt gekeken hoe lang een tekst er gemiddeld over doet om de TTR onder een bepaalde waarde te brengen. Dat gebeurt door iedere keer dat de TTR onder de ingestelde waarde zakt, hem weer op 1 te zetten. Een tekst passeert zodoende een aantal malen de TTR-drempel.

Een voorbeeld kan dit verhelderen. We nemen de MTLD van het volgende fragment, uitgaande van een drempelwaarde van .72 (dat is de standaardwaarde die aangeraden wordt in de literatuur):

```
Dit is een proefje. Dit is de tweede zin van het proefje.
```

De MTLD werkt zowel 'heen' (voorwaarts (VW) door de tekst) als 'terug' (achterwaarts, AW). De voorwaartse berekening staat in Tabel 3 in de kolommen 2-4, de achterwaartse in de kolommen 5-7; lees die laatste drie kolommen van onder naar boven.

Woord	**Aantal tokens tot zover VW**	**Aantal types tot zover VW**	**TTR tot zover VW**	**Aantal tokens tot zover AW**	**Aantal types tot zover AW**	**TTR tot zover AW**
Dit	1	1	1	12	8	.67 (reset)
Is	2	2	1	11	8	.73
Een	3	3	1	10	8	.8
Proefje	4	4	1	9	7	.78
Dit	5	4	.8	8	7	.88
Is	6	4	.67 (reset)	7	6	.86
De	1	1	1	6	5	.83
Tweede	2	2	1	5	5	1
Zin	3	3	1	4	4	1
Van	4	4	1	3	3	1
Het	5	4	1	2	2	1
Proefje	6	5	1	1	1	1

Tabel 3. De berekening van MTLD

Dit tekstje bereikt in voorwaartse richting in 6 woorden eenmaal de drempel. Na de reset komt de TTR niet meer onder de 1. In achterwaartse richting wordt de drempel ook precies eenmaal bereikt, namelijk bij het 12de woord. Dat betekent dat de tekst in beide richtingen 12 woorden nodig heeft om een maal over de drempel te komen. Dat geeft een MTLD in beide richtingen van 12; gemiddeld 12.

In echte teksten wordt de drempel natuurlijk veel vaker bereikt, en is er op het eind van de tekst een TTR van lager dan 1: een 'rest' dus. Die rest wordt meegenomen in de berekening. Neem een tekst van

90 woorden waarin de drempel vier maal wordt bereikt (4 resets) en waarin de TTR op het eind .86 is. Dat betekent dat die laatste keer de helft van de afstand tussen 1 en de drempelwaarde .72 is overbrugd (.14/.28). Dan gaat T-Scan ervan uit dat de drempel 4,5 maal is bereikt in 90 woorden, wat een MTLD geeft van 20.

De MLTD is een diversiteitsmaat met een heel kort geheugen, omdat bij elke reset de berekening opnieuw begint. Onderzoek van Koizumi (2012) laat zien dat de TTR sterk verschilt naargelang je 50, 100, 200 of 300 woorden uit een tekst neemt, terwijl dat de MLTD veel stabieler is. Wat de meest interessante maat is, wordt bepaald door de onderzoeksvraag; maar veelal zal de MLTD toch de voorkeur genieten.

Argumentoverlap

Argumentoverlap is in T-Scan gedefinieerd als het herhalen van referentiële uitdrukkingen binnen begrensde tekstregio's. Daarbij kan het gaan om herhalingen van uitdrukkingen uit de vorige zin of uit een in te stellen buffer van X woorden; de standaardbuffer telt 50 woorden.

Als referentiële uitdrukkingen ('argumenten') worden daarbij gezien:

- zelfstandige naamwoorden;
- namen;
- hoofdwerkwoorden;
- voornaamwoorden (maar niet aanwijzende). Door met lijstjes voornaamwoorden te werken worden ook twee voornaamwoorden van verschillend type maar in dezelfde persoon meegerekend als overlappende argumenten, bijvoorbeeld *ik* en *mijn* in de zinnen 'Gisteren kocht ik een exemplaar van 'De Avonden'. Ik was erg blij met mijn nieuwe boek.'

Bij de zinsoverlap-maten wordt voor iedere zin nagegaan welk van de argumenten in de vorige zin voorkomt. Komt in een enkelvoudige zin van 5 woorden 1 argument terug uit de vorige zin, dan is het aantal herhaalde argumenten per deelzin 1, en de dichtheid van zinsoverlap 1/5 x 1000 = 200.

Bij de bufferoverlap-maten wordt de buffer als een 'venster' over de tekst heen geschoven, en wordt telkens voor het eerste woord na de buffer de overlap bekeken. In een tekst van 100 woorden met een bufferinstelling van 50 wordt dus gekeken of woord 51 overlap vertoont met een woord uit de woorden 1-50, of woord 52 dat doet met woord 2-51, en zo door tot en met woord 100 versus woord 50-99. Er wordt voor deze maat alleen een waarde op tekstniveau gegeven; alinea's kunnen namelijk soms korter zijn dan 50 woorden. De buffer is standaard ingesteld op 50 woorden, maar kan aangepast worden. Wordt een buffer van 100 woorden gebruikt, dan stijgt de overlap natuurlijk veelal.

Er zijn zowel woord- als lemmavarianten van de overlapmaten beschikbaar.

Inhoudswoorden

Een indicatie van informatierijkdom die niet op woordherhaling is gebaseerd, is het aantal dan wel de proportie inhoudswoorden, ook wel aangeduid als 'lexical density' (zie bv. Johansson 2008). Daarbij gaat het dus om aantal zelfstandige naamwoorden, werkwoorden, adjectieven en bijwoorden per 1000 woorden, of per deelzin.

Terugverwijzende voornaamwoorden

Onder terugverwijzende voornaamwoorden rekenen we voornaamwoorden die naar alle waarschijnlijkheid verwijzen naar eerder genoemde referenten:

- persoonlijke voornaamwoorden van de derde persoon (*hij, zij, ze, hen*; maar niet *men*);
- bezittelijke voornaamwoorden van de derde persoon (*zijn, haar, hun*);
- aanwijzende voornaamwoorden (*die, deze, dit, dat*).

Onbepaalde naamwoordgroepen
Onbepaalde naamwoordgroepen (*een oude man ….*) verwijzen zeker niet altijd naar naar nieuwe referenten, maar wel vaker dan bepaalde naamwoordgroepen (*de oude man …*). Daarom tellen we het aantal onbepaalde naamwoordgroepen per deelzin en delen we het aantal op het totaal aantal naamwoordgroepen.

Informatie op woordniveau over referentiële cohesie en lexicale diversiteit
Op woordniveau is voor deze kenmerkgroep alleen informatie terug te vinden over de vraag of een woord al of niet een terugverwijzend voornaamwoord is (zie kenmerk 35 in paragraaf 3.2 hierboven).

3.6 Relationele coherentie en situatiemodelmaten

193.	Conn_d	Totale dichtheid van temporele, contrastieve, comparatieve en causale connectieven
194.	Conn_dz	Totaal aantal temporele, contrastieve, comparatieve en causale connectieven per deelzin
195.	Conn_TTR	Type-token-ratio voor temporele, contrastieve, comparatieve en causale connectieven
196.	Conn_MTLD	Measure of Lexical Diversity in Text voor temporele, contrastieve, comparatieve en causale connectieven
197.	Conn_temp_d	Dichtheid van temporele verbindingswoorden
198.	Conn_temp_dz	Temporele verbindingswoorden per deelzin
199.	Conn_temp_TTR	Type-token-ratio voor temporele verbindingswoorden
200.	Conn_temp_MTLD	Measure of Lexical Diversity in Text voor temporele verbindingswoorden
201.	Conn_reeks_wg_d	Dichtheid van reeksaanduiders voor woordgroepen
202.	Conn_reeks_wg_dz	Reeksaanduiders per deelzin voor woordgroepen
203.	Conn_reeks_wg_TTR	Type-token-ratio voor reeksaanduiders voor woordgroepen
204.	Conn_reeks_wg_MTLD	Measure of textual lexical diversity voor reeksaanduiders voor woordgroepen
205.	Conn_reeks_zin_d	Dichtheid van reeksaanduiders voor (deel)zinnen
206.	Conn_reeks_zin_dz	Reeksaanduiders per deelzin voor (deel)zinnen
207.	Conn_reeks_zin_TTR	Type-token-ratio voor reeksaanduiders voor (deel)zinnen
208.	Conn_reeks_zin_MTLD	Measure of textual lexical diversity voor reeksaanduiders voor (deelzinnen)
209.	Conn_contr_d	Dichtheid van contrastieve verbindingswoorden
210.	Conn_contr_dz	Contrastieve verbindingswoorden per deelzin
211.	Conn_contr_TTR	Type-token-ratio voor contrastieve verbindingswoorden
212.	Conn_contr_MTLD	Measure of Lexical Diversity in Text voor contrastieve verbindingswoorden
213.	Conn_comp_d	Dichtheid van comparatieve verbindingswoorden
214.	Conn_comp_dz	Comparatieve verbindingswoorden per deelzin
215.	Conn_comp_TTR	Type-token-ratio voor comparatieve verbindingswoorden
216.	Conn_comp_MTLD	Measure of Lexical Diversity in Text voor comparatieve verbindingswoorden
217.	Conn_caus_d	Dichtheid van causale verbindingswoorden
218.	Conn_caus_dz	Causale verbindingswoorden per deelzin
219.	Conn_caus_TTR	Type-token-ratio voor causale verbindingswoorden
220.	Conn_caus_MTLD	Measure of Lexical Diversity in Text voor causale verbindingswoorden
221.	Causaal_d	Dichtheid van causale inhoudswoorden
222.	Ruimte_d	Dichtheid van ruimtewoorden
223.	Tijd_d	Dichtheid van tijdwoorden
224.	Emotie_d	Dichtheid van emotiewoorden
225.	Causaal_TTR	Type-token-ratio voor causale inhoudswoorden
226.	Causaal_MTLD	Measure of textual lexical diversity voor causale inhoudswoorden
227.	Ruimte_TTR	Type-token-ratio voor ruimtewoorden
228.	Ruimte_MTLD	Measure of Lexical Diversity in Text voor ruimtewoorden
229.	Tijd_TTR	Type-token-ratio voor tijdwoorden
230.	Tijd_MTLD	Measure of Lexical Diversity in Text voor tijdwoorden
231.	Emotie_TTR	Type-token-ratio voor emotiewoorden
232.	Emotie_MTLD	Measure of textual lexical diversity voor emotiewoorden

Een voor de hand liggende indicator van relationele coherentie vormen de connectieven van verschillende klassen. T-Scan kijkt naar de volgende groepen verbindingswoorden (zie verder Bijlage C):

- Causale connectieven (incl. conditionele connectieven): *daarom, indien*
- Comparatieve connectieven: *zoals, dan* als voegwoord
- Contrastieve connectieven: *toch, desondanks*
- Opsommende connectieven: *en, daarnaast*
- Temporele connectieven: *voordat, eertijds*

Als connectief worden beschouwd de woorden die veelal complete zinnen in een betekenisrelatie met elkaar plaatsen. Meestal gaat het om voegwoorden en voornaamwoordelijke bijwoorden. Echter, bij de temporele connectieven zijn ook een aantal tijdbijwoorden meegenomen.

Een aantal veel voorkomende maar nogal flexibel inzetbare verbindingswoorden is niet opgenomen, zoals *als*. Dat woord kan zowel temporeel als causaal gebruikt worden.

Een veel voorkomend connectief is *en*. Vaak gaat het daarbij om nevenschikkingen op korte afstand (*appels en peren*), die weinig zeggen over tekstcoherentie. Om daarvoor enigszins te corrigeren, hebben onderscheid gemaakt tussen reeksconnectieven zoals *en,* die veelal woordgroepen verbinden, en reeksconnectieven zoals *bovendien* en *ten tweede,* die vaker gebruikt worden om zinnen of deelzinnen te verbinden. De woordgroepverbinders leveren de kenmerk *conn_reeks_wg_d* en *conn_reeks_wg_dz* op, de frequentie van de andere reeksconnectieven komt naar voren in *conn_reeks_zin_d* en *conn_reeks_zin_dz*. Het gaat hier natuurlijk om een benadering: zekerheid over de aard van de verbinding hebben we niet.

Om niet alleen zicht te krijgen op het aantal maar ook op de diversiteit van verbindingswoorden (en mogelijk dus van relaties), hebben we voor connectieven ook type-token-ratio's en MTLD's berekend. Als een tekst vooral een enkel connectief bevat, zijn die maten dus erg laag.

Naast de afzonderlijke groepen connectieven zijn ook totalen berekend van alle niet-opsommende connectieven (dus causaal, comparatief, contrastief en temporeel), evenals de type-token-ratio's en MTLD's voor deze hele groep.

In de coherentiegroep hebben we ook enkele maten geplaatst gebaseerd op woorden die zich richten op situatiemodel-dimensies. Wie een tekst leest, bouwt een situatiemodel op, waarin op verschillende dimensies de inhoud van de tekst wordt 'bijgehouden': tijd, plaats, causaliteit, intentionaliteit, en personages (Zwaan & Rapp 2006). We hebben lijsten samengesteld waarin voor de eerste drie genoemde dimensies:

- In de tijdwoordenlijst staan woorden die verwijzen naar tijdstippen, periodes en temporele relaties zoals opeenvolging (*continu, vandaag* enz.). Zelfstandige naamwoorden en adjectieven die naar tijd verwijzen zijn buiten beschouwing gelaten, omdat die nog aan de orde komen in de semantische classificaties naar concreetheid (zie 3.7 hierna).
- In de ruimtewoorden-lijst staan woorden die verwijzen naar plaatsen en ruimtelijke relaties en eigenschappen (*krap, dichtbij* enz.). Ook hier zijn weer zelfstandige naamwoorden en adjectieven buiten beschouwing gelaten (zie 3.7 hierna).
- In de causaliteitswoorden-lijst staan woorden die verwijzen naar causale verbanden (*oorzaak, gevolg, aanleiding, effect*, enz.).

Het samenstellen van een lijst met intentionaliteitswoorden leek ons niet eenvoudig. Termen die verwijzen naar personen zijn in andere T-Scankenmerken geïdentificeerd, dus dat leek minder nodig. Wel hebben we nog een lijst van 834 woorden samengesteld die verwijzen naar emoties en andere psychologische kenmerken van mensen, zoals *zwartgallig, weemoedig, wilskrachtig, wanhopig* enz.

Voor de tijd-, ruimte-, causaliteits- en emotiewoorden is ook weer de lexicale diversiteit berekend.

3.7 Semantische klassen, concreetheid en algemeenheid

3.7.1 Zelfstandige naamwoorden

233.	Conc_nw_strikt_p	Proportie van strikt-concrete naamwoorden
234.	Conc_nw_strikt_d	Dichtheid van strikt-concrete naamwoorden
235.	Conc_nw_ruim_p	Proportie van ruim-concrete naamwoorden
236.	Conc_nw_ruim_d	Dichtheid van ruim-concrete naamwoorden
237.	Pers_nw_p	Proportie van naamwoorden verwijzend naar personen
238.	Pers_nw_d	Dichtheid van naamwoorden verwijzend naar personen
239.	PlantDier_nw_p	Proportie van naamwoorden verwijzend naar planten en dieren
240.	PlantDier_nw_d	Dichtheid naamwoorden verwijzend naar planten en dieren
241.	Gebr_vw_nw_p	Proportie van naamwoorden verwijzend naar gebruiksvoorwerpen
242.	Gebr_vw_nw_d	Dichtheid van naamwoorden verwijzend naar gebruiksvoorwerpen
243.	Subst_conc_nw_p	Proportie van naamwoorden verwijzend naar concrete substanties
244.	Subst_conc_nw_d	Dichtheid van naamwoorden verwijzend naar concrete substanties
245.	Voed_verz_nw_p	Proportie van naamwoorden verwijzend naar voeding en verzorging
246.	Voed_verz_nw_d	Dichtheid van naamwoorden verwijzend naar voeding en verzorging
247.	Concr_ov_nw_p	Proportie van overige concrete naamwoorden
248.	Concr_ov_nw_d	Dichtheid van overige concrete naamwoorden
249.	Gebeuren_conc_nw_p	Proportie naamwoorden verwijzend naar concrete gebeurtenissen
250.	Gebeuren_conc_nw_d	Dichtheid naamwoorden verwijzend naar concrete gebeurtenissen
251.	Plaats_nw_p	Proportie van naamwoorden verwijzend naar plaatsen en ruimtes
252.	Plaats_nw_d	Dichtheid van naamwoorden verwijzend naar plaatsen en ruimtes
253.	Tijd_nw_p	Proportie van naamwoorden verwijzend naar tijd
254.	Tijd_nw_d	Dichtheid van naamwoorden verwijzend naar tijd
255.	Maat_nw_p	Proportie van naamwoorden verwijzend naar maten
256.	Maat_nw_d	Dichtheid van naamwoorden verwijzend naar maten
257.	Subst_abstr_nw_p	Proportie van naamwoorden verwijzend naar abstracte substanties
258.	Subst_abstr_nw_d	Dichtheid van naamwoorden verwijzend naar abstracte substanties
259.	Gebeuren_abstr_nw_p	Proportie naamwoorden verwijzend naar abstracte gebeurtenissen
260.	Gebeuren_abstr_nw_d	Dichtheid naamwoorden verwijzend naar abstracte gebeurtenissen
261.	Organisatie_nw_p	Proportie van naamwoorden verwijzend naar organisaties
262.	Organisatie_nw_d	Dichtheid van naamwoorden verwijzend naar organisaties
263.	Ov_abstr_nw_p	Proportie overige abstracte naamwoorden
264.	Ov_abstr_nw_d	Dichtheid overige abstracte naamwoorden
265.	Undefined_nw_p	Proportie van naamwoorden die ongedefinieerd blijven in de lijst
266.	Gedekte_nw_p	Proportie van naamwoorden en namen die in de lijst staan

Zelfstandige naamwoorden zijn opgedeeld in veertien klassen, zie Tabel 4. Uitgangspunt voor deze indeling is de geannoteerde nominalijst uit het Referentie Bestand Nederlands geweest. (Martin & Maks 2005). Deze lijst is echter door H. Pander Maat, N. Dekker en N. van Houten handmatig gecorrigeerd, gehergroepeerd en uitgebreid. Hij telt nu zo'n 83000 woorden, waarvan ruim 61000 samenstellingen. Het ontwerp van de lijst wordt verder toegelicht in Bijlage D. De oorspronkelijke lijst bevat veel woorden met meerdere lezingen (soms wel vijf of zes), die minutieus worden gecatalogiseerd. T-Scan kan echter niet kiezen tussen deze lezingen. Daarom zijn in de T-Scanversie van de lijst polyseme of ambigue woorden ongedefinieerd gelaten, wat een vijftiende groep oplevert, zie Tabel 4.

Klasse	Voorbeelden	Concreet of abstract?
1. Personen	*Leraar, schreeuwlelijk*	Strikt en ruim concreet
2. Planten en dieren	*Mus, eik*	Strikt en ruim concreet
3. Gebruiksvoorwerp	*Stoel, weefgetouw*	Strikt en ruim concreet
4. Concrete substanties	*Modder, kerrie*	Strikt en ruim concreet
5. Voeding en verzorging	*Melk, sigaret, bruistablet*	Strikt en ruim concreet
6. Concreet overig	*Galblaas, vulkaan*	Strikt en ruim concreet
7. Concreet gebeuren	*Aai, ademhaling*	Strikt en ruim concreet
8. Plaats	*Amsterdam, voorkamer*	Ruim concreet
9. Tijd	*Feestdag, periode*	Ruim concreet
10. Maat	*Euro, dB*	Ruim concreet
11. Abstracte substanties	*Fosfor, splijtstof*	Abstract
12. Abstract gebeuren	*Crisis, loonverlaging*	Abstract
13. Organisatie	*Werkgeversorganisatie, NATO*	Abstract
14. Abstract overig	*Christendom, motto*	Abstract
15. Undefined	*Kant, poot*	

Tabel 4. De classificatie van nomina in T-Scan

De eerste zeven klassen worden gezien als concreet in strikte zin, de eerste tien klassen als concreet in ruime zin. T-Scan geeft dichtheden en proporties voor alle veertien klassen, en voor strikt-concrete woorden (alle woorden uit de klassen 1-7) en ruim-concrete woorden (woorden uit de klassen 1-10).

Bij dit alles moeten we bedenken dat T-Scan in eerste instantie alleen lemma's met de woordsoort 'noun' of 'spec' opzoekt in de lijst (zie kenmerken op woordniveau voor een toelichting op de woordsoorten). Bij verkeerde herkenning van de woordsoort wordt er dus geen semantisch label toegekend. De nominalijst bevat wel voornamen, maar geen achternamen. In teksten met veel namen worden dus nogal wat woorden gemist. Daarom hebben we ervoor gezorgd dat ook woorden die door de Frog naamherkenning als persoonsnaam worden herkend, als 'persoon' worden gelabeld, ook al staan ze niet in de lijst. Evenzo krijgen plaatsnamen het label 'plaats', en organisatienamen het label 'organisatie'. Toch is het goed om voor een meer gedifferentieerd beeld van 'persoonlijke elementen' ook te kijken naar het kenmerk pers_namen (zie par. 3.8 over Persoonlijke elementen).

Onder de abstracte gebeurens en de overige abstracte woorden (klassen 12 en 14 in Tabel 4) bevindt zich een kleinere groep woorden die we 'algemene nomina' genoemd hebben. Voorbeelden van dat soort nomina in het Nederlands zijn *idee, methode, resultaat, probleem* en *consensus*. Kenmerken van die woorden is dat ze vaak in de context worden gespecificeerd (Pander Maat 2002 sprak daarom van 'invulelementen'). In die specificatie wordt dus duidelijk om *welk* idee, resultaat enz. het gaat. Flowerdew en Forest (2015) spreken van 'signalling nouns'. In T-Scan vinden wij deze woorden niet alleen interessant vanwege hun tekststructurerende functie maar ook omdat ze een indicatie zijn van de mate waarin de tekst een *algemeen, niet-domeingebonden vocabulaire* hanteert, waarin de nadruk ligt op expliciet argumenteren over het thema. In tegenstelling tot andere abstracte nomina zoals 'vermogensrendementsheffing' zijn algemene nomina niet gebonden aan bepaalde thema's.

Bijlage J geeft details over de manier waarop deze nomina geïdentificeerd zijn. Tabel 5 hieronder geeft een indruk van de semantische categorieën die erin terugkeren. Daaronder volgen de kenmerken die we uit deze groepen destilleren.

Afzonderlijke situaties		**Relaties tussen situaties**	
Groep	***Voorbeeld***	***Groep***	***Voorbeeld***
Belang-interesse	*relevantie*	Additie-alternatief	*combinatie*
Beschrijving	*kenmerk*	Contrast-variatie	*dichotomie*
Bestaan	*aanwezigheid*	Discussie	*debat, weerlegging*
Bewoording	*formulering*	Doelen-bereiken	*ambitie, realisatie*
Concept(systeem)	*categorie*	Handelingen-keuzes	*aanpassing, ingreep*
Feitelijke-juistheid	*debat, weerlegging*	Middel tot doel	*aanwending*
Gebeurtenis	*voorval*	Ontwikkeling-stabiliteit	*aanzet, mijlpaal*
Gedachte-standpunt	*aanname, idee*	Probleem-oplossing	*impasse, soelaas*
Gradatie	*hevigheid*	Redeneren-causaliteit	*betoog, uitzondering*
Informatie	*melding, gegeven*	Structuur	*component*
Interpretatie	*codering, strekking*		
Kennisverwerving	*inzicht, vraagstuk*		
Mogelijkheid	*gelegenheid, kans*		
Toestand	*Omstandigheid*		
Wenselijkheid	*ideaal, schrikbeeld*		

Tabel 5. Groepen algemene nomina

267.	Alg_nw_d	Dichtheid van algemene nomina (totaal)
268.	Alg_nw_p	Proportie van algemene nomina op alle nomina
269.	Alg_nw_afz_sit_d	Dichtheid van algemene nomina rond *afzonderlijke* situaties (zie linker kolom Tabel 5)
270.	Alg_nw_afz_sit_p	Proportie van algemene nomina rond *afzonderlijke* situaties op alle nomina (linker kolom Tabel 5)
271.	Alg_nw_rel_sit_d	Dichtheid van algemene nomina rond *relaties* tussen situaties (zie rechter kolom Tabel 5)
272.	Alg_nw_rel_sit_p	Proportie van algemene nomina rond *relaties* tussen situaties op alle nomina (rechter kolom Tabel 5)
273.	Alg_nw_hand_d	Dichtheid van nomina rond menselijk *handelen*
274.	Alg_nw_hand_p	Proportie van nomina rond menselijk *handelen* op alle nomina Het gaat bij dit en het vorige kenmerk om de groepen: doel en het bereiken daarvan; handelingen en keuzes; middel tot doel; probleem – oplossing
275.	Alg_nw_kenn_d	Dichtheid van nomina rond *kennis* (incl. de juistheid en verwerving daarvan)
276.	Alg_nw_kenn_p	Proportie van nomina rond *kennis* (incl. de juistheid en verwerving daarvan) Het gaat bij de kennisnomina om de groepen: concept(systeem), feitelijke juistheid, gedachte en standpunt. informatie, interpretatie, kennisverwerving, discussie, redeneren en causaliteit
277.	Alg_nw_disc_caus_d	Dichtheid van nomina rond discussie, redeneren en causaliteit
278.	Alg_nw_disc_caus_p	Proportie van nomina rond discussie, redeneren en causaliteit
279.	Alg_nw_ontw_d	Dichtheid van nomina over ontwikkeling en stabiliteit
280.	Alg_nw_ontw_p	Proportie van nomina over ontwikkeling en stabiliteit

3.7.2 Bijvoeglijke naamwoorden

281.	Waarn_mens_bvnw_p	Proportie van bijv. naamwoorden over waarneembare kenmerken van mensen
282.	Waarn_mens_bvnw_d	Dichtheid van bijv. naamwoorden over waarneembare kenmerken van mensen
283.	Emosoc_bvnw_p	Proportie van bijv. naamwoorden over emoties en sociaal gedrag
284.	Emosoc_bvnw_d	Dichtheid van bijv. naamwoorden over emoties en sociaal gedrag
285.	Waarn_nmens_bvnw_p	Proportie van bijv. naamwoorden verwijzend naar waarneembare kenmerken van dingen
286.	Waarn_nmens_bvnw_d	Dichtheid van bijv. naamwoorden die verwijzend naar waarneembare kenmerken van dingen
287.	Vorm_omvang_bvnw_p	Proportie van bijv. naamwoorden verwijzend naar vorm en omvang
288.	Vorm_omvang_bvnw_d	Dichtheid van bijv. naamwoorden verwijzend naar vorm en omvang
289.	Kleur_bvnw_p	Proportie van bijv. naamwoorden verwijzend naar kleur en licht
290.	Kleur_bvnw_d	Dichtheid van bijv. naamwoorden verwijzend naar kleur en licht
291.	Stof_bvnw_p	Proportie van bijv. naamwoorden verwijzend naar stoffen
292.	Stof_bvnw_d	Dichtheid van bijv. naamwoorden verwijzend naar stoffen
293.	Geluid_bvnw_p	Proportie van bijv. naamwoorden verwijzend naar geluid
294.	Geluid_bvnw_d	Dichtheid van bijv. naamwoorden verwijzend naar geluid
295.	Waarn_nmens_ov_bvnw_p	Proportie van bijv. naamwoorden verwijzend naar andere waarneembare kenmerken van dingen
296.	Waarn_nmens_ov_bvnw_d	Dichtheid van bijv. naamwoorden verwijzend naar andere waarneembare kenmerken van dingen
297.	Technisch_bvnw_p	Proportie van bijv. naamwoorden verwijzend naar technisch waarneembare kenmerken van dingen
298.	Technisch_bvnw_d	Dichtheid van bijv. naamwoorden verwijzend naar technisch waarneembare kenmerken van dingen
299.	Tijd_bvnw_p	Proportie van bijv. naamwoorden verwijzend naar tijd
300.	Tijd_bvnw_d	Dichtheid van bijv. naamwoorden verwijzend naar tijd
301.	Plaats_bvnw_p	Proportie van bijv. naamwoorden verwijzend naar plaats en ruimte
302.	Plaats_bvnw_d	Dichtheid van bijv. naamwoorden verwijzend naar plaats en ruimte
303.	Spec_positief_bvnw_p	Proportie van bijv. naamwoorden die specifiek positief evalueren
304.	Spec_positief_bvnw_d	Dichtheid van bijv. naamwoorden die specifiek positief evalueren
305.	Spec_negatief_bvnw_p	Proportie van bijv. naamwoorden die specifiek negatief evalueren
306.	Spec_negatief_bvnw_d	Dichtheid van bijv. naamwoorden die specifiek negatief evalueren
307.	Alg_positief_bvnw_p	Proportie van bijv. naamwoorden die algemeen positief evalueren
308.	Alg_positief_bvnw_d	Dichtheid van bijv. naamwoorden die algemeen positief evalueren
309.	Alg_negatief_bvnw_p	Proportie van bijv. naamwoorden die algemeen negatief evalueren
310.	Alg_negatief_bvnw_d	Dichtheid van bijv. naamwoorden die algemeen negatief evalueren
311.	Alg_ev_zr_bvnw_p	Proportie van bijv. naamwoorden die evalueren zonder richting
312.	Alg_ev_zr_bvnw_d	Dichtheid van bijv. naamwoorden die evalueren zonder richting
313.	Ep_positief_bvnw_p	Proportie van bijv. naamwoorden die epistemisch positief evalueren
314.	Ep_positief_bvnw_d	Dichtheid van bijv. naamwoorden die epistemisch positief evalueren
315.	Ep_negatief_bvnw_p	Proportie van bijv. naamwoorden die epistemisch negatief evalueren
316.	Ep_negatief_bvnw_d	Dichtheid van bijv. naamwoorden die epistemisch negatief evalueren
317.	Ov_abstr_bvnw_p	Proportie van de overige abstracte bijv. naamwoorden
318.	Ov_abstr_bvnw_d	Dichtheid van de overige abstracte bijv. naamwoorden
319.	Spec_ev_bvnw_p	Proportie van bijv. naamwoorden die specifiek evalueren
320.	Spec_ev_bvnw_d	Dichtheid van bijv. naamwoorden die specifiek evalueren
321.	Alg_ev_bvnw_p	Proportie van bijv. naamwoorden die algemeen evalueren
322.	Alg_ev_bvnw_d	Dichtheid van bijv. naamwoorden die algemeen evalueren
323.	Ep_ev_bvnw_p	Proportie van bijv. naamwoorden die epistemisch evalueren
324.	Ep_ev_bvnw_d	Dichtheid van bijv. naamwoorden die epistemisch evalueren
325.	Conc_bvnw_strikt_p	Proportie van strikt-concrete bijv. naamwoorden
326.	Conc_bvnw_strikt_d	Dichtheid van strikt-concrete bijv. naamwoorden
327.	Conc_bvnw_ruim_p	Proportie van ruim-concrete bijv. naamwoorden
328.	Conc_bvnw_ruim_d	Dichtheid van ruim-concrete bijv. naamwoorden
329.	Subj_bvnw_p	Proportie van subjectieve bijv. naamwoorden
330.	Subj_bvnw_d	Dichtheid van subjectieve bijv. naamwoorden die
331.	Undefined_bvnw_p	Proportie van bijv. naamwoorden die in de lijst ongedefinieerd zijn

332.	Gelabeld_bvnw_p	Proportie van bijv. naamwoorden die in de lijst een label krijgen
333.	Gedekte_bvnw_p	Proportie van bijv. naamwoorden die in de lijst staan

De adjectieven uit de oorspronkelijke RBN-lijst zijn voor T-Scan opnieuw geclassificeerd, en er zijn woorden aan toegevoegd. De nieuwe indeling is te vinden in Tabel 6. Zij indeling is gedetailleerder dan die uit het RBN: er zijn onderscheidingen toegevoegd tussen menselijke en niet-menselijke categorieën (1-2 versus 3-4), tussen al of niet zintuiglijk waarneembare kenmerken (3 versus 4), tussen evaluatieve en niet-evaluatieve woorden (7-9 versus 10) en tussen verschillende vormen van evaluatie (7 t/m 9). Meer details over de classificatie zijn te vinden in Bijlage E.

Klasse	**Voorbeelden**
1. Direct waarneembare kenmerken van personen	*doodsbleek, dwergachtig*
2. Emotionele kenmerken en sociaal gedrag	*gegriefd, goedgelovig*
3. Direct waarneembare kenmerken van dingen	*flanellen, geel*
4. Niet-direct waarneembare kenmerken	*teerarm, kiemvrij*
5. Tijd	*voorbijgaand, vrijdags*
6. Plaats	*binnenlands, Gelders*
7. Specifieke evaluatie (positief/negatief)	*onverslijtbaar; lawaaierig*
8. Algemene evaluatie (positief/negatief/zonder richting)	*mooi*; *verwerpelijk; aanmerkelijk*
9. Epistemische evaluatie (positief/negatief)	*steekhoudend; onzinnig*
10. Overige (niet-evaluatieve) abstracte adjectieven	*aanverwant, aandachtig*
11. Ongedefinieerd	*belastbaar, druk, smal*

Tabel 6. T-Scan adjectiefklassen

T-Scan levert proporties en dichtheden voor de 10 klassen, en een proportie voor de laatste ongedefinieerde groep. Verder zijn er enkel groepen onderscheiden op hoger niveau.

- Specifiek oordelende adjectieven (klasse 7; positief of negatief)
- Algemeen oordelende adjectieven (klasse 8; positief, negatief of zonder richting)
- Epistemische adjectieven (klasse 9; positief of negatief)
- Als strikt-concreet worden opgevat de adjectieven uit de klassen 1, 2 en 3.
- Als ruim-concreet vatten we naast 1, 2 en 3 ook de klassen 5 en 6 op.
- Subjectieve adjectieven: dat zijn de klassen 7 t/m 9.

Bij dit alles moeten we bedenken dat T-Scan alleen lemma's met de woordsoort 'adjective' opzoekt in de adjectievenlijst. Bij verkeerde herkenning van de woordsoort wordt er dus geen semantisch label toegekend.

3.7.3 Werkwoorden en totalen voor concreetheid

334.	Conc_ww_p	Proportie van concrete werkwoorden
335.	Conc_ww_d	Dichtheid van concrete werkwoorden op werkwoorden
336.	Abstr_ww_p	Proportie van abstracte werkwoorden
337.	Abstr_ww_d	Dichtheid van abstracte werkwoorden op werkwoorden
338.	Undefined_ww_p	Proportie van werkwoorden die in de lijst ongedefinieerd blijven
339.	Gedekte_ww_p	Proportie van werkwoorden die in de lijst staan
340.	Alg_ww_d	Dichtheid van algemene werkwoorden (totaal)
341.	Alg_ww_p	Proportie van algemene werkwoorden op alle werkwoorden
342.	Alg_ww_afz_sit_d	Dichtheid van algemene werkwoorden rond *afzonderlijke* situaties (zie linker kolom Tabel)
343.	Alg_ww_afz_sit_p	Proportie van algemene werkwoorden rond *afzonderlijke* situaties op alle werkwoorden (linker kolom Tabel 7)
344.	Alg_ww_rel_sit_d	Dichtheid van algemene werkwoorden rond *relaties* tussen situaties (zie rechter kolom Tabel 7)
345.	Alg_ww_rel_sit_p	Proportie van algemene werkwoorden rond *relaties* tussen situaties op alle werkwoorden (rechter kolom Tabel 7)
346.	Alg_ww_hand_d	Dichtheid van werkwoorden rond menselijk *handelen*
347.	Alg_ww_hand_p	Proportie van werkwoorden rond menselijk *handelen* op alle werkwoorden Het gaat bij dit en het voorgaande kenmerk om de volgende groepen: doel en het bereiken daarvan; handelingen en keuzes; middel tot doel; probleem – oplossing
348.	Alg_ww_kenn_d	Dichtheid van werkwoorden rond *kennis* (incl. de juistheid en verwerving daarvan)
349.	Alg_ww_kenn_p	Proportie van werkwoorden rond *kennis* (incl. de juistheid en verwerving daarvan) Het gaat bij dit en het voorgaande kenmerk om de volgende groepen: concept(systeem), feitelijke juistheid, gedachte en standpunt. informatie, interpretatie, kennisverwerving, discussie, redeneren en causaliteit
350.	Alg_ww_disc_caus_d	Dichtheid van werkwoorden rond discussie, redeneren en causaliteit
351.	Alg_ww_disc_caus_p	Proportie van werkwoorden rond discussie, redeneren en causaliteit
352.	Alg_ww_ontw_d	Dichtheid van werkwoorden over ontwikkeling en stabiliteit
353.	Alg_ww_ontw_p	Proportie van werkwoorden over ontwikkeling en stabiliteit
354.	Conc_tot_d	Gesommeerde dichtheid van concrete nomina, adjectieven en werkwoorden
355.	Conc_tot_p	Proportie van concrete woorden op het totaal van nomina, adjectieven en werkwoorden

Er is ook een RBN-lijst van 6.600 werkwoorden. In RBN worden deze niet geannoteerd voor concreetheid. Ten behoeve van T-Scan is dat wel gebeurd, zij het op zeer globale wijze. Daarbij is onder 'concreet' verstaan dat het werkwoord een zintuiglijke voorstelling oproept. Voor details, zie Bijlage F.

Voorbeelden van concreetheidscodes voor werkwoorden staan in de kolommen van Tabel 6. De rijen in deze tabel verwijzend naar de codering naar 'state of affairs'; zie daarover 3.9.2.

Actie/ proces / toestand	**Abstract**	**Concrete**	**Undefined**
Actie	*aanbesteden, afgelasten*	*kwetteren, lassen*	*verfrissen, verlichten*
Proces	*ineenstorten, meemaken*	*doorlekken, openrijten*	*leeglopen, losslaan*
Toestand	*toeschijnen, hopen*	*vriezen, maffen*	*ontbranden*
Actie / proces > proces	*ontkrachten, tekeergaan*	*doorboren, kronkelen*	*breken, neerslaan*
Actie / toestand > ongedefinieerd	*beantwoorden, letten*	*hobbelen*	*paren*
Proces / toestand > ongedefinieerd	*frustreren, meevallen*	*ruiken*	*horen*
Actie / process / toestand > ongedefinieerd	*bijdragen, verschaffen*		*hechten, maken*

Tabel 6. Werkwoorden gecodeerd naar concreetheid en 'state of affairs'

Onder de abstracte werkwoorden is er weer een kleinere groep die extra abstract van aard zijn, en die we 'algemene werkwoorden' genoemd hebben. We hebben bijna 800 van die werkwoorden geïdentificeerd en ze net zo ingedeeld als de algemene nomina. Meer daarover is te lezen in Bijlage K.

Tabel 7 hieronder illustreert de semantische categorieën onder de algemene werkwoorden.

Afzonderlijke situaties		**Relaties tussen situaties**	
Groep	***Voorbeeld***	***Groep***	***Voorbeeld***
Belang-interesse	*interesseren*	Additie-alternatief	*aanvullen*
Beschrijving	*karakteriseren*	Contrast-variatie	*homogeniseren*
Bestaan	*manifesteren*	Discussie	*Betwisten*
Bewoording	*betitelen*	Doelen-bereiken	*implementeren*
Concept(systeem)	*categoriseren*	Handelingen-keuzes	*aanpassing, ingreep*
Feitelijke-juistheid	*aandikken*	Middel tot doel	*verbruiken*
Gebeurtenis	*plaatsvinden*	Ontwikkeling-stabiliteit	*inluiden, evolueren*
Gedachte-standpunt	*gissen, aanprijzen*	Probleem-oplossing	*ondervangen*
Gradatie	*nuanceren*	Redeneren-causaliteit	*betoog, uitzondering*
Informatie	*Melden*	Structuur	*relateren*
Interpretatie	*duiden*		
Kennisverwerving	*nagaan, overzien*		
Mogelijkheid	*bijbrengen*		
Wenselijkheid	*behoren, schenden*		

Tabel 7. Groepen algemene nomina

Totalen voor concreetheid

Ten slotte zijn de dichtheden van strikt-concrete nomina, strikt-concrete adjectieven en concrete werkwoorden gesommeerd tot de totale dichtheid van concrete woorden. Daarnaast is een proportie concrete woorden berekend door de concrete dichtheid te delen op de totale dichtheid van nomina, adjectieven en werkwoorden.

3.7.4 Algemene en specifieke bijwoorden

356.	Alg_bijw_d	Dichtheid van algemene bijwoorden
357.	Alg_bijw_p	Proportie van algemene bijwoorden op bijwoorden
358.	Spec_bijw_d	Dichtheid van specifieke bijwoorden
359.	Spec_bijw_p	Proportie van specifieke bijwoorden op bijwoorden
360.	Gedekte_bw_p	Proportie van bijwoorden die in de lijst staan

Hoewel T-Scan ze als functiewoord ziet, nemen bijwoorden eigenlijk een tussenpositie in tussen functiewoorden en inhoudswoorden. Ten eerste is het aantal bijwoorden niet heel groot, maar toch groter dan dat van de typische functiewoorden: onze lijst telt er ruim 900. Ten tweede hebben sommige bijwoorden een specifieke inhoudelijke betekenis, terwijl andere algemeen van strekking zijn. Met 'algemeen' bedoelen we in dit verband dat het bijwoord met willekeurig welke propositionele inhoud gecombineerd kan worden. Voor specifieke bijwoorden geldt dat niet: zo beperkt een tijdbijwoord zich tot standen van zaken die in de tijd te plaatsten zijn, en geven de meeste bijwoorden van wijze nadere informatie over de uitvoering van handelingen. In de Tabel hieronder worden veertien typen bijwoorden onderscheiden, waarvan de eerste negen als algemeen worden opgevat, evenals de meerduidige groep.

In wat ruimere zin zijn bijwoorden van plaats, richting en tijd ook algemeen. Ten eerste gaat het om heel kleine, moeilijk uit te breiden woordensets; ten tweede zijn plaats, richting en tijd variabelen die toepasbaar zijn op erg veel standen van zaken. Bij de beperkte definitie van inhoudswoorden (zie

daarover 3.3) zijn daarom zijn ook deze bijwoorden buiten beschouwing gelaten. Alleen bijwoorden van wijze zijn in die definitie nog als inhoudswoord gezien.

Nr.	Type	Voorbeelden (evt. toelichting)	Algemeen / specifiek
1	Markering coherentierelatie	*Aldus, allereerst, bijgevolg, waartoe*	Algemeen
2	Abstract anaforisch	*Ertegen, hierover, waarvan*	Algemeen
3	Overig algemeen	*Enzovoorts, overeen*	Algemeen
4	Kwantificerend*	*Ongeveer, telkenmale*	Algemeen
5	Modaal	*Desnoods, liever, mogelijkerwijs*	Algemeen
6	Modaal of focuspartikel	*Eens, even, zelfs*	Algemeen
7	Negatiewoord	*Evenmin, niet, geenszins*	Algemeen
8	Graadaanduiding	*Allerwegen, enigszins, louter*	Algemeen
9	Tussenwerpsel	*Alstublieft, insgelijks, komaan* (door FROG vaak gezien als bijwoord)	Algemeen
10	Plaats of richting	*Achterin, huiswaarts*	Specifiek
11	Tijd	*Almaar, bijtijds, eergisteren*	Specifiek
12	Plaats/richting of tijd	*Achterna, dichterbij*	Specifiek
13	Wijze	*Desgevraagd, droogweg, gewapenderhand, ongelukkigerwijze*	Specifiek
14	Meerduidig	*Alleen* (kan 1, 6, of 13 zijn), *eerder* (1 of 11), *nader* (3 of 10)	Algemeen

Tabel 8. Semantische indeling van bijwoorden
* Dit soort bijwoorden verwijst naar de omvang van verzamelingen entiteiten, tijdstippen of plaatsen, of kwalificeert hoeveelheden.

Wat kun je nu hebben aan dit onderscheid? Ten eerste geven de algemene bijwoorden een indruk van de aandacht die de auteur besteedt aan het pragmatisch aanvaardbaar maken van zijn uiting. Maar een belangrijker reden in T-Scan is dat we soms willen afzien van deze woorden, namelijk wanneer we geïnteresseerd zijn in het aantal bijwoordelijke bepalingen. We willen in staat zijn om de algemene bijwoorden daarvan af te trekken en ons zo te beperken tot 'inhoudelijke' bepalingen. Om deze reden hebben we de bijwoorden alleen geteld wanneer ze een bijwoordelijke bepaling vormen of het hoofd daarvan zijn. Dat betekent dat *gisteren* in *het feest gisteren* niet meegenomen wordt, omdat het bijvoeglijk functioneert; en *bijna* in *ik ben bijna klaar* modificeert niet het werkwoord maar een predicaatscomplement, en blijft daarom ook buiten beschouwing.

3.8 Persoonlijke elementen

361.	Pers_ref_d	Dichtheid van verwijzingen naar personen
362.	Pers_vnw1_d	Dichtheid persoonlijke en bezittelijke voornaamwoorden eerste persoon
363.	Pers_vnw2_d	Dichtheid persoonlijke en bezittelijke voornaamwoorden tweede persoon
364.	Pers_vnw3_d	Dichtheid persoonlijke en bezittelijke voornaamwoorden derde persoon
365.	Pers_vnw_d	Dichtheid van alle persoonlijke en bezittelijke voornaamwoorden

T-Scan geeft dichtheden voor verschillende soorten verwijzingen naar personen. Allereerst is er een algemeen kenmerk dat verschillende persoonsverwijzingen bijeen neemt:

- Persoonlijke en bezittelijke voornaamwoorden
- Zelfstandige naamwoorden die naar een mens verwijzend (*bakker, uilskuiken*)
- Persoonsnamen (dus wel *Piet*, niet *Volvo*)

De voornaamwoorden die worden meegeteld als persoonlijk element kunnen zowel persoonlijke als bezittelijke voornaamwoorden zijn, zolang ze verwijzen naar personen. *Het* en *er* zijn niet meegeteld. Omdat het gaat om indicaties van het persoonlijke karakter van een tekst, is ook *men* buiten beschouwing gelaten, omdat het in principe naar niet-identificeerbare personen verwijst.

3.9 Andere lexicale informatie

3.9.1 Namen

366.	Pers_namen_p	Proportie van persoonsnamen op alle namen
367.	Pers_namen_p2	Proportie van persoonsnamen op alle namen en naamwoorden
368.	Pers_namen_d	Dichtheid van persoonsnamen
369.	Plaatsnamen_d	Dichtheid van plaatsnamen
370.	Org_namen_d	Dichtheid van organisatienamen
371.	Prod_namen_d	Dichtheid van productnamen
372.	Event_namen_d	Dichtheid van evenementnamen

Op basis van de Named Entity Recognition in Frog onderscheidt T-Scan tussen persoons-, plaats-, organisatie-, product- en evenementnamen. We moeten wel bedenken dat voor die laatste twee categorieën de kwaliteit van de herkenning wat minder goed is (Desmet & Hoste 2013).

Meer algemeen merken we op dat de herkenning van namen vrij sterk leunt op de aanwezigheid van hoofdletters. Daarom is voorzichtigheid nodig in het aanbieden van teksten met een afwijkend hoofdlettergebruik. In sommige juridische contexten is het bijvoorbeeld gebruik om allerlei termen van een hoofdletter te voorzien (bv. *Participatie, Toonder*). Dat kan in T-Scan leiden tot een overschatting van het aantal namen in de tekst. Omdat die namen soms als persoonsnamen worden beschouwd, kan dit ook weer leiden tot een overschatting van het aantal persoonsverwijzingen in de tekst. Teksten met afwijkend hoofdlettergebruik kunnen daarom beter vooraf 'genormaliseerd' worden.

3.9.2 Werkwoordkenmerken

373.	Actieww_p	Proportie van actiewerkwoorden op werkwoorden
374.	Actieww_d	Dichtheid van actiewerkwoorden
375.	Toestww_p	Proportie van toestandswerkwoorden op werkwoorden
376.	Toestww_d	Dichtheid van toestandswerkwoorden
377.	Proceesww_p	Proportie van proceswerkwoorden op werkwoorden
378.	Proceesww_d	Dichtheid van proceswerkwoorden
379.	Undefined_ATP_ww_p	Proportie van werkwoorden die ongedefinieerd blijven voor het kenmerk Actie/Proces/Toestand (ATP)
380.	Ww_tt_p	Proportie van tegenwoordige tijden op alle persoonsvormen
381.	Ww_tt_d	Dichtheid van werkwoorden in tegenwoordige tijd
382.	Ww_mod_d	Dichtheid van modale werkwoorden
383.	Ww_mod_dz	Aantal modale werkwoorden per deelzin
384.	Huww_tijd_d	Dichtheid van hulpwerkwoorden van tijd
385.	Huww_tijd_dz	Aantal van hulpwerkwoorden van tijd per deelzin
386.	Koppelww_d	Dichtheid van koppelwerkwoorden
387.	Koppelww_dz	Gemiddeld aantal koppelwerkwoorden per deelzin
388.	Infin_bv_d	Dichtheid van bijvoeglijke infinitieven
389.	Infin_bv_dz	Bijvoeglijke infinitieven per deelzin
390.	Infin_nw_d	Dichtheid van naamwoordelijke infinitieven
391.	Infin_nw_dz	Naamwoordelijke infinitieven per deelzin
392.	Infin_vrij_d	Dichtheid van vrijstaande infinitieven
393.	Infin_vrij_dz	Vrijstaande infinitieven per deelzin
394.	Vd_bv_d	Dichtheid van bijvoeglijke voltooid deelwoorden
395.	Vd_bv_dz	Bijvoeglijke voltooid deelwoorden per deelzin
396.	Vd_nw_d	Dichtheid van naamwoordelijke voltooid deelwoorden
397.	Vd_nw_dz	Naamwoordelijke voltooid deelwoorden per deelzin
398.	Vd_vrij_d	Dichtheid van vrijstaande voltooid deelwoorden
399.	Vd_vrij_dz	Vrijstaande voltooid deelwoorden per deelzin
400.	Ovd_bv_d	Dichtheid van bijvoeglijke onvoltooid deelwoorden

401.	Ovd_bv_dz	Bijvoeglijke onvoltooid deelwoorden per deelzin
402.	Ovd_nw_d	Dichtheid van naamwoordelijke onvoltooid deelwoorden
403.	Ovd_nw_dz	Naamwoordelijke onvoltooid deelwoorden per deelzin
404.	Ovd_vrij_d	Dichtheid van vrijstaande onvoltooid deelwoorden
405.	Ovd_vrij_dz	Vrijstaande onvoltooid deelwoorden per deelzin

Soorten 'state of affairs'
Deze kenmerken gaan over de 'state of affairs' (SoA) waaraan een werkwoord refereert. De aan het Referentie Bestand Nederlands ontleende lijst van werkwoorden bevat een SoA-classificatie, waarbij onderscheiden wordt tussen acties, processen en toestanden. Deze classificatie is handmatig gecontroleerd door H. Pander Maat. Meer details daarover zijn te vinden in Bijlage G.

Werkwoorden die meerdere lezingen hebben, zijn meestal in de categorie 'ongedefinieerd' geplaatst, met uitzondering van de werkwoorden die zowel een proces- als een actielezing toelaten. Die werkwoorden zijn als proces gecodeerd. Voorbeelden van werkwoordcoderingen zijn te vinden in Tabel 5 hierboven.

Tijden
Op basis van Frog-informatie is onder de persoonsvormen het aantal tegenwoordige-tijdsvormen (proportie op de persoonsvormen; dichtheid) berekend. Daarbij moeten we bedenken dat ook de hulpwerkwoorden hierbij meetellen, onder andere die van tijd. Een groot aantal tegenwoordige tijden kan dus gepaard gaan met een tekst die volledig in de voltooide tijd geschreven is.

Modale werkwoorden en hulpwerkwoorden van tijd
T-Scan geeft op basis van Frog dichtheden en tellingen per deelzin voor modale werkwoorden (zowel zelfstandig gebruikte als modale hulpwerkwoorden) en voor hulpwerkwoorden van tijd. Hierbij tellen ook infinitieven mee.

We merken op dat alle vormen van *zullen* worden opgevat als hulpwerkwoorden van tijd, ook in zinnen als 'ik zou het niet doen'. Een ander probleem is dat *hebben* en *zijn* soms ten onrechte als hulpwerkwoord van tijd worden gezien in gevallen waarin het gaat om 'zijn' als koppelwerkwoord en 'hebben' als hoofdwerkwoord.

Koppelwerkwoorden
Ook voor koppelwerkwoorden geeft T-Scan dichtheden en tellingen per deelzin. Kleinschalige tests leren dat daarbij soms gevallen van *schijnen* ('de zon schijnt fel') en *heten* ('hij heet Peter') ten onrechte als koppelwerkwoord worden gezien. Daarentegen wordt weer goed onderscheid gemaakt tussen *zijn* in 'hij is gek' (koppelwerkwoord) en 'hij is op kantoor' (geen koppelwerkwoord); evenzo voor *blijven* in 'hij blijft op de hoogte' (koppelwerkwoord) en 'het blijft maar stormen' (geen koppelwerkwoord); en voor *lijken* in 'het huis leek onbewoond' (koppelwerkwoord) en 'hij lijkt op zijn broer' (geen koppelwerkwoord).

Voorkomen als koppelwerkwoord wordt soms wel ('het komt me voor dat hij intelligent is') en soms niet ('hij komt me intelligent voor') niet herkend. Evenzo wordt *dunken* wordt soms wel ('het dunkt me dat …'; 'dat is onzin, dunkt me') en soms niet herkend ('dat dunkt me geloofwaardig').

Niet-vervoegde werkwoorden (deelwoorden en infinitieven)
T-Scan treft in teksten zowel vervoegde werkwoorden (persoonsvormen) aan als niet-vervoegde werkwoorden: infinitieven, voltooide deelwoorden en tegenwoordige deelwoorden. De werkwoordskenmerken worden in het algemeen berekend op al deze vormen, net als de later te behandelen woordsoortdichtheid 'werkwoord' (zie 3.9.4). Dat betekent dat bijvoorbeeld ook bijvoeglijk gebruikte infinitieven en deelwoorden meewegen in deze kenmerken.

Om de gebruiker de kans te geven om het aandeel van verschillende soorten niet-vervoegde werkwoorden in te schatten, zijn kenmerken gebouwd voor negen verschillende vormen. Daarbij blijkt Frog helaas niet altijd betrouwbaar; waar nodig wordt dat gemeld.

- Bijvoeglijk gebruikte infinitieven (*de te <u>lezen</u> post*); in tests wordt deze vorm echter vaak ten onrechte als 'vrij' beschouwd.
- Naamwoordelijk gebruikte infinitieven (*het <u>lezen</u> van post vind ik vervelend*)
- 'Vrij' gebruikte infinitieven (*hij zit te <u>lezen</u>*)
- Bijvoeglijk gebruikte voltooid deelwoorden (*de <u>geschilderde</u> muur*)
- Naamwoordelijk gebruikte voltooid deelwoorden (*de <u>verworpenen</u> der aarde*); deze vorm echter wordt door Frog niet betrouwbaar herkend
- 'Vrij' gebruikte deelwoorden, die het hoofdwerkwoord zijn in de zin (*de muur is <u>geschilderd</u>*)
- Bijvoeglijk gebruikte tegenwoordige deelwoorden (*de <u>fluitende</u> vogel*)
- Naamwoordelijk gebruikte tegenwoordige deelwoorden (*de <u>fluitende</u> vind ik het mooist*); deze vorm echter wordt door Frog niet betrouwbaar herkend
- 'Vrij' gebruikte deelwoorden (*<u>fluitend</u> liep hij over straat*)

Hoewel de maten voor naamwoordelijke deelwoorden en voor bijvoeglijke infinitieven dus niet betrouwbaar zijn, geven deze kenmerken een goede indruk van de overige klassen.

3.9.3 Imperatieven, ellipsen en vragen

406.	Imp_ellips_p	Proportie van gebiedende wijzen op persoonsvormen/deelzinnen
407.	Imp_ellips_d	Dichtheid van gebiedende wijzen
408.	Vragen_p	Proportie van vragen op zinnen
409.	Vragen_d	Dichtheid van vragen

Het zou heel goed zijn als T-Scan werkwoorden in de gebiedende wijs zou herkennen, maar dat kan niet. Wat wel kan, is Alpino laten zoeken naar persoonsvormen zonder onderwerp. Dat zijn soms zinnen met imperatieven, maar soms ook elliptische zinnen ('eerst de uien fruiten'; 'heb de hele dag gewerkt'). Corpusonderzoek zal moeten uitwijzen of het bij dit soort zinnen in overwegende mate gaat om imperatieven dan wel om ellipsen.

Vragen worden herkend op basis van een vraagteken. Deze triviale methode is beter dan hij lijkt, tenminste zolang we bedenken dat we op deze wijze eerder vragen (de taalhandeling) dan vraagzinnen identificeren. Zo hebben de volgende vraagzinnen geen vraagteken, maar het kan worden betwijfeld of het om vragende taalhandelingen gaat:

- Dacht ik het niet!
- Hoe is het mogelijk!

Omgekeerd kan een vraagteken een mededeling tot vraag maken:

- Ik vroeg me af of je nog meeging vanavond?

3.9.4 Woordsoorten

410.	Bvnw_d	Dichtheid van bijvoeglijke naamwoorden
411.	Vg_d	Dichtheid van voegwoorden
412.	Vnw_d	Dichtheid van voornaamwoorden
413.	Lidw_d	Dichtheid van lidwoorden
414.	Vz_d	Dichtheid van voorzetsels
415.	Bijw_d	Dichtheid van bijwoorden
416.	Tw_d	Dichtheid van telwoorden
417.	Nw_d	Dichtheid van zelfstandige naamwoorden
418.	Ww_d	Dichtheid van werkwoorden
419.	Tuss_d	Dichtheid van tussenwerpsels
420.	Spec_d	Dichtheid van namen en andere speciale typen woorden
421.	Interp_d	Dichtheid van interpunctietekens

Bij de Frog-herkenning van woordsoorten moeten we bedenken dat deze (terecht) geen rekening houdt met syntactische posities. Daarom kunnen:

- onder werkwoorden ook niet-vervoegde vormen vallen;
- onder bijvoeglijke naamwoorden en telwoorden zowel predicaatsnomina, bijvoeglijke bepalingen als bijwoordelijke bepalingen vallen;
- en voorzetsels zowel voorzetselvoorwerpen, bijvoeglijke bepalingen als bijwoordelijke bepalingen inleiden.

De dichtheden worden bepaald door per eenheid de frequentie van een woordsoort te delen op het totaal van de woordsoortfrequenties, waarin leestekens niet zijn opgenomen. De dichtheid van leestekens wordt, paradoxalerwijze, ook op deze manier bepaald: het gaat hier dus om de verhouding tussen leestekens en woorden.

Bij de dichtheid van 'speciale typen woorden' gaat het meestal (80-90%) om namen; verder gaat het om buitenlandse woorden, cijfers, tijden, romeinse cijfers, URL's, tekens en om letterreeksen die abusievelijk niet herkend zijn als woord, zoals *da's* en *enz*.

3.9.5 Afkortingen

422.	Afk_d	Dichtheid van alle afkortingen tezamen
423.	Afk_gen_d	Dichtheid van generieke afkortingen
424.	Afk_int_d	Dichtheid van internationale afkortingen
425.	Afk_jur_d	Dichtheid van juridische afkortingen
426.	Afk_med_d	Dichtheid van medische afkortingen
427.	Afk_ond_d	Dichtheid van onderwijsafkortingen
428.	Afk_pol_d	Dichtheid van politieke afkortingen
429.	Afk_ov_d	Dichtheid van afkortingen overig
430.	Afk_zorg_d	Dichtheid van zorgafkortingen

T-Scan identificeert afkortingen aan de hand van een lijst met 1725 items, onderscheiden naar 'domein':

- Generiek: niet-domeinspecifieke afkortingen (a.s., a.u.b.)
- Internationaal: afkortingen verwijzend naar nationaliteiten (BE, UK) of internationale organisaties (OPEC, IMF)
- Juridisch: afkortingen verwijzend naar wetten, regelingen en juridische organisaties (AAW, Anw)
- Medisch: afkortingen verwijzend naar aandoeningen (add, ALS)
- Onderwijs (HEAO, DUO)
- Politiek: afkortingen verwijzend naar overheids- of politieke organisaties (AIVD, DS'70)
- Zorg: afkortingen verwijzend naar organisaties in de zorg (BION, KNMP)
- Overig: afkortingen die wel domeinspecifiek zijn maar niet benoemd naar domein (ADSL, KEMA, pdf)

Naast de klassen wordt een dichtheid voor alle afkortingen tezamen gegeven (afk_d).

3.9.6 Voorzetseluitdrukkingen en oude naamvalsvormen

431.	Vzu_d	Dichtheid van voorzetseluitdrukkingen
432.	Vzu_dz	Aantal voorzetseluitdrukkingen per deelzin
433.	Arch_d	Dichtheid van archaïsche naamvalsvormen

T-Scan identificeert voorzetseluitdrukkingen aan de hand van een lijst met 101 items als *ten behoeve van* en *in tegenstelling tot*; zie Bijlage H.

Verder worden oude naamvalsvormen (zoals *des* en *mijner*) geteld op basis van voornaamwoorden die door Frog als genitief of datief gemarkeerd worden.

3.9.7 Intensiveerders

Intensiveerders zijn woorden en uitdrukkingen die een hoge graad van een eigenschap aangeven of de interpretatie versterken van de uiting waarin ze staan. Bij het tellen van intensiveerdersintensiveerder put T-Scan uit een lijst van ongeveer 3700 sterke uitdrukkingen. De laatste versie van de lijst telt ongeveer 1120 adjectieven (bv. *zielsgelukkig*), 35 adjectieven die in 'bijwoordelijk' gebruik een versterker zijn (*knap*), zo'n 125 bijwoorden (*zienderogen*), 220 combinaties (*zeker en vast*), ongeveer 1535 nomina (*zenuwpees, stortregen*), 650 werkwoorden (*wemelen*) en zo'n 35 tussenwerpsels (*ammehoela*). Bij enkele van de adjectieven gaat het om woorden die ook als nomen voorkomen (*reactionair*), maar omdat de adjectieflezingen blijkens het SoNaR-corpus frequenter zijn, zijn deze woorden in de intersiveerderslijst als adjectief gelabeld; dit om de analyse zo eenvoudig mogelijk te houden.

Op basis van de lijst worden de volgende kenmerken gedefinieerd; zie voor een toelichting Bijlage I.

434.	Int_d	Dichtheid van alle intensiveerders uit de lijst bij elkaar
435.	Int_bvnw_d	Dichtheid van de intensiverende adjectieven
436.	Int_bvbw_d	Dichtheid van de intensiverende adjectieven die bijwoordelijk worden gebruikt
437.	Int_bw_d	Dichtheid van de intensiverende bijwoorden
438.	Int_combi_d	Dichtheid van de intensiverende woordcombinaties
439.	Int_nw_d	Dichtheid van de intensiverende naamwoorden
440.	Int_tuss_d	Dichtheid van de intensiverende tussenwerpsels
441.	Int_ww_d	Dichtheid van de intensiverende werkwoorden

3.10 Probabiliteitsmaten

442.	Log_prob_fwd	Logaritme van de voorwaartse trigram-probabiliteit
443.	Log_prob_fwd_inhwrd	Logaritme van de voorwaartse trigram-probabiliteit, alleen inhoudswoorden
444.	Log_prob_fwd_zn	Logaritme van de voorwaartse trigram-probabiliteit, zonder namen
445.	Log_prob_fwd_inhwrd_zn	Logaritme van de voorwaartse trigram-probabiliteit, alleen inhoudswoorden en zonder namen
446.	Entropie_fwd	Entropie, voorwaarts
447.	Entropie_fwd_norm	Entropie, voorwaarts, gecorrigeerd voor zinslengte
448.	Perplexiteit_fwd	Perplexiteit, voorwaarts
449.	Perplexiteit_fwd_norm	Perplexiteit, voorwaarts, gecorrigeerd voor zinslengte
450.	Log_prob_bwd	Logaritme van de achterwaartse trigram-probabiliteit
451.	Log_prob_bwd_inhwrd	Logaritme van de achterwaartse trigram-probabiliteit, alleen inhoudswoorden
452.	Log_prob_bwd_zn	Logaritme van de achterwaartse trigram-probabiliteit, zonder namen
453.	Log_prob_bwd_inhwrd_zn	Logaritme van de achterwaartse trigram-probabiliteit, alleen inhoudswoorden en zonder namen
454.	Entropie_bwd	Entropie, achterwaarts
455.	Entropie_bwd_norm	Entropie, achterwaarts, gecorrigeerd voor zinslengte
456.	Perplexiteit_bwd	Perplexiteit, achterwaarts
457.	Perplexiteit_bwd_norm	Perplexiteit, achterwaarts, gecorrigeerd voor zinslengte

Hoe minder waarschijnlijk een woord of een tekstfragment, hoe lastiger het waarschijnlijk te verwerken zal zijn. Daarom is het interessant om iets te weten over de probabiliteit van woorden en zinnen. T-Scan biedt op dat punt drie maten, alle drie ontleend aan WOPR (Berck & Van den Bosch 2009). WOPR is een taalmodel dat op elke tekstplaats het volgende woord voorspelt. T-Scan maakt gebruikt van een WOPR model dat is getraind op het krantendeel van het SoNaR-corpus.

Allereerst geeft T-Scan de voorwaartse *trigram-probabiliteit*: dat wil zeggen de kans dat een woord (of een leesteken) zich voordoet, afgaand op de twee woorden die eraan voorafgaan. Van die waarschijnlijkheid wordt de logaritme genomen. Het laatste woord van de zin *ik houd van voetbal* is bijvoorbeeld waarschijnlijker dan dat in *ik houd van kasten*, zonder dat *voetbal* frequenter is dan *kasten*. De voorwaartse 'logprob' wordt ook gegeven voor alleen de inhoudswoorden, zonder de namen mee te tellen, en met beide beperkingen tegelijk.

Probabiliteiten kunnen ook worden berekend op basis van de woorden die volgen op het woord in kwestie; we spreken dan van achterwaartse probabiliteiten, die ook weer in vier varianten worden gegeven.

De probabiliteiten van woorden worden op hogere tekstniveaus verwerkt tot gemiddelden. Op zinsniveau wordt het gemiddelde genomen van de woorden, op tekstniveau het gemiddelde van de zinnen.

T-Scan geeft ook in beide richtingen de *entropie* en de *perplexiteit*; dat gebeurt alleen op zins- en hogere niveaus. Entropie is een maat voor onzekerheid in een taal. Hoe onverwachter een taaluiting is, hoe hoger de entropie ervan. In de informatietheorie wordt de entropie ook wel gezien als het aantal bits dat nodig is om bepaalde informatie te coderen. Zijn er twee mogelijke gebeurtenissen, dan volstaat 1 bit. Bij vier mogelijkheden zijn 2 bits nodig, bij acht mogelijkheden 3, enzovoort. Het aantal mogelijkheden wordt ook wel de perplexiteit van het model genoemd. Algemeen geldt: als je 2^{entropie} (2 tot-de-macht-entropie) neemt, krijg je de perplexiteit. En andersom is de entropie de logaritme met grondtal 2 van de perplexiteit. Voor een simpele uitleg, zie https://www.nemokennislink.nl/publicaties/de-voorspelbaarheid-van-taal/; iets minder simpel is https://en.wikipedia.org/wiki/Perplexity.

Nu zijn er twee verfijningen nodig. Ten eerste zijn niet alle gebeurtenissen even waarschijnlijk: het ene woord is veel frequenter dan het andere; en dat geldt ook voor bepaalde opeenvolgingen van woorden (een zelfstandig naamwoord wordt vaak voorafgegaan door *de*). Als je wilt berekenen hoe

waarschijnlijk een zin is, moet je daarom iets weten over de frequentie van woorden en woordopeenvolgingen. In T-Scan baseren we ons zoals gezegd op een WOPR-taalmodel op basis van het SoNaR-krantencorpus. Het is belangrijk om te weten dat WOPR niet verder kijkt dan twee woorden naar links en naar rechts: het hanteert een zogenaamd trigram-model (Berck & Van den Bosch 2009). De waarschijnlijkheid van woorden wordt dus vrij lokaal bepaald.

Ten tweede groeit het aantal mogelijke sequenties naarmate een sequentie langer wordt. Daarom wordt de waarschijnlijkheid van een bepaalde sequentie kleiner naarmate deze langer wordt. En daarom is ook de entropie van een lange zin automatisch hoger dan die van een korte zin. De correlatie tussen entropie en zinslengte is in grotere corpora erg hoog: boven de .90. Wie zinnen wil vergelijken op entropie, moet daarom delen door het aantal woorden in de zin. Voor meer uitleg informatie over het rekening houden met frequenties en het normaliseren, zie Goldsmith (2007).

In T-Scan werkt, in navolging van WOPR, met entropie en perplexiteit op het niveau van de zin. Deze maten worden op tekst- en paragraafniveau dus niet apart bepaald; in plaats daarvan worden de gemiddelden van de waarden voor de zinnen gegeven. Daarom kan de tekst-entropie beter genormaliseerd worden door te delen door de gemiddelde zinslengte dan door te delen door het aantal tekstwoorden. Omdat perplexiteit een exponentiële functie is van de entropie, is het denkbaar om deze te niet te delen de zinslengte, maar door het kwadraat daarvan.

Voor wie meer geïnteresseerd is in de werking van WOPR, citeren we hieronder Van den Bosch en Berck (2009, 23):

> To calculate the perplexity of a sentence, we feed it to WOPR and see which words it predicts for each word in the sentence. The perplexity is calculated from the estimated probabilities of each prediction. A prediction is a classification by IGTREE based on a local context of preceding words. In contrast with how IGTREE is used in the WOPR translation module, the word predictor classifier in WOPR produces class distributions (with more than one class if the classification occurs at a non-ending node).
>
> Thus, WOPR usually returns more than one word for a given sequence, together with a probability based on frequency counts. This distribution of possible answers is used to calculate a perplexity value. There are three possibilities: (1) If the distribution returned by IGTREE contains the correct word, we take the probability of the word in the distribution; (2) If the distribution does not contain the correct word, we check if it is in the lexicon. If it is, the lexical probability is taken; (3) If it is not in the lexicon, a probability for unseen items is used that is estimated through Good-Turing smoothing. WOPR calculates the sum of –p log2(p) of all the probabilities (one for each word in the sentence) (…). The perplexity value is two to the power of this sum.

Twee opmerkingen hierbij. IGTREE is een algoritme voor het weergeven van opties in beslissingsbomen (zie Daelemans, Van de Bosch & Weijters 1997). En in de oorspronkelijke setting van WOPR wordt de som van de probabiliteiten gedeeld door het aantal woorden in de zin. In T-Scan gebeurt dit vooralsnog niet.

3.11 Eigen classificatie

458.	Eigen classificatie	Komt het woord voor in een zelf ingevoerde lijst? Of: / hoeveel woorden uit deze zin/alinea/tekst komen voor in die lijst?

T-Scan biedt de optie om een eigen woordclassificatie uit te voeren op basis van een woordenlijst die je zelf invoert. T-Scan kijkt dan voor ieder woord uit de tekst of het voorkomt op jouw lijst. Zo ja, dan krijgt het woord een ‘1’ bij het kenmerk ‘eigen classificatie’, zo nee dan krijgt het daar een ‘0’.

4. Kenmerken op woordniveau

Kenmerken op hoger niveau zijn vaak gebaseerd op kenmerken die op woordniveau worden toegekend. Om te kunnen zien welke woordkenmerken dat zijn, levert T-Scan ook output op woordniveau. Die output is kwalitatief van aard: het gaat om nominale variabelen. Men kan daarin zien welke woorden bijdragen aan een bepaalde hoog of laag scorende maat.

In onderstaand overzicht vermelden we in de rechterkolom de groepen waar een kenmerk bij hoort:

0. Algemeen
1. Woordmoeilijkheid
2. Zinscomplexiteit (hieronder niet relevant)
3. Referentiële coherentie en woordenrijkdom (hieronder niet relevant)
4. Relationele coherentie
5. Semantische klassen en woordconcreetheid
6. Persoonlijke elementen
7. Andere informatie over woorden en uitdrukkingen
    a. Namen
    b. Werkwoordkenmerken
    c. Woordsoorten (POS-tags)
    d. Afkortingen
8. Probabiliteitsmaten
9. Intensiveerders

Nr	Naam	Toelichting	Groep
1.	Inputfile	Naam van de ingevoerde tekstfile	0
2.	Segment	Tekstsegment waarvoor de featurewaarde geldt	0
3.	Woord	Het woord waarom het gaat	0
4.	Lemma	Het lemma daarvan	0
5.	Voll_lemma	Het volledige lemma, inclusief woorddelen die elders staan (bv. 'uit' bij 'uitnodigen')	0
6.	Morfemen	De kleinste betekenisdragende eenheden van het woord, bijvoorbeeld [be][volk][ing][s][onderzoek]	0
7.	Samenst_delen_Frog	We experimenteren met een samenstellingensplitser op basis van Frog. Deze geeft telkens als hij een samenstelling herkent, de woordsoorten van het eerste en het laatste onderdeel. De gebruikte afkortingen zijn: A = bijvoeglijk naamwoord (inclusief bijwoordelijk gebruikte adjectieven); B = bijwoord; LID = lidwoord; N = zelfstandig naamwoord (noun); T = telwoord; P = voorzetsel; V = werkwoord Op dit moment is de splitser uitgeschakeld vanwege bugs.	
8.	Wrdsoort	De woordsoort (Part-Of-Speech), als volgt afgekort: ADJ = bijvoeglijk naamwoord (inclusief bijwoordelijk gebruikte adjectieven); BW = bijwoord; LET = interpunctieteken; LID = lidwoord; N = zelfstandig naamwoord (noun); SPEC = speciale eenheden, vooral namen*; TW = telwoord; VG = voegwoord; VNW = voornaamwoord; VZ = voorzetsel; WW = werkwoord	7c
9.	Afk	Afkorting (1=ja; 0=nee)	7d
10.	Let_per_wrd	Letters per woord	1
11.	Wrd_per_let	Woorden per letter	1
12.	Let_per_wrd_zn	Letters per woord, zonder namen	1
13.	Wrd_per_let_zn	Woorden per letter, zonder namen	1
14.	Morf_per_wrd	Morfemen per woord	1
15.	Wrd_per_morf	Woorden per morfeem	1
16.	Morf_per_wrd_zn	Morfemen per woord, zonder namen	1
17.	Wrd_per_morf_zn	Woorden per morfeem, zonder namen	1
18.	Sam_delen_per_wrd	Samenstellingsdelen per woord	1

Nr	Naam	Toelichting	Groep
19.	Sam_d	Samenstellingsdichtheid	1
20.	Samenst	Gaat het om een samenstelling (1=ja, 0=nee)	1
21.	Samenst_delen	Aantal delen van de samenstelling (als het geen samenstelling betreft, zijn deze en de volgende kenmerken 'NA')	1
22.	Let_per_wrd_hfdwrd	Woordlengte in letters voor het hoofdwoord	1
23.	Let_per_wrd_satwrd	Woordlengte in letters voor het satellietwoord	1
24.	Wrd_freq_log_hfdwrd	Woordfrequentie (logaritme) van het hoofdwoord	1
25.	Wrd_freq_log_satwrd	Woordfrequentie (logaritme) van het satellietwoord	1
26.	Wrd_freq_log_(hfd_sat)	Gemiddelde van de logaritmen van de woordfrequentie van hoofdwoorden en satellietwoorden in de nominale samenstellingen	1
27.	Freq1000_hfdwrd	Hoort het hoofdwoord bij de meest frequente 1000 woorden	1
28.	Freq5000_hfdwrd	Hoort het hoofdwoord bij de meest frequente 5000 woorden	1
29.	Freq20000_hfdwrd	Hoort het hoofdwoord bij de meest frequente 20000 woorden	1
30.	Freq1000_satwrd	Hoort het satellietwoord bij de meest frequente 1000 woorden	1
31.	Freq5000_satwrd	Hoort het satellietwoord bij de meest frequente 5000 woorden	1
32.	Freq20000_satwrd	Hoort het satellietwoord bij de meest frequente 20000 woorden	1
33.	Wrd_freq_log	Woordfrequentie, logaritme	1
34.	Wrd_freq_log_corr	Woordfrequentie, waarbij voor samen-stellingen de frequentie van het hoofd-woord wordt genomen	1
35.	Wrd_freq_zn_log	Woordfrequentie zonder namen, logaritme	1
36.	Wrd_freq_zn_log_corr	Woordfrequentie zonder namen, waarbij voor samen-stellingen de frequentie van het hoofd-woord wordt genomen	1
37.	Lem_freq_log	Lemmafrequentie, logaritme	1
38.	Lem_freq_zn_log	Lemmafrequentie zonder namen, logaritme	
39.	Freq1000	Hoort het woord bij de meest frequente 1000 woorden	1
40.	Freq2000	Idem voor de meest frequente 2000 woorden	1
41.	Freq3000	Idem voor de meest frequente 3000 woorden	1
42.	Freq5000	Idem voor de meest frequente 5000 woorden	1
43.	Freq10000	Idem voor de meest frequente 10000 woorden	1
44.	Freq20000	Idem voor de meest frequente 20000 woorden	1
45.	Conn_type	Type verbindingswoord (als het om een verbindingswoord gaat): Causaal, Comparatief, Contrastief, Opsommend (wg= woordgroep; zin=zin) dan wel Temporeel	4
46.	Conn_wrdcombi	Maakt het woord deel uit van een woordcombinatie die als 1 verbindingswoord wordt geteld, zoals 'dan' in 'dan ook'? (1=ja; 0=nee)	4
47.	Vnw_ref	Terugverwijzend voornaamwoord (1=ja; 0=nee)	3
48.	Semtype_nw	Het semantische type van een zelfstandig naamwoord	5
49.	Alg_nw	Het semantische type van het nomen als het om een algemeen naamwoord gaat (0 = niet van toepassing)	5
50.	Conc_nw_strikt	Is het nomen concreet in strikte zin? (1=ja; 0=nee)	5
51.	Conc_nw_ruim	Is het nomen concreet in ruime zin? (1=ja; 0=nee)	5
52.	Semtype_bvnw	Het semantische type van een bijvoeglijk naamwoord	5
53.	Conc_bvnw_strikt	Is het adjectief concreet in strikte zin? (1=ja; 0=nee)	5
54.	Conc_bvnw_ruim	Is het adjectief concreet in ruime zin? (1=ja; 0=nee)	5
55.	Semtype_ww	Het semantische type van een werkwoord; daarbij wordt zowel de concreetheid als het type 'state of affairs' gegeven	5
56.	Alg_ww	Het semantische type van het werkwoord als het om een algemeen werkwoord gaat (0 = niet van toepassing)	5
57.	Semtype_bw	Het semantische type van een bijwoord (algemeen/specifiek/niet gevonden)	
58.	Pers_ref	Verwijzing naar personen	6
59.	Pers_vnw1	Eerste-persoonsvoornaamwoord	6
60.	Pers_vnw2	Tweede-persoonsvoornaamwoord	6
61.	Pers_vnw3	Derde-persoonsvoornaamwoord	6
62.	Pers_vnw	Persoonlijk of bezittelijk voornaamwoord	6
63.	Naam_POS	Is het woord een naam? (1=ja; 0=nee) Het gaat hier om de Frog-woordsoort (POS-tag) *spec, eigen*.*	7a
64.	Naam_NER	Welk soort naam is het woord volgens de Named Entity Recognition (NER) module? LOC = plaatsnaam	7a

Nr	Naam	Toelichting	Groep
		ORG = organisatienaam PER = persoonsnaam PRO = productnaam MISC = andersoortige naam	
65.	Imp_ellips	Gaat het om een werkwoord in een zin zonder subject?	7b
66.	Ww_vorm	Om welk soort werkwoord gaat het? HOOFDWW = hoofdwerkwoord KOPPELWW = koppelwerkwoord MODAALWW = modaal werkwoord PASSIEFWW = hulpwerkwoord van de lijdende vorm TIJDWW = hulpwerkwoord van tijd	7b
67.	Ww_tt	Gaat het om een vorm in de tegenwoordige tijd? (1=ja; 0=nee)	7b
68.	Vol_dw	Gaat het om een voltooid deelwoord? (1=ja; 0=nee)	7b
69.	Onvol_dw	Gaat het om een onvoltooid deelwoord? (1=ja; 0=nee)	7b
70.	Infin	Gaat het om een infinitief? (1=ja; 0=nee)	7b
71.	Archaisch	Gaat het om een archaïsche naamvalsvorm? (1=ja; 0=nee)	1
72.	Log_prob	De waarschijnlijkheid van dit woord gegeven de 2 woorden die eraan voorafgaan (hiervan: de logaritme)	8
73.	Intens	Is een woord een intensiveerder of maakt het deel uit van een intensiverende combinatie (1 = ja; 0 = nee)	9

* Het kan behalve namen (80-90%) ook gaan om buitenlandse woorden, cijfers, tijden, romeinse cijfers, URL's, tekens en om letterreeksen die abusievelijk niet herkend zijn als woord, zoals *da's* en *enz*.

Literatuur

Berck, P., and Van den Bosch, A. (2009). Memory-based machine translation and language modeling. *The Prague Bulletin of Mathematical Linguistics* 91, pp. 17-26.

Bouma, G., Van Noord, G., & Malouf, R. (2001). Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers* 37(1), 45-59.

Daelemans, W., Van den Bosch, A. 2005. *Memory-Based Language processing*. Cambridge University Press.

Camblin, C., Ledoux, K. Boudewyn, M., Gordon, P.C. & Swaab, T.Y. (2007). Processing new and repeated names: Effects of coreference on repetition priming with speech and fast RSVP. *Brain Research*, 1146, p. 172-184.

Covington, M.A., He, C., Brown, C., Naçi, L. & Brown, J. (2006). *How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level scale*. CASPR Research Report 2006-01, Artificial Intelligence Center, The University of Georghia.

Daelemans W., Van den Bosch A., Weijters A. (1997). IGTree: using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review*, 11, 407-423.

Desmet, B. & Hoste, V. (2013). Fine-Grained Dutch Named Entity Recognition. *Language Resources and Evaluation* 48(2), 307-343.

Eynde, F. van (2004). *Part of Speech tagging en lemmatisering van het Corpus Gesproken Nederlands*. Centrum voor Computerlinguïstiek, KU Leuven.

Flowerdew, J. & Forest, R.W. (2015). *Signalling nouns in English. A corpus-based approach.* Cambridge University Press, Cambridge.

Gibson, E. (2000). The Dependency Locality Theory: a distance based theory of linguistic complexity. In In Y. Miyashita, A. P. Marantz & W. O'Neil (eds.), *Image, language, brain* Cambridge: MIT Press, 95–126.

Goldsmith, J. (2007). Probability for linguists. *Mathématiques et sciences humaines. Mathematics and social sciences*, (180), 73-98.

Graesser, A.C., McNamara, D., Louwerse, M.M. and Cai, Z.. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers*, (36), pp. 193-202.

Haas, W. de & Trommelen, M. 1993. *Morfologisch handboek van het Nederlands*. SDU Uitgeverij, Den Haag.

Hendrickx, I, and Van den Bosch, A. (2003). Memory-based one-step named-entity recognition: Effects of seed list features, classifier stacking, and unannotated data. In Proceedings of CoNLL-2003, the Seventh Conference on Natural Language Learning, Edmonton, Canada, 2003, pp. 176-179.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers* 53, 61-79. Lund University, Department of Linguistics and Phonetics.

Keuleers, E., Brysbaert, M. & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods, 42*(3), 643-650.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology, 68(8),* 1665-1692.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction* 1(1), 60-69.

Kraf, R. & Pander Maat, H. (2009). Leesbaarheidsonderzoek: oude problemen en nieuwe kansen. *Tijdschrift voor Taalbeheersing* 31(2), 97-123.

Manning, C.D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge Mass. / London.

Martin, W. & Maks, I. (2005). Referentie Bestand Nederlands. Met medewerking van S. Bopp en M. Groot.

McCarthy, P.M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42(2), 381-392.
Oostdijk, N., Reynaert, M. Hoste, V. & Heuvel, H. van den (2013). *SoNaR User Documentation*. Version 1.0.4.
Pander Maat, H., Kraf, R., Bosch, A. van den, Dekker, N., Gompel, M. van, Kleijn, S., Sanders, T.J.M. & Sloot, K. van der (2014). T-Scan: a new tool for analyzing Dutch text. *Computational Linguistics in the Netherlands Journal 4,* 53-74.
Pander Maat, H. (2002). *Tekstanalyse. Wat teksten tot teksten maakt.* Coutinho, Bussum.
Schmitt, N., Jiang, X. & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal,* 95, 26-43.
Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument.* Cito.
Temperley, D. (2007). Minimization of dependency length in written English. *Cognition* 105 (2), 300-333.
Van den Bosch, A., and Daelemans, W. (1999). Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 285-292.
Van den Bosch, A., Busser, G.J., Canisius, S., and Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In P. Dirix, I. Schuurman, V. Vandeghinste, and F. Van Eynde (Eds.), *Computational Linguistics in the Netherlands 2006: Selected Papers of the Seventeenth CLIN Meeting*. Utrecht: LOT, 191-206.
Weiskopf, D. (2007). Compound nominals, context and compositionality. *Synthese* 156, 161-204.
Zwaan, R.A., & Rapp, D.N. (2006). Discourse comprehension. In: M.A. Gernsbacher & M.J. Traxler (Eds.). *Handbook of psycholinguistics*, hoofdstuk 18 (pp. 725-764). San Diego, CA: Elsevier.

Bijlagen

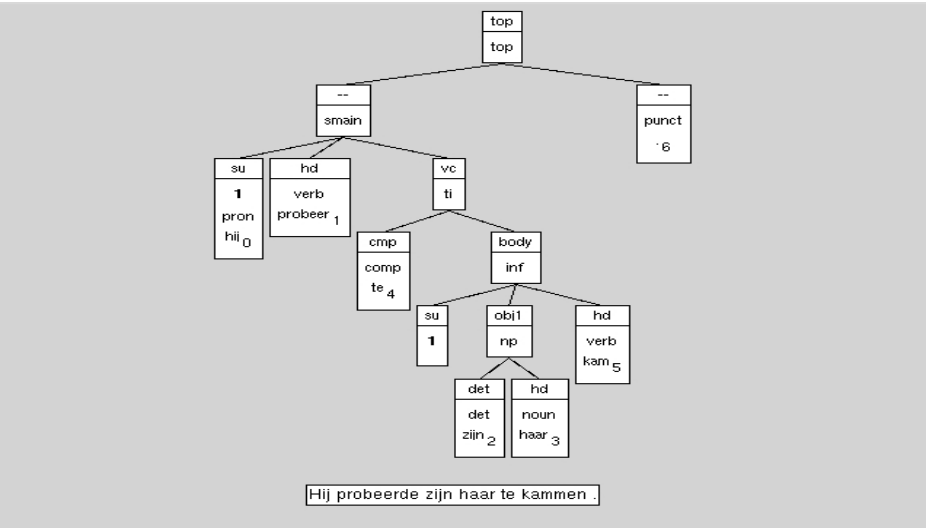
Bijlage A. De implementatie van D-level in T-Scan

D-level is een maat voor syntactische complexiteit ontworpen door Rosenberg en Abbeduto (1987) die gebaseerd is op taalverwervingsonderzoek. Voor T-Scan hielden we ons aan de implementatie door Covington et al (2006). We hebben een D_level schaal geïmplementeerd waarbij vanaf het hoogste niveau (7) gekeken wordt of de zin op dit niveau past, indien niet wordt steeds een niveau gedaald tot level 0.

Hieronder volgt een korte omschrijving van de schalen. Merk op dat T-Scan bij het hoogste niveau begint toe te wijzen; wanneer een zin voldoet aan de kenmerken voor niveau 6, wordt de test voor niveau 5 (of een lager niveau) niet meer gedaan.

Level 0
– Simpele zinnen (hoofdzin zonder bijzin), inclusief vraagzinnen.
– Elliptische zinnen, bijv. *"Daar ja."*
Implementatie: T-Scan telt het aantal finiete werkwoorden (persoonsvormen). Als het aantal persoonsvormen kleiner dan of gelijk is aan 1, wordt level 0 toegekend.

Level 1
– Zinnen met een infinitief die het subject deelt met de persoonsvorm, bijv. "Hij probeerde zijn haar te kammen."
Implementatie: T-Scan doorzoekt Alpino bomen op zoek naar een infinitief (een verbal-complement van een head-werkwoordsknoop met een *ti*- of *oti*-label). Vervolgens wordt in de dochterknopen recursief naar een subject gezocht, en moet het subject een index delen met het subject van de persoonsvorm (de head-verb van de main-clause).

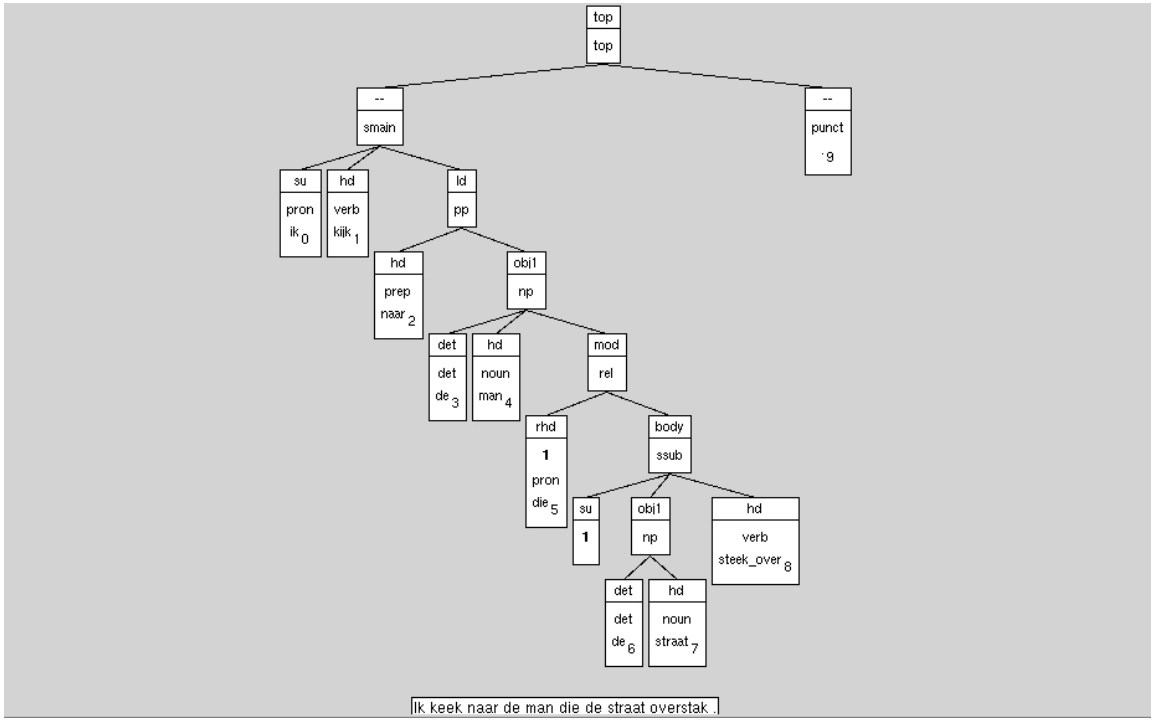


Voorbeeld van een zin op D_level 1.

Level 2
– Een zin opgebouwd uit meerdere nevengeschikte zinnen, bijv. *"Ik ging naar huis en Piet liep weg."*
– Zinnen met NPs die een nevenschikking bevatten, bijv. *"Jan en Marie liepen naar huis."*
– Andere zinnen met een nevenschikking, bijv. *''Hij sprong en schreeuwde het uit van vreugde."*
Implementatie: T-Scan controleert zinnen op de aanwezigheid van nevenschikkende voegwoorden, die door Frog aangeduid worden.

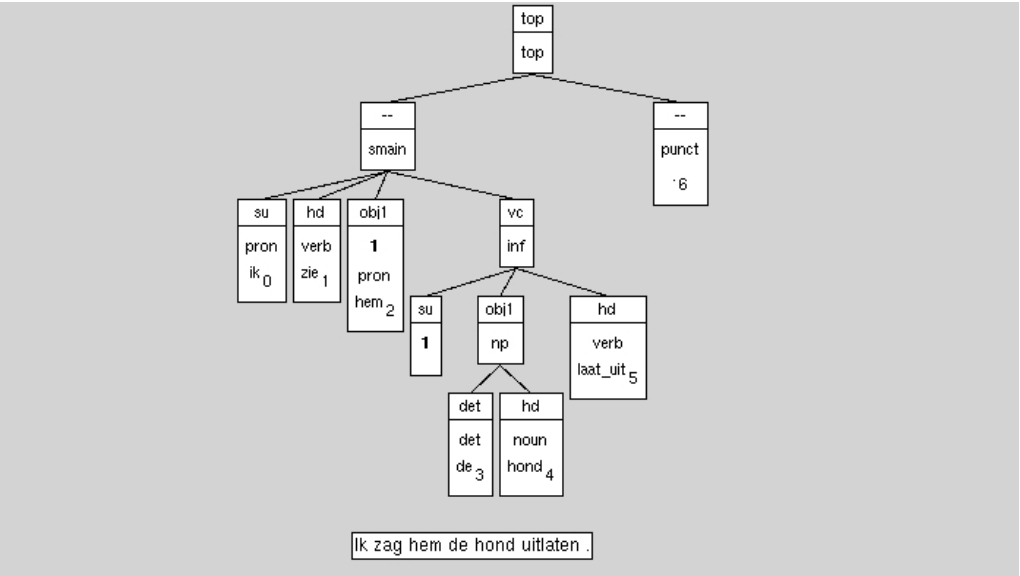
Level 3

– Zinnen met een betrekkelijke bijzin die het object modificeert, bijv. *"Ik keek naar de man die de straat overstak."*
– Zinnen met een bijzin die als object van de hoofdzin fungeert, bijv. *"Ik wist dat hij boos was."*
Implementatie: T-Scan doorzoekt de Alpino boom naar betrekkelijke bijzinnen (cat=*rel*), die onder een objectknoop vallen (zie het voorbeeld), of naar een complementizer phrase die een *vc*(verbal complement) is van een werkwoordshoofd.



Voorbeeld van een zin op D_level 3

Level 4
– Zinnen met een infinitiefcomplement waarvan het subject overeen komt met het object van de persoonsvorm, bijv. *"Ik zag hem de hond uitlaten"*
– Zinnen met een comparatief die een object van vergelijking bevat, bijv. *"Hij is ouder dan Karel."*
Implementatie: T-Scan doorloopt Alpino bomen op zoek naar *vc*-knopen. Als die gevonden zijn wordt gezocht of de *vc*-knoop een subject dochter heeft die een index deelt met een object van de persoonsvorm. Daarnaast worden comparatieven met object van vergelijking gevonden door in de Alpino boom te zoeken naar *obcomp*-relaties.



Voorbeeld van een zin op D_level 4

Level 5

– Zinnen die een onderschikkend voegwoord bevatten.

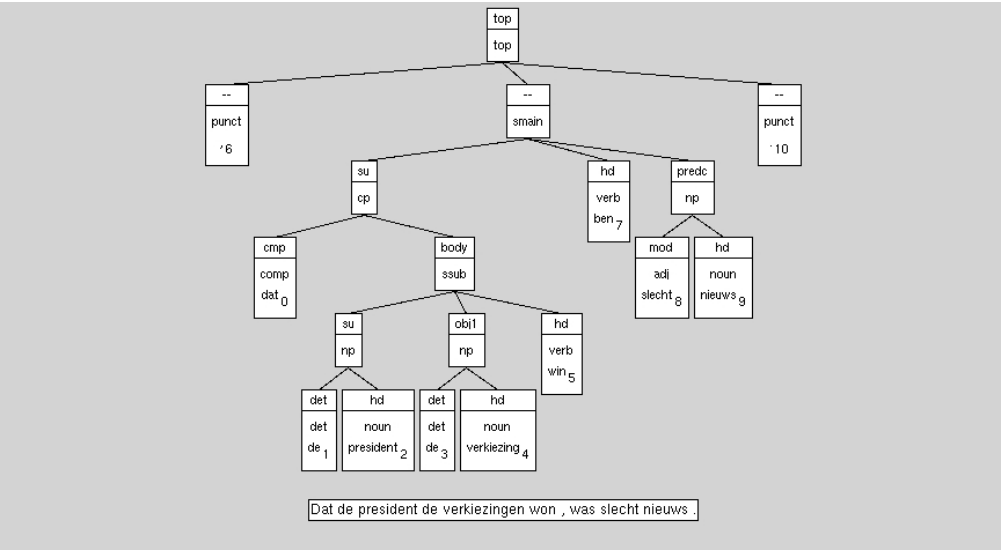
Implementatie: T-Scan controleert zinnen op de aanwezigheid van onderschikkende voegwoorden, die door Frog aangeduid worden.

Level 6

– Zinnen met een betrekkelijke bijzin die het subject modificeert, bijv. *"De man die de straat overstak knoopte zijn jas dicht."*

– Zinnen met een bijzin die als subject van de hoofdzin fungeert, bijv. *"Dat de president de verkiezingen won, was slecht nieuws."*

Implementatie: T-Scan doorzoekt de Alpino boom naar betrekkelijke bijzinnen (cat=*rel*), die onder een subjectknoop vallen), of naar complementizerphrase die een *su*(subject) is van een werkwoordshoofd.



Voorbeeld van een zin op D_level 6.

Level 7

– Zinnen die dubbele nestingen bevatten (bijzinnen in bijzinnen), bijv. “Karel dacht dat hij Marie, die haar haren rood geverfd had, op straat had Zien lopen.”

Implementatie: T-Scan telt of meer dan één werkwoordshoofd onder een *smain*-knoop vallen.

Bijlage B. Nominalisatiesuffixen die T-Scan gebruikt

Niet alle nominaliseringen maken de tekst abstracter. We hebben geprobeerd om alleen die suffixen te selecteren die inderdaad tot abstracte woorden lijken te leiden. Deze abstracte woorden zouden de tekst moeilijker kunnen maken, en zijn dus interessant voor T-Scan om te kunnen detecteren. Hieronder volgende twee lijsten met suffixen die wel en niet tot abstracte woorden leiden. De totale lijst aan suffixen komt uit De Haas en Trommelen (1993).

Suffixen die T-Scan gebruikt:

- -ing
- -sel
- -(e)nis
- -heid
- -te (incl. ge-….-te)
- -schap
- -dom
- -sie uitgezonderd namen en landen
- -tie uitgezonderd namen en landen
- -de
- -iek
- -iteit
- -isme (doen we op stringniveau)
- -age
- -ose (*prognose*); als string herkend
- -ase (*extase*); string
- -ese (*hypothese*) ; string
- -esse

Niet door T-Scan gedetecteerd:

- -erij: levert geen abstracte nouns op, eerder concrete nouns
- -st: makkelijk te verwarren met overtreffende trap
- -t (teelt): te verwarren met 3e persons-uitgang
- -ief: *statief*, *explosief* e.d.: doet Frog niet, is ook niet altijd abstract
- -ade: niet altijd abstract (*limonade*, *balustrade*)
- -uur; Frog raakt in de war; bovendien zijn er concrete voorbeelden: *frituur*, *literatuur* is concreter dan *literair*, e.d.
- -ure: *blessure*, *gravure*; vaak niet abstract
- -aire: *documentaire*; vaak niet abstract
- -oir: *urinoir*; vaak niet abstract
- -erie; *parfumerie*, vaak niet abstract
- -et; vaak niet abstract
- -ma, -eem, -con, -um, -ex, -ide, -ium, -ale, -alis, -ix, -oide, -itis, chemische suf- fixen, -edie, -on, -isie, -ooi, -akel, -ate, -iel, -ance, -ence, -ande, -enda, -ens, -oop, -gram, -droom, -staat, -ude, -rama, -theek, -tiek, -taria, -tel, -iet, -ijn, -aat, -eur, -or, -ier, iere, -ee, -ette, -ine, -ant, -ent, -us, -ica, -aal, -eel, -ans, -ioen, : vaak niet abstract en/of nominaliseringskarakter onduidelijk.

Verder moeten we niet vergeten dat ‘stam’-nominaliseringen (*verraad, betoog*) niet door T-Scan worden gevonden, gezien het ontbreken van een suffix.

Bijlage C. Connectievenlijsten in T-Scan

BW/VG/VZ = dit woord wordt alleen geteld als bijwoord / voegwoord / voorzetsel

Causaal

alleen dan
aangezien
anders
bijgevolg
blijkens
daar
daardoor
daarmee
daarom
daartoe
daarvoor
dan ook
dankzij
derhalve
dientengevolge
doordat
dus
ergo
ermee
erom
ertoe
getuige VZ
gezien VZ
hierdoor
hiermee
hierom
hiertoe
hiervoor
immers
in verband met
indien
ingeval
ingevolge
krachtens
met behulp van
middels
mits
namelijk
om VG
omdat
opdat
teneinde
vanwege
vermits
waardoor
waarmee
waarom
waartoe
wanneer VG
want
wegens
zodat
zodoende
zolang
dan ook
tengevolge van
vandaar dat
zo VG
zo ja
zo nee
zo niet
zodoende

Comparatief

alsof
dan VG
meer
meest
minder
minst
naargelang
naarmate
zoals
meer dan
minder dan
net zo min

Contrastief

al VG
alhoewel
althans
anderzijds
behalve
behoudens
daarentegen
daarvan
desondanks
doch
echter
enerzijds
evengoed
evenwel
hoewel
hoezeer
integendeel
maar VG
niettegenstaande
niettemin
nochtans
ofschoon
ondanks
ongeacht
tenzij
terwijl
uitgezonderd
weliswaar
in plaats
laat staan
ook al
in plaats daarvan
in tegenstelling tot
zonder dat

Opsommend (kolom 1: woordgroepniveau; kolom 2-4: zinsniveau)

alsmede
alsook
annex
en
evenals
noch
of
ofwel
respectievelijk
zowel
bovenal
bovendien
buitendien
daarenboven
daarnaast
eveneens
evenmin
hetzij
hierenboven
hoofdzakelijk
met name
nog eens
om te beginnen
ook
ook nog
ook nog eens
op de eerste plaats
op de tweede plaats
op de derde plaats
op de vierde plaats
temeer
tevens
vooral
voornamelijk
voorts
waarnaast
zelfs
zowel
daarbij komt
dan wel
ten eerste
ten tweede
ten derde
ten vierde
ten overvloede

Temporeel

aanstonds
achtereenvolgens
aldoor
aleer
altijd
alvast
alvorens
bijtijds
binnenkort
daarna
daarnet
daarstraks
daartussendoor
dadelijk
destijds
eensklaps
eer VG
eerdaags
eerdat
eerlang
eerst BW
eertijds
eindelijk
ertussendoor
hoelang
indertijd
ineens
ingaande
inmiddels
meestal
meteen
nadat
naderhand
nadezen
nadien
net BW
olim
onderwijl
onlangs
opeens
pardoes
pas BW
plots
plotsklaps
recentelijk
reeds
sedertdien
sinds
sindsdien
steeds
strakjes
straks
subiet
tegelijk
tegelijkertijd
terstond
tevoren
tezelfdertijd
thans
toen
toenmaals
toentertijd
totdat
vervolgens
vooraf
vooraleer
vooralsnog
voordat
voorheen
wederom
weer BW
weldra
weleer
zodra
zo-even
zo even
zo gauw
zogauw
zojuist
zolang
zonet
zopas
a la minute
hic et nunc

Bijlage D. Semantische klassen voor zelfstandige naamwoorden

1 Inleiding

Voor het vaststellen van woordconcreetheid zijn we uitgegaan van een zeer waardevolle databron: de semantische typen uit het Referentiebestand Nederlands (RBN) (zie http://tst-centrale.org/images/stories/producten/documentatie/rbn_documentatie_nl.pdf).

Bij nadere beschouwing bleek het RBN wel een behoorlijk aantal fouten en inconsequentheden te bevatten. Wat betreft de fouten, een voorbeeld daarvan is dat vaak reeksen woorden met een bepaald begin in dezelfde categorie terechtkomen. Alle woorden die beginnen met *straat-* zijn bijvoorbeeld in de categorie Place gezet, inclusief *straathandelaar, straatklinker* en *straatprostitutie*. Een voorbeeld van een inconsequentheid is dat de ene samenstelling met het grondwoord *-commissie* als Human is gezien en de andere niet. Een ander probleem is dat veel woorden een groot aantal lezingen hebben. Een keuze uit die lezingen is vrij willekeurig. Om deze redenen is de lijst handmatig gecorrigeerd; bij een hoge mate van ambiguïteit of polysemie is het woord ongedefinieerd gelaten.

Verder bepaalt de klasse niet geheel hoe concreet een woord is. Daarom zijn de dynamische nomina en de substantiewoorden onderverdeeld in abstracte en concrete gevallen. Verder zijn bepaalde artefacten, substanties en planten en dieren geplaatst in een nieuwe categorie: voeding en verzorging.

Ten slotte zijn aan de RBN-woorden ongeveer 9000 nieuwe woorden toegevoegd op basis van geanalyseerde teksten, de naamwoorden uit een frequentielijst van het SoNaR-corpus, en eigen initiatief. De laatste versie van de lijst telt bijna 46000 zelfstandige naamwoorden.

In het schema hieronder bevat de eerste kolom de categorieën, gevolgd door de veelal Engelse termen waarmee worden aangeduid in de lijsten waar T-Scan mee werkt. In de derde kolom staan de termen waarmee de betreffende klasse nomina wordt aangeduid in namen van T-Scankenmerken. In die kenmerknamen staat achter die termen een '_d' bij dichtheden en een '_p' bij proporties.

	Categorie	**Voorbeeld**	**T-Scankenmerk**
1	Personen (human)	*Leraar, schreeuwlelijk*	Pers_nw
2	Planten en dieren (nonhuman)	*Mus, eik*	PlantDier_nw
3	Gebruiksvoorwerp (artefact)	*Stoel, weefgetouw*	Gebr_vw_nw
4	Concrete substanties (substance_conc)	*Olie, cellofaan*	Subst_conc_nw
5	Voeding en verzorging (voed_verz)	*Melk, sigaret, bruistablet*	Voed_verz_nw
6	Concreet overig (concrother)	*Galblaas, vulkaan*	Concr_ov_nw
7	Concreet gebeuren (dynamic_conc)	*Aai, ademhaling*	Gebeuren_conc_nw
8	Plaats (place)	*Amsterdam, voorkamer*	Plaats_nw
9	Tijd (time)	*Feestdag, periode*	Tijd_nw
10	Maat (measure)	*Euro, dB*	Maat_nw
11	Abstracte substanties (substance_abstr)	*Fosfor, splijtstof*	Subst_abstr_nw
12	Abstract gebeuren (dynamic abstr)	*Crisis, loonverlaging*	Gebeuren_abstr_nw
13	Organisatie (institut)	*Werkgeversorganisatie*	Organisatie_nw
14	Abstract overig (nondynamic)	*Christendom, motto*	Ov_abstr_nw
15	Undefined	*Schot, kant*	Undefined_nw

Er zijn twee overkoepelende klassen gevormd.

- Als concreet in strikte zin (conc_nw_strikt) zijn opgevat de klassen 1 tot en met 7;
- Als concreet in ruimte zin (conc_nw_ruim) zijn opgevat de klassen 1 tot en met 10; dat wil zeggen dat plaatsen, tijden en maten wel concreet-ruim zijn maar niet concreet-strikt.

De codering vindt plaats op basis van lemma's, zodat meervouden en verkleinwoorden worden gecodeerd op basis van het basiswoord, als dat in de lijst staat. Als de lemmatisering foutief is, kan er dus een treffer in de lijst gemist worden. Een voorbeeld: 'werkwijze' wordt foutief gelemmatiseerd als 'werkwijz'. Daardoor rapporteert T-Scan helaas abusievelijk dat dit woord niet gevonden wordt.

2 Hoe de codering is uitgevoerd

Ambiguïteit en polysemie bij de herziening van de lijst

Er zijn allerlei woorden met een groot aantal betekenissen. In een eerdere versie van T-Scan werden woorden die minimaal een concrete lezing hebben, als concreet gezien. Dat levert een overtelling op. In de nieuwe lijst zijn ambigue of polyseme woorden onder handen genomen als ze zowel concrete als niet-concrete lezingen hebben. Dus een woord als *amfibie,* dat zowel op een dier als op een artefact (voertuig) kan slaan, is ongemoeid gelaten.

De 'zwaar' ambigue of polyseme woorden daarentegen zijn ofwel leeg gemaakt (voorzien van het label 15, Undefined), ofwel er is een dominante lezing gekozen. Meestal gaat het om leeg maken: in dat geval blijft het woord in de lijst staan, maar zonder type erbij. Voorbeelden van leeg gemaakte woorden:

- *Baken, basispakket, goed, geval, hoop, straal, tip, spel, golf, stroom, weer, rand, vlek, schreef, provisie, scheut, vertering, commando, organisatie, instelling, prikkel, gelegenheid, type, sopraan, harmonie, raad, staat, gezag* enz.

Voorbeelden van ambigue woorden waarin een lezing is verwijderd:

- Voor *beroep* staan de volgende lezingen genoteerd: nondynamic, dynamic, dynamic, artefact. Omdat ik de artefact-lezing niet begreep, is deze verwijderd. Daardoor is het woord niet-concreet geworden. De dubbelzinnigheid tussen nondynamic (beroep als bezigheid) en dynamic (beroep als juridisch protest) is behouden.
- Voor *ambtenarij* staat zowel een nondynamic- als een human-lezing. Daarvoor in de plaats is het label instituut gezet.
- Voor *boom* staat ook een 'human'-lezing (*een boom van een vent*). Die is verwijderd.
- Voor *criterium* staat is de nondynamic-lezing (maatstaf) gehandhaafd, en de dynamische lezing (wielerwedstrijd) verwijderd.
- Voor *Chileen, Albanees* e.d. staat ook telkens zowel een nondynamic- als een human-lezing. Daarvan is gekozen voor de human-lezing.
- Voor *fysica, fenomeen, geleide* en *gevolg* staat ook telkens zowel een nondynamic- als een human-lezing. Bij die woorden is gekozen voor de nondynamic-lezing.
- Voor *effect* bestaat een nondynamic- en een artefact-lezing (beleggingsvorm). Omdat het een frequent woord betreft waarvan een van de lezingen veel minder frequent is, is die minder frequente lezing geschrapt. Een soortgelijke keuze is gemaakt voor woorden als *herinnering, status, rede* en *concessie*.

Een speciaal probleem vormen de woorden met saillante figuurlijke lezingen, zoals *melkkoe* en *spagaat*. Dat soort woorden zijn als 'undefined' gecodeerd. Van woorden als *speldenprik, spruitjeslucht* en *spitwerk* is de letterlijke lezing zo op de achtergrond geraakt, dat ze als abstract ('nondynamic') zijn gelabeld.

Ten slotte zijn er woorden die een andere betekenis hebben wanneer de beginletter een hoofdletter is. *Bermuda* is een eiland, een *bermuda* is een kledingstuk. T-Scan maakt dit onderscheid zolang het woord midden in de zin staat. Wanneer het woord aan begin van de zin staat, wordt de lezing gekozen van de spelling met een grote letter.

We lichten nu de afzonderlijke klassen verder toe.

1 Personen (human)

Heel veel 'human'-termen zijn familietermen, beroepsaanduidingen (*psycholoog*), functionele rollen (*discussieleider*), opvattingen (*dogmatist*), probleemgroepen (*drankzuchtige*), herkomstaanduidingen (*Gouwenaar*), hobby's (*hondenliefhebber*). Soms ook gaat het om kwalificaties die vooral op mensen worden toegepast (*lelijkerd, doordrammer, honnieponnie*). Soms ook gaat het om bekende personages (*Homerus, Horatius*).

Aanduidingen van kleine groepen (*docententeam, meidenband*) zijn ook als 'human' gecodeerd. Verenigingen daarentegen zijn als organisatie gezien, evenals woorden die naar sportverenigingen verwijzend (*tweedeklasser, middenmoter*).

Woorden als *aandeelhouder* kunnen zowel naar personen als organisaties verwijzen. Dat leidt tot lastige keuzes. In de lijst is *aandeelhouder* als persoonlijk gezien, maar *hoofdaandeelhouder* als organisatie, omdat het veelal niet individuen zijn die het merendeel van de aandelen bezitten. Evenzo zijn *automatiseerder* en *netbeheerder* als aanduidingen van organisaties gezien. Datzelfde geldt voor *fabrikant* dat meestal als stamwoord in samenstellingen voorkomt.

Vaktermen voor mensenrassen zijn niet als 'human' gezien (*hominidae*). Dat geldt ook voor collectiviteiten als: *minderheid, meerderheid, pressiegroep, publieksgroep, rennersveld.* Het gaat hier om ongeorganiseerde groepen; deze zijn leeg gelaten, omdat niet duidelijk is of ze concreet gebruikt worden of niet. Samenstellingen met *–personeel* lijken vaker concreet gebruikt te worden ('het winkelpersoneel is ontevreden'), dus zijn deze wel als 'human' gerekend.

Andere grensgevallen komen hieronder nog aan bod bij 'organisatie'.

2 Planten en dieren (non-human)

Het gaat hier om waarneembare niet-menselijke organismen; in de praktijk betreft het vooral dieren en planten (eetbare planten zijn onder voeding gevat, zie hieronder). Organismen die alleen zichtbaar zijn onder de microscoop (*amoebe,* virus) zijn hier niet opgenomen; zij zijn geplaatst bij de abstracte substanties.

Een aantal woorden verwijzend naar dieren wordt vaak gebruikt om te verwijzen naar mensen: *baardaap, beest, huismus.* Omdat 'human' een belangrijk kenmerk is om persoonlijkheid van de tekst vast te stellen, is per woord geprobeerd een lezing te kiezen. Voor *baardaap* en *huismus* is dat 'human', geworden, voor *beest* non-human.

Ook woorden die refereren aan religieuze of fictieve wezens (*aartsengel, aardgeest*) zijn in de categorie non-human geplaatst.

3 Gebruiksvoorwerpen (artefact)

Als artefact zijn gedefinieerd tastbare en duurzame entiteiten die geproduceerd of gewonnen zijn voor menselijk gebruik. Voor voedings- en verzorgingsproducten is een aparte klasse gevormd, zie hieronder.

Buiten artefacten vallen verder:

- Technische voorzieningen die op zich genomen onzichtbaar zijn (*internetverbinding*)
- Geografische locaties
- Muziekstukken
- Woorden verwijzend naar papieren of digitale teksten zijn meestal als abstract en non-dynamisch gezien, dus als 'informatiedragers' (*beleidsnota, website*); daarentegen vallen identiteitsdocumenten die getoond of overhandigd moeten worden (*paspoort, ticket*), weer wel onder de artefacten. Hetzelfde geldt voor andere woorden die verwijzen naar teksten en documenten die een bepaald visueel beeld oproepen, zoals *factuur, formulier* en *flyer*.

Sommige woorden kunnen algemeen verwijzen naar artefacten: *spul(len), rommel, rotzooi* en *troep*. We hebben ervoor gekozen om *spullen* als artefact te zien, en *rommel, rotzooi* en *troep* niet. *Troep* is meerduidig. *Rommel* en *rotzooi* kunnen ook overdrachtelijk gebruikt worden, als negatieve kwalificatie van niet-concrete zaken.

Artefacten zijn meestal vrij kleine objecten, maar er is een uitzondering. Vervoermiddelen zijn ook als artefact gezien, ook hele grote vervoermiddelen zoals *vliegdekschip*. Ook bouwwerken als *brug* zijn als artefact beschouwd. Dit in tegenstelling tot gebouwen, die zoals hieronder zal blijken als plaats zijn gecodeerd.

4/11 Concrete en abstracte substanties

Artefacten hebben een vaste vorm en kunnen per stuk worden waargenomen, substanties zijn vormloos of bestaan uit verzamelingen van kleine eenheden (*rijst*). Dus materialen, vloeistoffen en poeders zijn substanties. Eetbare substanties vallen onder voeding en verzorging, zie hieronder.

Alle substanties nemen ruimte in, maar niet alle substanties zijn zintuiglijk waarneembaar. 'Chemische' en 'farmaceutische' substanties (*fosfor, virusremmer*) zijn niet waarneembaar in de zin van zintuiglijk herkenbaar als zodanig. *Hout* is dat bijvoorbeeld wel. Daarom is *fosfor* een abstracte substantie en *hout* een concrete.

Vaak zijn termen uit de chemische en farmaceutische sfeer abstracte substanties. Maar niet alle termen verwijzend naar geneesmiddelen zijn abstracte substanties, want bijvoorbeeld pillen en tabletten (*Rennies, aspirientjes*) worden onder voeding en verzorging (zie hieronder) gevat.

Als substantie worden alleen de substantieterm zelf gecodeerd. Zo wordt *olie* als concrete substantie gezien, maar *olievoorraad* of *olievervuiling* zijn abstract.

5 Voedings- en verzorgingsmiddelen

Sommige artefacten, substanties en organismen worden gegeten, gedronken of anderszins concreet zelf toegediend op dagelijkse basis: voedsel, drank, genotmiddelen, concreet voorstelbare geneesmiddelen en producten voor persoonlijke verzorging. We hanteren verder de term 'voeding en verzorging'. Deze categorie impliceert dat de artefacten, substanties, planten en dieren in onze classificatie niet-consumeerbaar zijn.

Heldere voorbeelden van woorden uit de groep voedsel en drank zijn *aalbes, aalbessensap, aardappel, aardappelmeel, aardappelpuree, aardbei, aardnoot, abrikoos, achterham, amandel, amandelbroodje, amandelolie, ananas, andijvie, anijs en ansjovis*. Onder de geneesmiddelen vallen allerlei pillen en drankjes. De verzorgingsproducten zijn nogal eens crèmes en zalfjes.

Twijfelgevallen doen zich vooral voor bij dieren. Zo zijn *rund, zuigkalf, rundvlees, kalfsvlees* en *zeebaars* in als voedingsmiddel gezien, maar *koe, kalf, hert* en *snoekbaars* niet. Als koeien besproken worden als eetbaar, worde veelal over *rund* gesproken. Kalveren en herten worden gegeten, maar kunnen ook als dieren aan de orde komen. Zeebaarzen worden vaker als eetbaar besproken dan snoekbaarzen. Samenstellingen met –vee daarentegen (*rundvee, pluimvee*) worden niet als voedsel gezien. Aanduidingen van gewassen en bomen (*voedergewas, appelboom*) evenmin.

Pillen en tabletten worden ook in de oraal ingenomen categorie geplaatst, dit in tegenstelling tot aanduidingen van de werkzame stof in geneesmiddelen. Een uitzondering vormt *anti-conceptiepil*, een term die meer met de werking dan met het in te nemen object wordt geassocieerd. Ook genotmiddelen als rookwaren (*sigaret, shag*) en drugs (*cocaïne*, *xtc*) worden in deze groep geplaatst.

6 Concreet-overig (concrother)

Er zijn concrete woorden die geen artefact, plant, dier of substantie zijn. Het gaat dan om bijvoorbeeld:

- zichtbare lichaamsdelen van mensen en dieren (*neus, schouder*); niet *brein* of *nier* want die delen van het lichaam zijn onzichtbaar;
- zaken die door mensen of dieren worden uitgescheiden (*uitwerpsel, zweet*);
- delen van planten en vruchten (*achillespees, meeldraad, okkernoot, pitje, boon*);
- 'onwillekeurige' fysieke verschijnselen (*aardbol, berghelling*);
- zichtbare medische klachten en uiterlijke kenmerken (*blaren, vlekken, blos* e.d.);
- vormaspecten (*ribbels, splinter, spaander, spleet*); onzichtbare klachten daartentegen (*spierscheuring, kuitblessure*) zijn als abstract gecodeerd;
- geluiden (*bijgeluiden, grafstem*), geuren (*dennengeur*), en woorden refererend naar kleuren (*herfstkleur*) en licht (*kaarslicht*);
- weers- en natuurverschijnselen zoals *motregen* en *zonsopgang*.
- visuele voorstellingen (*gezicht, vergezicht*);
- gezichtsuitdrukkingen (*glimlach, grimas*);
- bekende afbeeldingen (*smiley, emoticon, emoji)*;
- lichaamshoudingen en –bewegingen (*kleermakerszit, pirouette*).

Er zijn nogal wat biologische en natuurkundige verschijnselen die buiten deze groep vallen. Bijvoorbeeld: *atoom, celkern, spierweefsel.*

Meer algemeen is voor de klassen Artefact, Substance en Concrother dat de woorden een zintuiglijke voorstelling oproepen en daartoe moeten ze *specifiek* van betekenis zijn. Er zijn worden die in het algemeen een klasse concrete entiteiten aanduiden. Zo roept *haarverzorgingsproduct* de gedachte op aan *gel* en *shampoo*. Maar omdat deze gedachten nog verschillende beelden kunnen oproepen, is het woord niet met *substance_conc* gecodeerd. Hetzelfde geldt voor woorden als *voortplantingsorgaan*.

7/12 Dynamic-concrete en dynamic-abstract

Daaronder vallen woorden die verwijzen naar een gebeurtenis die in de tijd geplaatst kan worden. *Circusvoorstelling* is een evident dynamisch woord, een *waanvoorstelling* niet, en *voorstellingsvermogen* evenmin. Je kunt je zinnen voorstellen waarin aan de circusvoorstelling een tijdsbepaling wordt gekoppeld, of waarin 'na de circusvoorstelling' zelf een tijdsbepaling is. Dat kan niet met non-dynamische woorden.

Maar dynamische woorden kunnen niet alleen gebeurtenissen zijn, maar ook processen (*transformatie*), inclusief processen die zich langdurig herhalen (*ademhaling, stofwisseling, busvervoer, energiegebruik*). Je kunt die woorden niet in een tijdsbepaling gebruiken, maar je kunt er wel van zeggen dat ze op zeker moment ophouden, belemmerd worden of voltooid zijn. Een ander zinsframe waarin dynamische nomina kunnen worden gebruikt is 'er vindt (een) X plaats'.

Er blijven twijfelgevallen. Van b*eleid* bijvoorbeeld kun je niet zeggen dat het belemmerd wordt, maar wel dat het ophoudt. Daarom zijn woorden op *-beleid* toch dynamisch gecodeerd. Duidelijker dynamisch zijn *beleidsvorming* en *beleidsmaatregel.*

Alleen woorden met een prominente gebeurtenis/proces-lezing krijgen het predicaat 'dynamisch'. Het woord *leestoets* bijvoorbeeld kan dynamisch opgevat worden ('vrijdag moet Jan de leestoets doen') maar ook nondynamisch ('de leestoets is te moeilijk'). Daarom krijgt het woord het predicaat nondynamisch. Evenzo is *voorlichting* nondynamisch gecodeerd (vgl. 'de voorlichting vond dinsdag plaats' versus 'de voorlichting is niet te volgen'), net als *vergunningverlening*. Evenzo is *inrichting* niet als dynamisch gecodeerd, maar *herinrichting* wel.

Er zijn ook woorden met twee of meer helder verschillende lezingen, waaronder dynamische en niet-dynamische. Die lezingen zijn blijven staan in de lijst. Voorbeelden zijn *zending* en *productie*. Deze woorden worden geteld als niet-dynamisch; dat is namelijk de meest algemene lezing. Als we die kiezen, lopen we niet het risico om de dynamiek in een tekst te overschatten.

Onder de dynamische woorden valt te onderscheiden tussen woorden die een zintuiglijke voorstelling oproepen, zoals *aai, ademhaling, hoofdpijn* en *afgraving*, en woorden die dat niet doen, zoals *crisis, intrede* en *transformatie*. De eerste groep krijgt als label 'dynamic-concrete', en de tweede is 'dynamic-abstract'. Er zijn natuurlijk grensgevallen. Zijn bijvoorbeeld een *concert* en *vergadering* concreet? We hebben ervoor gekozen die woorden alleen als zodanig te labelen als ze extra informatie bij wordt gegeven die een visuele, auditieve of anderszins zintuiglijke voorstelling oproept. Daarom is zijn *lunchvergadering* en *galaconcert* concreet, maar *vergadering* en *concert* niet. Evenzo zijn *loopgravenoorlog* en *vuistgevecht* concreet, maar *oorlog en gevecht* niet. En zijn *voetbalwedstrijd* en *tennismatch* concreet, maar *wedstrijd* en *match* niet. En *tennisles* is concreet, maar *proefles* niet. Bij een *feest* is daarentegen altijd sprake van grote groepen mensen die praten en/of dansen onder het genot van drank en voedsel; daarom zijn samenstellingen met *–feest* als concreet gezien.

Aanduidingen van sporten vormen sowieso een dilemma: een term als *voetbal* kan zowel de sector als de activiteit kan aanduiden (vgl. 'in het voetbal gaan enorme bedragen om' met 'ik ben gek op voetbal'). Omdat de meeste sportaanduidingen in staat zijn een visuele voorstelling op te roepen, zijn ze alle als dynamisch-concreet gecodeerd.

8 Plaats (place)

Het gaat hier om woorden met een dominant plaatselijke dan wel ruimtelijke interpretatie:

- aardrijkskundige namen (*Parijs, Afrika*);
- landschappelijke eenheden (*lagune, kust, laagland, toendra, woestijn, zeehaven, rotstuin*);
- gebouwde eenheden (*woning, gebouw, vliegveld, haven, booreiland, autobaan*);
- samenstellingen met als stam *kamer, ruimte, kamp, terrein, kelder, kantoor, gebouw, zone, zaal, kant, post, plaats, hoek* e.d.;
- ruimtelijke vormen (*diagonaal, diameter, rechthoek, vierkant, kromming*).

Woorden die niet-letterlijk lokaal zijn, zoals *luilekkerland, rustpunt, oase* en *mekka* en *toevluchtsoord* krijgen een abstract label (nondynamic). Dat geldt ook voor woorden die vooral als predicaat worden gebruikt: *broeinest, rovershol*.

Wegen, tunnels, paden en straten zijn als plaats beschouwd, en niet als artefacten. De gedachte is: wordt een ruimtelijk situatiemodel opgeroepen? Dat is niet per definitie het geval bij bovengenoemde woorden, maar ze zijn wel altijd goed denkbaar in het kader van een plaatsbepaling in een zin.

Een dilemma hebben we bij woorden die zowel een plaats als een instituut als een artefact kunnen zijn: -*winkels*, -*centra* en –*huizen* bijvoorbeeld.

Daarbij hebben we als volgt gehandeld. Als termen die dominant locaal zijn, beschouwen we *huis, woning, stalling, depot, kamp, verblijf, winkel, shop* en *tehuis*. Er zijn wel uitzonderingen: *webwinkels* zijn geen plaatsen, en *workshops* ook niet.

Als Instituut worden gezien:

- woorden met uitgang: -*wezen*, -*ziekenhuis*, -*afdeling*, -*kolonie*; deze keuze blijft discutabel, want in sommige contexten wordt hier de plaats-lezing bedoeld en in andere de institutionele lezing. Maar omdat die laatste lezingen bij deze woorden vrij regulier lijken (meer dan bij –*winkel*), is ervoor gekozen deze woorden a priori niet als concreet te zien.

Per woord is gekeken naar afleidingen eindigend op X-*erij*. Veelal betekenen die 'plaats waar ge-X-t wordt': *drukkerij, bakkerij* enz. Dat geldt ook voor –*theek* (maar niet *hypotheek*). Maar *visserij* en *uitgeverij* zijn uitzonderingen en zijn als Instituut gecodeerd.

Leeg gemaakt zijn ten slotte de woorden die ambigu zijn tussen Place en Abstract (*splitsing*), tussen Place en Nondynamic (*disco*) of tussen Place en Institut (*centrum*). Wanneer die woorden als grondwoord in een samenstellingen voorkomen, is per geval bekeken welke lezing dominant lijkt. Zo is *buurtcentrum* als plaats gezien, en *afkickcentrum* als organisatie.

9 Tijd (time)

Hieronder vallen tijdseenheden, woorden die betrekking hebben op begin, einde en verloop, kalenderdagen e.d. Aan de RBN-lijst zijn ruim vijftig tijdwoorden toegevoegd, zoals *dinsdagmorgen, dinsdagochtend, dinsdagmiddag, dinsdagavond* en *dinsdagnacht* maar ook woorden als *groeiseizoen, einddatum* enzovoort.

10 Maat (measure)

Hieronder vallen alle maten die niet ruimte of tijd betreffen.

13 Organisatie (institut)

Hieronder vallen organisaties, verenigingen, en zakelijke instellingen. Scholen, bedrijven en kerken zijn gecodeerd als instituten. Dat geldt ook voor samenstellingen met als stam:

- *-wezen, -industrie, -ziekenhuis, -shop, -museum, -commissie, -bestuur, -vereniging, -beweging, -afdeling, -kolonie, gevangenis*.

Sommige woorden duiden zowel groepen mensen aan als organisaties: *comités, commissies, orkesten, koren* e.d. Die woorden krijgen de code 'instituut'. Het onderscheid is subtiel, maar bij wat kleinere collectieven als *team* en *elftal* is gekozen voor 'human'. Vergelijk 'het team is een gezellige groep mensen' met 'de beoordelingscommissie is een gezellige groep mensen'. De tweede zin ligt toch minder voor de hand dan de eerste.

Er zijn ook woorden die als organisatie en als plaats kunnen worden opgevat, zoals –*fabriek* en –*centrale.* Woorden op –*fabriek*, -*school*, -*museum* en –*filiaal* zijn als plaats gezien, bij –*centrale* is per woord bekeken welke lezing het meest prominent lijkt. Zo doet *bewonerscentrale* denken aan een organisatie maar *energiecentrale* aan een plaats met een gebouw.

Daarentegen zijn georganiseerde groepen zoals *bewonersgroep* of *belangengroep* als instituties gezien. Hetzelfde geldt voor maatschappelijke groepen zoals *klootjesvolk, indianenstam* of *middenklassegroep*, en voor alle woorden die eindigen op –*bevolking*. Die groepen (*moslimbevolking, wereldbevolking*) zijn weliswaar niet altijd erg georganiseerd, maar we kunnen ons er meestal geen individuen bij voorstellen.

Termen eindigend op –*land* of –*wereld* waarin een sector als geheel wordt aangeduid, zijn ook als organisatie gecodeerd (*radioland, kunstwereld*). Bij extensie zijn ook termen zoals *visserij* als organisatie gecodeerd. Onder deze klasse vallen dus mensen en activiteiten die gegroepeerd of georganiseerd verlopen.

14 Overige abstracte woorden

Onder deze groep vallen alle woorden die niet in de groepen hierboven vallen. Het gaat dus om abstracte woorden die niet verwijzend naar een gebeuren, substantie of organisatie. Er vallen niet alleen maar bijzonder abstracte woorden onder, maar ook bijvoorbeeld medische en psychische problemen (*kanker, autisme*). Voorbeelden van niet-dynamische woorden beginnend met 'huis': *huishuur, huisnummer, huisregels, huisstijl, huisvestingsbeleid.* Dynamische woorden beginnen met 'huis' zijn bijvoorbeeld *huiszoeking* en *huisarrest*.

Hoe gaan we om met woorden die twee labels hebben?
Een groot aantal ambiguïteiten is uit de lijst gehaald, maar er blijven er een kleine 500 over. Die nopen T-Scan tot keuzes. In de lijst zijn in kolom B zijn de uiteindelijke keuzes gegeven, terwijl in kolom E de oude ambiguïteiten zijn vermeld.

In keuzes tussen concrete klassen hebben we Concrete substanties het primaat gegeven, gevolgd door Artefacten, Planten/Dieren, Voeding en verzorging, Concreet Overig en Personen.

In de zeldzame keuzes tussen concrete klassen en de abstracte klassen Organisatie primeert de concrete klasse als het een Artefact is, maar niet als het een Persoon is. Zo wordt de persoonlijkheid van een tekst niet overschat.

In keuzes tussen abstracte klassen krijgt Organisatie de voorrang boven een Abstract Gebeuren en een niet-dynamische interpretatie de voorrang boven de dynamische.

Zie het overzicht in onderstaande tabel.

Treft T-Scan bij een woord de volgende labels:	*… dan kiest het als label:*	*Voorbeeld*
Substance_conc,artefact	Substance_conc	*Baksteen*
Substance_conc,concrother	Substance_conc	*Aarde*
Substance_conc,nonhuman	Substance_conc	*Nerts*
Artefact,concrother	Artefact	*Bassin*
Artefact,concrother,nonhuman	Artefact	*Bies*
Artefact,dynamic_conc	Dynamic_conc	*Voetbal*
Artefact,human	Artefact	*Duiker*
Artefact,nonhuman	Artefact	*Amfibie*
Nonhuman,concrother	Nonhuman	*Palm*
Nonhuman,human	Nonhuman	*Luiaard*
Voed_verz,concrother	Voed_verz	*Berenklauw*
Concrother,human	Concrother	*Bierbuik*
Artefact,institut	Artefact	*Golfclub*
Human,institut	Institut	*Grenswacht*
Dynamic_abstr,dynamic_conc	Dynamic_abstr	*Nek-aan-nek-race*
Dynamic_abstr,institut	Institut	*Bestuur*
Dynamic_abstr,nondynamic	Nondynamic	*Productie*

Bijlage E. Semantische klassen voor adjectieven

1 De nieuwe kenmerken

Uitgaande van de lijst die het RBN aanlevert (bijna 9.000 adjectieven), is een nieuwe classificatie opgesteld. De opbouw daarvan is als volgt (de onderstreepte termen komen voor in de lijst). Verderop wordt de classificatie toegelicht en geïllustreerd.

1. Waarn_mens: de waarneembare kenmerken van mensen
2. Emosoc: emotionele kenmerken en sociaal gedrag van mensen
3. Waarn_niet_mens: waarneembare kenmerken van stoffen, objecten en organismen. Subgroepen daarbij (in de derde kolom) zijn:
    a. Vorm_omvang
    b. Kleur
    c. Stof
    d. Geluid
    e. Waarn_niet_mens_ov: overige waarneembare kenmerken
4. Technisch: kenmerken van stoffen, objecten en organismen die alleen met technieken waarneembaar zijn
5. Time
6. Place
7. Specifiek evaluatief
    a. Spec_positief: inhoudelijk positief (*onoverwinnelijk, onverslijtbaar*)
    b. Spec_negatief: inhoudelijk negatief (*lawaaierig, demagogisch, onrechtmatig*)
8. Algemeen evaluatief (algemene oordelen over (on)wenselijkheid, (on)toelaatbaarheid, effectiviteit of schoonheid).
    a. Alg_positief: evaluatief positief (*aanbevelenswaard, effectief, mooi*)
    b. Alg_negatief: evaluatief negatief (*onverstandig*)
    c. Alg_evaluatief: evaluaties zonder vaste richting (*aanmerkelijk*)
9. Epistemisch evaluatief
    a. Ep_positief: epistemisch positief (*steekhoudend*)
    b. Ep_negatief: epistemisch negatief (*onzinnig*)
10. Overig abstract: de niet-evaluatieve abstracte woorden
11. Undefined

De volgende groeperende kenmerken worden gevormd:

- Specifiek oordelende adjectieven (7a en 7b)
    - Spec_oordeel_bvnw_p
    - Spec_oordeel_bvnw_d
- Algemeen oordelende adjectieven (8a, 8b en 8c)
    - Alg_bvnw_p
    - Alg_bvnw_d
- Epistemische adjectieven (9a en 9b)
    - Ep_bvnw_p
    - Ep_bvnw_d
- Strikt-concrete adjectieven: dat betreft klassen 1,2 en 3
    - Conc_bvnw_strikt_p
    - Conc_bvnw_strikt_d
- Ruim-concrete adjectieven: dat betreft klasen 1, 2, 3, 5 en 6
    - Conc_bvnw_ruim_p
    - Conc_bvnw_ruim_d

- Subjectieve adjectieven: dat zijn de klassen 7 t/m 9
    - Subj_bvnw_p
    - Subj_bvnw_d

Ten slotte hebben we enkele maten die aangeven hoe de dekking is van onze lijst in de tekst:
- De proportie adjectieven die 'undefined' blijft:
    - Undefined_bvnw_p
- De totale proportie die een specifieke lezing krijgt:
    - Gelabeld_bvnw_p
- De totale proportie adjectieven die in de lijst staat:
    - Gedekte_bvnw_p: Gelabeld_bvnw_p + Undefined_bvnw_p

2 Toelichting en voorbeelden bij de semantische typen

Waarneembare en fysieke kenmerken van mensen
Het gaat hier om het menselijk lichaam in ruime zin:
- Kleding en verzorging (*poedelnaakt, aangekleed, morsig*)
- Fysieke kenmerken (*welgeschapen, rijzig, roodharig, bebloed, besneden*)
- Fysieke condities en klachten (*rillerig, sneeuwblind, verkouden, soezerig, bekaf, hardhorend, invalide*) en effecten daarop (*vermoeiend*)
- Lichaamshoudingen (*schrijlings, kruipend*)

Emotionele kenmerken en sociaal gedrag van mensen
Het gaat hier om:
- Emoties in strikte zin (*aangedaan, aangeslagen, overgelukkig*)
- Stemmingen in ruimere zin (*radeloos, panisch, opgetogen*)
- Veroorzakers daarvan (*aandoenlijk, aangrijpend, afschrikwekkend*)
- Karaktereigenschappen (*roekeloos, praatgraag, rebels*)
- De houding waarmee mensen dingen doen (*achteloos, routineus, pretentieloos, onverstoorbaar*)
- De houding van mensen tegenover anderen (*vriendelijk, respectvol, onuitstaanbaar*)
- Misleidend gedrag (*steels, stiekem*)

Een adjectief kan alleen 'emosoc' krijgen als het met name voor personen gebruikt wordt. Dat geldt bijvoorbeeld niet voor 'dominant', 'spannend' en 'diplomatiek'.
Onder 'emosoc' vallen niet:
- opvattingen (*conservatief*) en liefhebberijen (*cinefiel*);
- objectieve kenmerken van personen als *Fransprekend, chassidisch, woordblind, dakloos* of *woonachtig*;
- lichamelijke of psychische condities als *ongesteld, vermoeid, ontoerekeningsvatbaar, bipolair*;
- cognitieve kenmerken als 'onwetend' of 'intelligent'.

Waarneembare kenmerken van stoffen, objecten en organismen
Het gaat hierbij om niet-menselijke entiteiten, met vijf subgroepen van kenmerken.
1. Omtrek: hiermee zijn omvang en vorm bedoeld (*metershoog, flinterdun, achthoekig, bolvormig, duinachtig*)
2. Kleur: hiermee is gedoeld op kleur, glans, en licht/zichtbaarheid meer in het algemeen (*blauw, asgrauw, stikdonker*)
3. Stof: materiaal, vochtigheid, substantie, oppervlak, transparantie (*bakstenen, doornat, klonterig, ribbelig, doorschijnend*)
4. Geluid: het gaat hier om een kleine groep woorden als *gedempt, (on)hoorbaar, (on)verstaanbaar, luid, (half)luid, gehorig, muisstil, nasaal, sonoor*
5. Overig: een vrij heterogene categorie, waarin we kenmerken aantreffen zoals temperatuur (*koud*; incl. weersomstandigheden (*zonnig*)), smaak, geur, gewicht; bewerkingen (*ongebrand;*

opblaasbaar, roestbestendig, bebouwbaar, braakliggend) (phyper) en functionaliteit (*defect, kaduuk, onklaar)*.

Technische kenmerken
Veel kenmerken van stoffen, objecten en organismen zijn alleen met technieken waarneembaar:

- Chemische, biologische, natuurkundige, elektrische en medische eigenschappen (*afbreekbaar, radioactief, bloeddrukverhogend, brachiaal*, *elektropneumatisch ongewerveld, brandgevaarlijk*). Onder biologisch vatten we ook adjectieven die naar een bepaald lichaamsdeel verwijzen, voor zover niet zichtbaar: *hepatisch, nefrotisch.*
- Ingrediënten, bestanddelen (*siliciumhoudend*)

Het onderscheid tussen waarneembare en technische kenmerken is van belang om concreetheid te definiëren: technische kenmerken zijn wel materieel in de zin van stoffelijk, maar niet concreet in de zin van zonder hulpmiddelen zintuiglijk waarneembaar.

Time
Het gaat hier om woorden die verwijzen naar:

- tijdsduur (*achturig, avondvullend*)
- tijdstippen (*dinsdags, nachtelijk*)
- leeftijd (*aloud*)
- historische perioden (*napoleontisch, naoorlogs*)
- begin en voltooiing (*aanvankelijk, afgerond*)
- verandering en continuïteit (*blijvend, chronisch, acuut*)
- snelheid (*bliksemsnel, treuzelachtig*)
- periodiciteit (*cyclisch, geregeld*)
- volgorde (*eerstkomend, successief*)
- en naar verleden, heden en toekomst (*voormalig, komend, huidig*).

Place
Het gaat hier om:

- relatieve locaties (*aangrenzend, afgelegen, buitenst*)
- geografische locaties (*Afrikaans, Ardenner, gewestelijk, multinationaal*)
- (wind)richtingen (*oostelijk, benedenwaarts, overdwars*)
- kenmerken van locaties of gebieden (*onoverdekt, bebouwbaar, bosrijk, ongelijkvloers*)

Specifiek evaluatief
Het gaat hier om woorden die aan een bepaalde kwaliteit refereren en daaraan een positief of negatief oordeel koppelen.

- Inhoudelijk positief (*baanbrekend, bedreven, bekoorlijk, doortimmerd, evenwichtig*)
- Inhoudelijk negatief (*lawaaierig, demagogisch, onrechtmatig*)

Algemeen evaluatief
Het gaat hier allereerst om algemene oordelen over (on)wenselijkheid, (on)toelaatbaarheid, effectiviteit of schoonheid. Er is niet duidelijk een kwaliteit aanwijsbaar die de basis vormt voor het oordeel. Bijvoorbeeld: *onverstandig* is algemeen evaluatief, terwijl *onrechtmatig* een juridische onwenselijkheid aangeeft.

- Evaluatief positief (*aanbevelenswaard, effectief, mooi*)
- Evaluatief negatief (*onverstandig*, *voorbeeldig*)

Er is soms twijfel tussen emotionele woorden (emosoc) en evaluatieve: *deerniswekkend* is letterlijk genomen emosoc, maar lijkt met name negatief-evaluatief te worden gebruikt. Hetzelfde geldt voor *aangenaam* en *onaangenaam*. *Afschrikwekkend* daarentegen is als emotioneel gezien.

Naast de evaluaties met een duidelijk positieve of negatieve richting zijn er evaluatieve adjectieven die wijzen op het belang, de omvang of de intensiteit van een verschijnsel: *aanmerkelijk, volslagen, tomeloos, minimaal*. De 'sterke' woorden in deze groep komen vaak bij de intensiveerders terug als intensiverend adjectief (zie Bijlage I). In het kader van de adjectiefclassificatie worden ze als algemene

evaluaties gezien. Maar onder de algemene evaluaties vallen dus ook adjectieven die juist de geringe omvang van iets aangeven.

Epistemisch evaluatief
Bij epistemische evaluaties gaat het om het waarheidsgehalte of de plausibiliteit van uitspraken of om kenmerken van mensen die hen ertoe brengen om in onjuistheden te geloven.

- Epistemisch positief (*steekhoudend, accuraat, evident, gegrond*)
- Epistemisch negatief (*aangedikt, aanvechtbaar, omstreden*)

Overige abstracte woorden
Woorden die bij geen van de voorgaande groepen zijn onder te brengen, worden als 'overige abstracte' betiteld, het gaat om zeer verschillende woorden als *aansprakelijk, aanspeelbaar, aangeboren, accentloos, academisch, actuarieel*.

Niet-gedefinieerde woorden
Er zijn allerlei adjectieven met veel lezingen. Om geen al te grote onnauwkeurigheden te introduceren, zijn die woorden ongedefinieerd gelaten. Een paar voorbeelden:

- *Aardig* kan een emotioneel woord zijn, een evaluatief woord, of een versterker.
- *Diep* kan van alles betekenen, van ruimtelijke lezingen tot versterkende lezingen.

Er zijn ook een paar woorden ongedefinieerd gelaten omdat ze door Frog als adjectief gelabeld worden, maar veelal als verbindingswoord gebruikt worden. *Eerder* is een tijdsadjectief maar ook vaak een contrastief verbindingswoord. Hetzelfde geldt voor *allereerst*. En *anders* heeft een lezing als conditioneel verbindingswoord.

Bijlage F. Concreetheid van werkwoorden

1 Hoe zijn de groepen onderscheiden?

Als uitgangspunt is genomen de lijst van 6657 werkwoorden die T-Scan ook gebruikt om te bepalen of het gaat om acties, processen of toestanden. Die lijst is simpelweg verdeeld in concreet, abstract en 'undefined'. In de herziene werkwoordenlijst is het feature 'concreetheid' te vinden in kolom D.

Als concreet gelden alle werkwoorden die een zintuiglijke voorstelling oproepen. Het gaat dus om acties, processen en toestanden die je kunt zien, horen of voelen. Het gaat dan om werkwoorden die verwijzen naar:

- Fysieke acties van personen jegens anderen (*aaien, aanbellen, aanblikken, begluren, doodschieten;* maar niet *doden* en *executeren*, omdat daar vele methoden voor zijn)
- Fysieke acties van personen jegens objecten (*afstoffen, afruimen, amputeren, blancheren, bladeren, doorzeven, dorsen, draineren* enz.)
- Fysieke acties en reacties (*ademen, hoesten, proesten*)
- Het produceren van objecten (incl. afbeeldingen): *etsen, tekenen, fabrieken, flansen*
- Non-verbaal gedrag (*bekkentrekken, glimlachen, bescheuren*)
- Spelletjes, sporten en bewegingen die een visueel beeld oproepen (*skiën, schaken, buikdansen, crawlen, dobbelen, eenendertigen*)
- Eten en drinken (*oppeuzelen, bedrinken, brunchen, doorslikken*; maar niet *slikken*, want dat heeft ook een niet-concrete lezing)
- Zichtbare ingrepen in het landschap (*afdammen, omheinen, asfalteren*; maar niet *indammen*, want dat heeft ook een niet-concrete lezing)
- Wijzen van spreken met een bepaalde klank (*snauwen, prevelen, mompelen, ginnegappen*)
- Geluiden maken (*burlen, tsjilpen, claxonneren, croonen, gakken*)
- Iets verplaatsen en zich voortbewegen (*aanmeren, aanrennen, afnokken, moven, banjeren, rondwandelen, douwen, verjagen*)
- Trajecten afleggen (*cirkelen, omvaren*)
- Zintuiglijke waarneming (niet *horen, zien* e.d. omdat die werkwoorden veelal abstract zijn; wel *ontwaren, achteromkijken*, *afluisteren*)
- In een enkel geval gaat het om andere zintuigen dan ogen of oren, zoals bij *verwarmen, opwarmen, afkoelen, kleumen, vernikkelen*.

Twijfelgevallen zijn collectieve of minder zichtbare acties, zoals *belegeren* en *bemalen*. Omdat bij *belegeren* een visuele voorstelling eenvoudiger gevormd wordt dan bij *bemalen* is alleen *belegeren* als concreet gezien.

Emoties (bv. *schrikken, griezelen* of *volschieten*) zijn voorlopig niet als concreet gelabeld. Overigens gaat het hier om een klein aantal werkwoorden, die vaak ook niet-concrete lezingen hebben (bv. *teleurstellen* hoeft geen emotie aan te duiden; het kan ook verwijzend naar een evaluatie).

Concreet is iets anders dan bekend. Net als de lijst met nomina en adjectieven bevat de werkwoordenlijst bevat heel wat woorden die vrijwel onbekend zijn, maar toch concreet. *Roten* betekent 'het blootstellen van vlasstengels aan water, zodat de vlasvezels vrijkomen', *wiegelen* betekent 'deinen, schommelen'.

Net als bij de adjectieven is het predicaat concreet niet toegekend aan kenmerken of processen die alleen technisch waarneembaar zijn, zoals *infecteren* of *fluoreren*.

Omdat bij de nomina en de adjectieven de plaats- en tijdwoorden tot concreet-in-ruime-zin worden gerekend, is het goed om bij de werkwoorden ook op deze categorieën te letten. Maar de afbakening valt hier anders uit.

De werkwoorden die 'ruimtelijk' zijn, zijn meestal ook visueel voorstelbaar, dus die bevinden zich al in de concrete categorie. Wat nu te doen met 'tijd-werkwoorden' als *uitstellen, vervroegen, versnellen, vertragen, vervroegen*? Door de beperkte tijd die beschikbaar is voor de codering, zien we er bij de

werkwoorden vanaf om onderscheid te maken tussen strikt-concreet (in strikte zin zijn tijdwoorden niet concreet) en ruim-concreet (in ruimere zin zijn tijdwoorden dat wel). We beperken ons tot het markeren van strikt-concrete woorden: tijdwoorden vallen daarbuiten.

Het criterium van voorstelbaarheid is ook in andere opzichten strikt gebruikt. Een werkwoord als *versturen* bijvoorbeeld is niet als concreet gezien. Het versturen van een brief is niet voorstelbaar: het *posten* ervan wel.

Heel wat werkwoorden worden zowel in concrete als in niet-concrete betekenissen gebruikt, zoals *graven* (kan ook 'onderzoeken' zijn), *gooien* (kan ook op abstracte objecten slaan; dat geldt niet voor bijvoorbeeld *omvergooien* of *afgooien*), *herademen* ('opgelucht zijn'), *liggen* ('dat ligt gevoelig'), *staan* ('het staat er goed voor'), *regenen* ('het regent klachten'), *verpakken* ('hij verpakt de boodschap handig'), *verschuilen* ('zij verschuilt zich achter haar superieuren / haar principes'), *vertrappen* ('onze rechten worden vertrapt'), *zitten* ('hij zit mij te dicht op de huid') enz. En *toejuichen* heeft een prominente niet-concrete lezing is ('ergens tevreden over zijn'), net als bijvoorbeeld *touwtrekken* ('ergens langdurig over in conflict zijn'), *trappelen* ('ongeduldig zijn'), *afbouwen* ('een activiteit beëindigen'), *uitkleden* en *uitknijpen* ('iemand financieel benadelen') enzovoort.

Dit soort werkwoorden heeft het 'undefined' gekregen, wat zoveel wil zeggen als 'mogelijk abstract'. Het gaat hier dus om twee categorieën werkwoorden:

- werkwoorden die zowel concrete als niet-concrete lezingen hebben; beide lezingen zijn conventioneel verbonden met het werkwoord;
- werkwoorden die open zijn van betekenis, zoals *zitten, staan* en *liggen*. Die lenen zich voor een groter aantal lezingen, die niet direct op te sommen vallen.

In een enkel geval lijkt de niet-concrete lezing duidelijk minder prominent dan de concrete. Dat geldt bijvoorbeeld voor de niet-concrete lezing van *begraven* ('de strijdbijl ….'). Daarom is *begraven* als concreet gelabeld. Uiteraard zouden al deze beslissingen met corpusevidentie ondersteund moeten worden; daarvoor was helaas geen tijd. In die zin is de lijst noodzakelijkerwijs op intuïties gebaseerd.

Er zijn ook werkwoorden die altijd abstract zijn: *argumenteren, verantwoorden, bezweren* enzovoort. Maar ook v*ertroetelen* is abstract, omdat niet duidelijk is met welke attenties het gebeurt; en *grossieren*, omdat het staat voor 'over iets in overvloed beschikken'. In het geval *ondersneeuwen* is besloten dat de abstracte lezing van het voltooid deelwoord 'ergens te weinig aandacht voor hebben' zwaarder weegt dan de concrete lezing. Voor situaties waarin iets daadwerkelijk met sneeuw bedekt is, wordt veelal *insneeuwen* en niet *ondersneeuwen* gebruikt.

Een fragment uit de lijst met voorbeelden van concrete en niet-concrete woorden volgt in onderstaande tabel, met enkele actiewerkwoorden. De niet-concrete woorden kunnen abstract zijn of ongedefinieerd.

Werkwoord	**Concreet?**	**Toelichting**
Aanbevelen	Nee	
Aanbinden	Ja	Dit ondanks de uitdrukking *de kat de bel aanbinden*
Aanhechten	Ja	
Aankruipen	Ja	
Aankweken	Nee	Je kunt talenten aankweken (abstract)
Aanlanden	Ja	Een strikt ruimtelijke betekenis
Aanleren	Nee	
Aanmerken	Nee	
Aanpoten	Nee	Kan concreet en niet-concreet zijn ('zijn best doen')
Aanpraten	Nee	
Aanroepen	Nee	Betreft spreken, maar geeft niet aan hoe dit spreken klinkt
Aanroeren	Nee	
Aanscherpen	Nee	Je kunt ook beleid of normen aanscherpen

Aanschrijven	Nee	
Aanschroeven	Ja	
Aansmeren	Nee	
Aansnijden	Nee	Je kunt een taart en een thema aansnijden
aanstellen	Nee	Je kunt je op allerlei manieren aanstellen
aantonen	Nee	
aanvallen	Nee	Kan fysiek zijn of overdrachtelijk
aanvegen	Ja	Dit ondanks de uitdrukking *de vloer aanvegen met iemand*
accepteren	Nee	
achterhouden	Nee	Je kunt een object achterhouden of een stuk informatie
achternazitten	Ja	
achtervolgen	Nee	Je kunt iemand fysiek en overdrachtelijk achtervolgen

2 Kenmerken

Op basis van dit alles worden zes kenmerken onderscheiden. Daarnaast noemen we nog twee samengestelde concreetheidskenmerken, die over drie woordgroepen heen aggregeren.

Conc_ww_p	Proportie van concrete werkwoorden
Conc_ww_d	Dichtheid van concrete werkwoorden op werkwoorden
Abstr_ww_p	Proportie van abstracte werkwoorden
Abstr_ww_d	Dichtheid van abstracte werkwoorden op werkwoorden
Undefined_ww_p	Proportie van werkwoorden die in de lijst ongedefinieerd blijven op werkwoorden
Gedekte_ww_p	Proportie van werkwoorden die in de lijst staan op werkwoorden
Conc_strikt_tot_d	Opgetelde dichtheden van strikt-concrete naamwoorden, strikt-concrete bijvoeglijke naamwoorden en concrete werkwoorden
Conc_ruim_tot_d	Opgetelde dichtheden van ruim-concrete naamwoorden, ruim-concrete bijvoeglijke naamwoorden en concrete werkwoorden

Bijlage G. De classificatie van werkwoorden naar actie, proces of toestand

1 Hoe is geclassificeerd?

In de herziene werkwoordenlijst zijn de werkwoorden opgedeeld in vier waarden: action, process, state en undefined. (In T-Scan wordt overigens gewerkt met de Nederlandse termen actie, proces en toestand.)

Uitgangspunt voor de T-Scanlijst was de RBN-lijst die zo'n 6600 werkwoorden telt. De RBN-classificatie (Martin & Maks 2005) wordt gepresenteerd is als voortkomend uit twee parameters, Dynamiek en Controle.

	Action	**Process**	**State**
Dynamic	+	+	-
Control	+	-	-

2.1 Het onderscheid naar dynamiek

Dynamic en nondynamic worden bijvoorbeeld geïndiceerd door combineerbaarheid met *is aan het X-en*. Dit testframe lijdt soms tot twijfel: het staat woorden als *tochten* en *wriemelen* toe, die in de oorspronkelijke lijst als State gelabeld zijn. Maar wellicht zijn dit ook processen.

Maar ook afgezien daarvan lijkt het beter om Non-dynamism te definiëren als wel of geen gebeurtenis cq. gebeuren; dat betekent niet meer dan iets wat in de tijd naar zijn aard begrensd is. Enkele voorbeelden van werkwoorden die in de oorspronkelijke lijst als State voorkomen, maar nu op basis van het criterium wel/geen verandering zijn omgecodeerd tot Process: *laaien, ontkomen, terugdeinzen*.

Dynamische woorden zijn, ook volgens de RBN-logica, combineerbaar met *langzamerhand / geleidelijk*. Toch werkt het criterium niet probleemloos. Neem *mislukken*; dat lijkt slecht te combineren met *langzamerhand*:

- *Mijn poging mislukte langzamerhand

Toch is *mislukken* duidelijk een gebeuren. Een gebeuren combineert met een tijdsbepaling:

- Mijn poging mislukte gisteren

Een toestand als *beseffen* laat zo'n bepaling minder makkelijk toe:

- ?Ik besefte gisteren dat ik fout zat

Hiermee worden cognitiewoorden uitgesloten als proces. Interessant is dat de tijdsbepaling wel combineert met het modale *blijken*:

- Gisteren bleek / vandaag blijkt dat ik fout zat.

Ook emotionele ervaringen lenen zich slecht voor tijdsbepalingen: *griezelen, hallucineren, ijzen*. Maar die ervaringen lijken naar hun aard tijdelijk, en roepen daarmee het idee van een gebeurtenis op. Hetzelfde geldt voor *plaatsvinden*.

Dat ligt anders voor woorden als *mokken, piekeren, sappelen* en *tobben*, die een cyclisch proces voor de geest roepen. Wel is er nog steeds een proces aan de orde. Maar je kunt wel zeggen:

- Het was sappelen, afgelopen jaar.

Sommige woorden hebben zowel een proces- als een state-lezing, zoals *dromen*. Dat woord kan slaan op een tijdelijke geestestoestand waar iemand doorheen gaat, maar ook in overdrachtelijke zin ('hij droomt ervan beroemd te zijn') op stabiel gekoesterde aspiraties.

Twijfel tussen proces en state is er ook bij woorden als *generen, frapperen* en *frustreren*:

- Het frustreert mij dat ik kaal ben (state)
- Het frustreerde mij dat hij zijn ongelijk niet toegaf (proces)

Telkens kan het woord verwijzend naar een op zeker moment plaatshebbend mentaal effect of een gedurige mentale toestand of attitude. Dat geldt niet voor *tevredenstellen*, dat de state-lezing minder heeft.

Voorbeelden van andere werkwoorden die oorspronkelijk State waren, maar zijn omgelabeld tot Proces: *fikken, watertanden, ontwaren, jeuken, verjaren*. Telkens gaat het om gebeurens. Ook geluiden als *knerpen* en *knarsen* kennelijk deze tijdelijkheid.

2.2 Het onderscheid naar controle

Bij Control wordt geen testframe of diagnostiek genoemd. Het is echter aannemelijk dat action-werkwoorden kunnen functioneren in de volgende zinsframes:

- Ik *ben van plan om* te X-en
- Ik X *met opzet*
- *Ik probeer te* X-en

In de oude lijst staan als action genoemd werkwoorden als *compromitteren, corroderen, institutionaliseren, ontnuchteren, verontrusten, verongelijken* en *desillusioneren*. Dat zijn eigenlijk processen, geen acties, omdat er geen controle is. Dat geldt ook voor een aantal werkwoorden die duidelijk menselijke activiteiten aanduiden, maar wederom zonder controle: *achteruitdeinzen, blunderen, hannesen, hoesten, indommelen, ineenkrimpen, kotsen, morsen, neerzijgen, raaskallen, uitgillen, ijlen* e.d. Dat zijn immers dingen die nooit met opzet worden gedaan. In een enkel geval wordt iets veelal niet met opzet gedaan, maar kan dat wel: *ademhalen* ('na 30 seconden haalde ik pas weer adem'), *geeuwen* ('hij geeuwde onbeschaamd'), *nagelbijten, huilen, vervuilen* (hoewel dat laatste woord ook een proces-lezing heeft: 'die stof vervuilt ons drinkwater').

In tegenstelling tot gevallen als *blunderen* worden intentionele maar mislukte acties wel als actie gezien: *overbelichten, misslaan, verstappen*.

Dat geldt ook voor diergedrag (*blaten, burlen, koeren, kwaken, klapwieken, kwispelstaarten*); die gedragingen zijn als actie gelabeld, hoewel de intentionaliteit van dit gedrag twijfelachtig is.

Een probleem lijkt dat sommige acties, dingen dus die met opzet gedaan kunnen worden, niet erg dynamisch lijken: *weerstaan, stilzitten, handhaven, opblijven, vrijhouden, zwijgen*. Het zijn zeldzame gevallen, maar ze betekenen dat wel/geen controle het hoofdcriterium is voor het actielabel. Het gaat om een 'wilsbesluit'. Nu zou je kunnen beweren dat voor het overeind houden van dat besluit een zekere 'tegendruk' tegen situationele invloeden nodig is, zodat uiteindelijk twee krachten elkaar in evenwicht houden; het voorkómen van een gebeurtenis kun je ook een gebeurtenis noemen.

Wie dat niet overtuigend vindt, kan een tweede kolom toevoegen aan ons schema ten behoeve van acties die niet dynamisch zijn:

	Action	**Action**	**Process**	**State**
Dynamic	+	-	+	-
Control	+	+	-	-

Het gaat hier om een klein aantal gevallen. En er is iets voor te zeggen om gevallen als 'tegenhouden' dynamisch te noemen. Daarom is afgezien van deze extra kolom.

2.3 Meerduidigheden

De oude lijst bevat veel meerduidige woorden, met tot wel 7-8 lezingen. Als het gaat om lezingen van hetzelfde type (bv. 'state,state,state') zijn deze blijven staan. Meerduidigheden als 'action,action,process') zijn opnieuw bekeken. Deze coderingen zijn eerst vereenvoudigd tot vier soorten combinaties van de drie hoofdtypen. Als sprake is van een meerduidigheid, is die te vinden in kolom C in de herziene werkwoordenlijst.

De tabel aan het eind van deze bijlage geeft voorbeelden van eenduidige en meerduidige werkwoorden. De meest voorkomende meervoudige lezing is action / process. Heel wat werkwoorden kunnen zowel een intentionele handeling beschrijven als een proces dat zich buiten de mens om voltrekt; dit is een reguliere bron van polysemie.

Minder frequent zijn combinaties van een action- en een state-lezing. Woorden als *paren* ('combineren' / 'copuleren') en *letten* ('aandacht besteden' / 'weerhouden') zijn eerder ambigu dan polyseem. Meer regulier zijn taalhandelingswerkwoorden die ook gebruikt worden als beschrijving van betekenisrelaties: *beantwoorden, tegenspreken*.

In een enkel geval is een minder frequente lezing veronachtzaamd. Bij *snappen* is bijvoorbeeld gekozen voor de state-lezing, en de archaïsche lezing 'een overtreding opsporen' buiten beschouwing gelaten. Bij *toelachen* (actie) is de uitdrukking 'het geluk lacht ons toe' genegeerd. En bij *uitnodigen* (actie) is de minder frequente state-lezing ('dat nodigt uit tot geweld') veronachtzaamd. Maar bij *spreken* (actie) is dat niet gedaan voor de state-lezing ('dat spreekt vanzelf / dat spreekt tegen zijn standpunt'), die frequenter lijkt dan de state-lezing van *uitnodigen*. Dit soort beslissingen blijft enigszins discutabel.

Zoals gezegd zijn de meerduidigheden zijn gemeld in een aparte kolom (C). Een ervan is gedesambigueerd in kolom B: Action / Proces > proces. Zo wordt voor een bepaalde tekst wellicht het aantal acties onderschat en het aantal processen overschat. Dat is onnauwkeurig, maar zo wordt althans het kenmerk dynamisch correct benoemd.

De andere typen meerduidigheid zijn diepgaander van aard: het zijn vaak totaal verschillende lezingen (zie onderstaande tabel voor voorbeelden). Daarom zijn deze meerduidigheden in kolom B omgezet in het label Undefined:

- Action / state > undefined
- Process / state > undefined
- Action / process/ state > undefined.

Lezing	**Voorbeelden**
Action	Doorzeuren, aanbesteden, afgelasten, bedotten, wegstoppen
Process	Ineenstorten, meemaken, omhooggaan, (mis)lukken, tanen, doorlekken, openrijten, nekken, blameren, bezuren, opleven, ontspinnen
State	Vriezen, toeven, toeschijnen, hopen, stoelen, verdrieten, waarderen, beangstigen, beseffen, dralen, suffen
Action / process Omgezet in Action	Ontkrachten, tekeergaan, meesleuren, verschaffen, dooddrukken, doorboren, omranden, creëren, kenmerken, isoleren, normaliseren, ontbinden, transformeren, verrassen, kronkelen, vatten, wegnemen, opslaan, aanbreken, aanhouden, breken, neerslaan Bijvoorbeeld: - Hij gaat tegen haar tekeer (actie) - De storm gaat tekeer (proces) - Hij breekt het brood (actie) - Hij breekt twee glazen (proces)
Action / state Omgezet in Undefined	Paren, beantwoorden, letten, hobbelen, aanvaarden, corresponderen, dienen, dreigen, overeenkomen, vloeken Bijvoorbeeld: - Hij beantwoordt de vraag daarmee niet (action) - Dit beantwoordt aan onze eisen (state)
Process / state Omgezet in Undefined	Dromen, frustreren, meevallen, verbazen, meevallen, ontroeren, teleurstellen, toekomen, uitwijzen, ruiken, horen, verwonderen, verstaan Bijvoorbeeld: - Zijn antwoord verbaasde mij (proces) - Zijn werklust verbaast mij telkens weer (state)
Action / process / state Omgezet in Undefined	Bijdragen: - Ik draag graag mijn steentje bij (actie) - Dat droeg bij aan hun verwijdering (proces) - Dat draagt bij aan mijn tevredenheid (state) Leven: - Ik wil groots en meeslepend leven (actie) - Deze plant leeft nog (proces) - Het leeft onder de mensen (state) Verschaffen: - Hij verschafte haar een alibi (actie) - Onze zuinigheid verschaft ons de ruimte voor nieuw beleid (proces) - Dat verschaft geen vrijbrief voor geweld (state) Hechten: - Hij hechtte de wond (actie) - Zij hechtte zich aan haar stiefmoeder (process) - Ik hecht sterk aan discretie (state) Maken: - Ik maak graag nasi - Dat maakt veel tongen los - Hij maakt het goed

Bijlage H. Voorzetseluitdrukkingen

aan de hand van
aan het adres van
afgezien van
al naargelang
al naargelang van
als gevolg van
bij de gratie van
bij monde van
bij wijze van
buiten medeweten van
door gebrek aan
door middel van
door toedoen van
gezien het feit dat
in antwoord op
in de loop van
in de richting van
in de trant van
in een poging om
in geval van
in het geval dat
in het kader van
in het licht van
in naam van
in opdracht van
in overeenstemming met
in overleg met
in plaats van
in reactie op
in strijd met
in tegenstelling met
in tegenstelling tot
in termen van
in verband met
in verhouding tot
in weerwil van
in zoverre als
met als gevolg dat
met behoud van
met behulp van
met betrekking tot
met dank aan
met gebruikmaking van
met het oog op
met het doel om
met inachtneming van
met ingang van
met medewerking van
met medeweten van
met uitzondering van
met weglating van
na afloop van
na verloop van
naar aanleiding van
naargelang van
naarmate van
omwille van
ondanks het feit dat
onder invloed van
onder leiding van
onder verwijzing naar
op advies van
op basis van
op de volgende wijze
op grond van
op het gebied van
op het stuk van
op initiatief van
op kosten van
op uitnodiging van
op vertoon van
op verzoek van
op voorspraak van
te midden van
tegen betaling van
ten aanzien van
ten bate van
ten bedrage van
ten behoeve van
ten gerieve van
ten gevolge van
ten gunste van
ten koste van
ten nadele van
ten opzichte van
ten overstaan van
ten tijde van
ten voordele van
ter attentie van
ter gelegenheid van
ter hoogte van
ter wille van
ter zake van
uit een oogpunt van
uit het oogpunt van
uit hoofde van
uit kracht van
uit naam van
van de kant van
van de zijde van
voor rekening van

Bijlage I. Intensiveerders in T-Scan

1 Inleiding

T-Scan put uit een lijst van ruim 3700 sterke uitdrukkingen. De laatste versie van de lijst telt ongeveer 1120 adjectieven (bv. *zielsgelukkig*), 35 adjectieven die in 'bijwoordelijk' gebruik een versterker zijn (*knap*), zo'n 125 bijwoorden (*zienderogen*), 220 combinaties (*zeker en vast*), ongeveer 1535 nomina (*zenuwpees, stortregen*), 650 werkwoorden (*wemelen*) en zo'n 35 tussenwerpsels (*ammehoela*).

Eerst behandelen we definities en tests, daarna de gebruikte bronnen, en daarna de 7 woordsoorten die intensiveerders opleveren. Aan het eind worden de kenmerken omschreven die T-Scan met behulp van deze lijst oplevert.

2 Definities en tests

Onder intensiveerders verstaan we:

- *Sterke* woorden of uitdrukkingen (verder: woorden), woorden dus die verwijzen naar een bijzonder hoge graad van een bepaalde eigenschap (bijvoorbeeld *fenomenaal*)
- *Versterkende* woorden, die de interpretatie van versterken van de uiting waar ze in staan (bijvoorbeeld *hogelijk*).

Of iets een intensiveerder is, kan gecontroleerd worden door verschillende testframes. Ze zijn niet per stuk doorslaggevend, maar geven wel indicaties.

1. De *zelfs*-test gaat ervan uit dat *zelfs* dat het zinsdeel dat volgt argumentatief sterker is dan het voorgaande zinsdeel. van Die houdt in dat de volgende sequentie acceptabel moet zijn (N is een meer neutrale uitdrukking, I de intensivering):

- *N. Zelfs I.*
- Hij was gelukkig. Zelfs zielsgelukkig.

2. Met nominale kwalificaties is *zelfs* lastiger toe te passen. Beter toepasbaar is daarbij de s*terker nog*-test:

- *N. sterker nog, I.*
- Hij is ongemanierd. Sterker nog, hij is een barbaar.

3. Het omgekeerde patroon kunnen we zien bij de *niet eens*-test:

- *Niet I. Niet eens N.*
- Nee, ik vind hem geen barbaar. Hij is niet eens ongemanierd.

4. Een ander patroon dat past bij intensiveringen is de metalinguïstische negatie. Als die bruikbaar is, voor een uitdrukking, gaat het in principe om een versterking:

- Hij is niet (gewoon) ongemanierd, hij is een barbaar.
- Ik ben niet gelukkig, ik ben extatisch.

5. Gradeerbare intense elementen kunnen veelal niet gecombineerd worden met bijwoordelijke verzwakkers, maar wel met bijwoordelijke versterkers:

- ? Hij is een beetje een lul / Hij is een enorme lul
- ? Ik ben een beetje uitgeteld / Ik ben helemaal uitgeteld

6. De *I is erg N*-test. Je kunt sterke woorden definiëren door *erg* te zetten voor een neutraal woord:

- Een 'lul' is een *erg vervelende* man
- Een 'genie' is een *erg intelligent* mens
- Een 'barbaar' is een *erg ongemanierd* mens.
- 'Uitgeteld' is *erg moe*.
- 'Fenomenaal' is *erg goed*.
- 'Heerlijk is *erg lekker.*

3 Hoe zijn de intensiveerders verzameld?

Er zijn eerst drie lijsten doorgenomen.

1. De RBN-lijsten met nomina, adjectieven en werkwoorden zijn doorgenomen op sterke woorden.
2. De site onderwoorden.nl bevat een Woordenboek van Nederlandse Intensiveringen, waarvan dankbaar gebruik is gemaakt. Overigens zijn niet alle items uit dit woordenboek overgenomen.

    a. Het woordenboek bevat veel uitdrukkingen (bv. 'zoeken naar een speld in een hooiberg'), die terzijde geschoven zijn omdat we er niet zeker van zijn of die betrouwbaar opgespoord kunnen worden.
    b. Het woordenboek bevat ook veel specifieke versterkers, die achterwege zijn gelaten. Zo wordt bijvoorbeeld *als kabeltouw* genoemd als mogelijke versterker van *dik*.
3. Sterke bijwoorden zijn gekozen uit een verzameling van bijwoorden uit een SoNaR-frequentielijst.
4. Voor de zo verzamelde woorden is bekeken of ze afleidingen hebben die ook versterkend zijn:
    a. Voor sterke nomina die afgeleid zijn van een werkwoord, werden de ermee corresponderende werkwoorden en adjectieven in overweging genomen (*overheersing > overheersen, overheersend*);
    b. hetzelfde geldt voor sterke adjectieven (*betoverend > betoveren*)
    c. en voor sterke verba (*verpletteren > verpletteren, verplettering*).

    Deze procedure leidt overigens zeker niet altijd tot nieuwe intensiveerders. Waar *brutaliteit* een sterk woord is, geldt dat voor *brutaal* minder.

Vervolgens zijn in de loop der jaren specifieke intensiveerders toegevoegd die we in bepaalde teksten tegenkwamen.

4 Een indruk van de sterke adjectieven

4.1 Adjectieven met voorvoegsels

Bij de adjectieven zijn voorvoegsels belangrijker dan bij de nomina en de werkwoorden. Ruim 670 van de 11 adjectieven worden voorafgegaan door voorvoegsels. Daarbij valt op dat veel voorvoegsels slechts 1 of 2 adjectieven kunnen modificeren, zoals *aal-* en *druip-*.

Voor-voegsel	**Adjectief**
Aal-	Glad
Aarde-	Donker
Aarts-	Lui
Al	Machtig
Alom	Tegenwoordig
Aller	Liefst
Alles	Bepalend
Ape	Trots
As	Grauw
Beeld	Schoon
Bere	Goed
Bloed	Mooi
Boven	Matig
Brem	Zout
Brood	Nodig
Buiten	Gewoon
Diep	Treurig
Dol	Blij
Dood	Kalm
Door	Dringend
Drijf	Nat
Druip	Nat
Duimen	Dik
Ellen	Lang
Foei	Lelijk
Giga	Groot
Git	Zwart
Glas	Helder
Gort	Droog
Goud	Eerlijk
Haar	Fijn
Hart	grondig
Hemels	Breed
Honds	Moe
Hoog	Lopend
Hoogst	Persoonlijk
Huizen	Hoog
Hyper	Modern
IJs	Koud
IJzer	Sterk
In	Goed
Inkt	Zwart
Kei	Leuk
Kern	Gezond
Kip	Lekker
Klaar	Wakker
Kledder	Nat
Klets	Nat
Klink	Klaar
Knetter	Gek
Knoeper	Hard
Knots	Gek
Kots	Beu
Kraak	Helder
Kurk	Droog
Ladder	Zat
Lang	Verwacht
Lelie	Blank
Levens	Gevaarlijk
Lijk	Bleek
Lijn	Recht
Loep	Zuiver
Lood	Zwaar
Mega	Leuk
Mes	Scherp
Mijlen	Ver
Modder	Vet
Mud	Vol
Muis	Stil
Nagel	Nieuw
Oer	Degelijk
Olie	Dom
Over	Bezorgd
Piek	Fijn
Piemel	Naakt
Piep	Jong
Pijl	Snel
Pik	Donker
Pimpel	Paars
Pis	Link
Poedel	Naakt
Poep	Duur
Pot	Dicht
Prins	Heerlijk
Punt	Gaaf
Ras	Echt
Razend	Snel
Regel	Recht
Rete	Goed
Reuze	Gezellig
Roet	Zwart
Rood	Gloeiend
Rots	Vast
Schuw	Lelijk
Spijker	Hard
splinter	Nieuw
Spin	Nijdig
Spot	Goedkoop
Spuug	Lelijk
Stamp	Vol
Stapel	Gek
Steen	Goed
Stervens	Druk
Stik	Chagrijnig
Stok	Oud
Stom	Vervelend

Straat	Arm
Stront	Eigenwijs
Super	Lekker
Tjok	Vol
Toeter	Zat
Tonnetje	Rond
Toren	Hoog
Veder	Licht
Veel	Voorkomend
Vet	Cool
Vliegens	Vlug
Vlijm	Scherp
Vogel	Vrij
Water	Vlug
Wel	Gemeend
Wereld	Schokkend
Wijd	Verspreid
Wit	Heet
Wonder	Mooi
Ziels	Gelukkig
Zijp	Nat
Zonne	Klaar
Zuur	Verdiend
Zwaar	Bewaakt

Naast voorvoegsels, die aan woorden kunnen worden toegevoegd, zijn er ook prefixen en voorzetsels die veel voorkomen in sterke woorden.

Prefix /voorzetsel	**Woorden met dit voorvoegsel**	**Toelichting**
On-	Onverzoenlijk, onbespreekbaar	Dat ten aanzien van een bepaalde situatie of object een bepaalde handeling onmogelijk is, impliceert een extreme eigenschap
Uit-	Uitnemend, uitputtend, uitgekakt	'Uit' geeft aan dat iets erbovenuit steekt, of volledig geconsumeerd of voltooid is
Ex-	Excellent, excessief	'Ex' is Latijn voor 'uit'
Door-	Doordringend, doorlopend, doornat	'Door' geeft temporele continuatie aan of 'penetratie' van een eigenschap
Per-	Persistent, perfect	'Per' is Latijn voor 'door'
Vol-	Volkomen, volleerd	'Vol' geeft aan dat iets maximaal van toepassing is

Ook het woordeinde kan een indicatie voor een sterk woord zijn:

Postfix	**Woorden met dit achtervoegsel**	**Toelichting**
-loos	Sprakeloos, roerloos	Dat iets totaal afwezig is, is een sterke uitspraak

4.2 Adjectieven zonder voorvoegsel

Toch zijn er ook 430 adjectieven zonder voorvoegsel als 'sterk' gelabeld. Bijvoorbeeld:

- *Aanhoudend*: net als *voortdurend* en *onafgebroken* geeft dit woord temporele continuïteit aan.
- *Aanmerkelijk*: dit woord versterkt (een verschil wordt groter als het een *aanmerkelijk verschil* is, iets wordt in hogere mate beter als het *aanmerkelijk beter* wordt).
- *Abominabel*: vergelijk 'zijn prestatie was zwak, zelfs abominabel'
- *Afschuwelijk*: afschuw is een sterke negatieve emotie

Voorbeelden van woorden die overwogen zijn maar uiteindelijk uit de lijst zijn verwijderd:

- *Benauwend*: je kunt vrij goed zeggen 'enigszins benauwend'
- *Geprononceerd*: je kunt goed zeggen 'enigszins geprononceerd'
- *Homerisch*: in de combinatie met *gelach* is *Homerisch* een versterker, maar in andere combinaties niet.
- *Meervoudig*: dit woord kent ook niet-intensiverende gebruiksgevallen.

5 Adjectieven die alleen bijwoordelijk versterken

Sommige adjectieven zijn alleen in bijwoordelijk gebruik versterkend. In bijvoeglijk (attributief of predicatief) gebruik is het woord neutraal, of heeft het minstens neutrale lezingen, zoals *aanzienlijk* en *zuiver*.

Woord	**Gebruikscontexten (*= niet versterkend)**
Aanzienlijk	Een aanzienlijke toename / *Een aanzienlijk man / Aanzienlijk toegenomen
Beslist	*Een beslist optreden /We zien beslist verbetering
Bijzonder	* Een bijzonder mens / Een bijzonder groot succes
Breed	* Een brede rivier / Een breed gedragen initiatief
Dik	* Een dikke man / We hebben dik gewonnen
Driftig	* Een driftig karakter / Hij is driftig bezig om meer invloed te krijgen
Duidelijk	* Zijn boodschap was duidelijk / Hij is duidelijk afgevallen
Flink	* Een flinke jongen / Hij is flink geraakt
Fors	* Een fors gebouwde vrouw / Dat is fors toegenomen
Gegarandeerd	* Een gegarandeerd rendement van 3% /Hij zal je gegarandeerd uitschelden
Gloeiend	* Een gloeiende sigaret / Ik ben het daar gloeiend mee eens
Knap	* Een knappe jongen / Dat is knap lastig
Lelijk	* Een lelijke jongen / Dat is lelijk misgegaan
Lustig	(geen bijvoeglijk gebruik) / Het zonnetje scheen er lustig op los
Mega	(geen bijvoeglijk gebruik) Het was mega gezellig vanmiddag (eigenlijk een spelfout, maar als dit voorkomt zal 'mega' waarschijnlijk als bijwoord gezien worden)
Nauw	* Een nauwe pantalon / Hij is nauw betrokken bij dit project
Opmerkelijk	* Een opmerkelijk voorval / Dat is opmerkelijk toegenomen
Opvallend	* Een opvallend type / Opvallend snel verbeterd
Rap	* Een rappe jongen / Hij heeft zich rap verbeterd
Roerend	* Roerende goederen / Ik ben het roerend met hem eens
Ruim	* Een ruime kamer / Je hebt het ruim gehaald
Snel	* Een snelle jongen / Dat is snel verbeterd
Stellig	* Hij maakt een stellige indruk / Dat is stellig toegenomen
Stevig	* Dat is een stevige constructie / Dat is stevig toegenomen
Über	(geen bijvoeglijk gebruik) / Het was über leuk vanmiddag (eigenlijk een spelfout)
Veel	* Te veel eten is niet goed / Het weer is veel beter nu
Vet	* Een vet stuk vlees / Het was vet leuk vanmiddag
Vierkant	* Een vierkante kamer / Daar ben ik vierkant tegen
Waarachtig	* Zijn liefde is waarachtig / Dat is waarachtig zo.
Zeer	* Mijn been doet zeer / Dat is zeer snel verbeterd
Zeker	* Hij voelt zich nog niet zeker / Hij is zeker vooruitgegaan
Zeldzaam	* Een zeldzame postzegel / Een zeldzaam slechte prestatie
Ziek	* Een zieke jongen / Ziek goed dansen
Zuiver	Dat is een zuivere penalty / * Een zuiver beeld van de feiten / Dat is zuiver plantaardig en veilig
Zwaar	* Een zware last / Hij is zwaar gefrustreerd

Een woord dat die aanvankelijk op deze lijst stond maar verwijderd is, is *substantieel*; dat woord intensiveert namelijk niet alleen in bijwoordelijk maar ook in bijvoeglijk gebruik (*een substantieel verschil*). Hetzelfde geldt voor *erg*.

Een ander geval is *schreeuwend*. Dat intensiveert juist in bijvoeglijk gebruik, niet in bijwoordelijk gebruik:

- Schreeuwend kwamen ze naar buiten (niet intensiverend)
- Een schreeuwend tekort aan leraren

Tot dusver is vrij globaal gesproken van bijwoordelijk gebruik van de intensiveerder. Daaronder vallen echter verschillende grammaticale contexten:

1. bepaling bij een attributief adjectief (hij presteert *zeer* goed)
2. bepaling bij een predicatief adjectief (zijn prestatie is *zeer* goed)
3. bepaling bij een adjectief dat bepaling is bij het werkwoord (dat is *zeer* snel verbeterd)
4. bepaling bij een bijvoeglijk gebruikt voltooid deelwoord (hij is een *erg* getroubleerde man)
5. bepaling bij een predicatief gebruikt voltooid deelwoord (de relatie is *volkomen* verstoord)
6. bepaling bij een voltooid deelwoord dat bepaling is bij het werkwoord (hij reageerde *erg* geïrriteerd)
7. bepaling bij een voltooid deelwoord in vrije positie (hij is *zeker* vooruitgegaan)
8. bepaling bij een bijvoeglijk gebruikt tegenwoordig deelwoord (hij is een *erg* meelevende man)
9. bepaling bij een predicatief gebruikt tegenwoordig deelwoord (dat is *heel* innemend van hem)
10. bepaling bij een tegenwoordig deelwoord dat een bepaling is bij het werkwoord (hij spreekt *erg* vleiend over mij)
11. bepaling bij een vervoegd werkwoord (we zien *beslist* verbetering)
12. bepaling bij een infinitief (hij wil *gegarandeerd* verbouwen)

Soms intensiveert een woord niet in alle bijwoordelijke contexten, zoals *behoorlijk*:

- Hij presteert behoorlijk goed (niet intensiverend)
- Zijn prestatie is behoorlijk goed (niet intensiverend)
- Dat is behoorlijk snel gegaan (intensiverend)
- Hij presteert behoorlijk (niet intensiverend)
- Dat is behoorlijk toegenomen (intensiverend)

Zulke woorden nemen we niet op in de lijst. Het zou te veel werk vergen om deze contexten te gaan onderscheiden.

De adjectieven die mogelijk bijwoordelijk versterken vormen een aparte klasse in de lijst (met label 'bvbw'. Om te zien of het adjectief in een bepaald gebruiksgeval al of niet bijwoordelijk versterkt, hebben we informatie uit de Alpino-parser nodig. We kijken daarbij niet naar het zinsdeel-label van het adjectief zelf (dat kan bv. een 'mod' zijn, of een 'predc') of naar het vormlabel (altijd 'ad'). We kijken naar het label *erboven* in de Alpino-boom (zie http://www.let.rug.nl/vannoord/bin/alpino). De regel is vervolgens:

Adjectieven zijn bijwoordelijk versterkend als ze hangen aan een zinsdeel met

- de vorm AP, PPART, PPRES of INF (dus aan een adjectief, voltooid deelwoord, tegenwoordig deelwoord, of infinitief);
- type SMAIN of SSUB (dus aan een vervoegd werkwoord).

We geven hieronder een paar intensiverende en niet-intensiverende voorbeelden van de adjectieven *duidelijk* en *erg* met bijbehorende Alpino-labels. Overigens zijn de Alpino-analyses soms incorrect.

Voorbeeld	**Alpino-label voor het onderstreepte woord**	**Intensiveerder?**
Hij fraudeert duidelijk.	*Een 'mod' onder een 'smain'*	*Ja*
Hij is duidelijk.	Een 'predc' onder een 'smain'	Nee
Het is duidelijk dat hij zo lang is.	Een 'predc' onder een 'smain'	Nee
Het punt is dat hij duidelijk is.	Een 'predc' onder een 'ssub'	Nee
Het punt is dat hij duidelijk fraudeert.	*Een 'mod' onder een 'ssub'*	*Ja*
Hij is een duidelijke man.	Een 'mod' onder een 'predc' met vorm 'np'	Nee
Zij heeft een duidelijke man.	Een 'mod' onder een 'obj1' met vorm 'np'	Nee
Hij geeft dat aan een duidelijke man.	Een 'mod' onder een 'obj1' met vorm 'np'	Nee
Ik gaf deze duidelijke man een boek.	Een 'mod' onder een 'obj2' met vorm 'np'	Nee
Een duidelijke man gaat over lijken.	Een 'mod' onder een 'su' met vorm 'np'	Nee
Hij is erg slim.	*Een 'mod' onder een 'predc' met vorm 'ap'*	*Ja*
Hij is duidelijk slim.	*Een 'mod' onder een 'smain'*	*Ja*
Hij fietst erg hard.	*Een 'mod' onder een 'mod' met vorm 'ap'*	*Ja*
Hij fietst duidelijk hard.	*Een 'mod' onder een 'smain'*	*Ja*
Hij is een duidelijk slimme man.	*Een 'mod' onder een 'mod' met vorm 'ap'*	*Ja*
Hij kijkt erg verstoord.	*Een 'mod' onder een 'predc' met vorm 'ppart'*	*Ja*
Hij is duidelijk verstoord.	*Een 'mod' onder een 'smain'*	*Ja*
Hij keek duidelijk verstoord.	*Een 'mod' onder een 'smain'*	*Ja*
Hij is erg enthousiamerend.	*Een 'mod' onder een 'predc' met vorm 'ppres'*	
Hij is duidelijk enthousiasmerend.	*Een 'mod' onder een 'smain'*	*Ja*
Hij spreekt duidelijk enthousiasmerend.	*Een 'mod' onder een 'smain'*	*Ja*
Er wordt duidelijk gesnoeid.	*Een 'mod' onder een 'vc' met vorm 'ppart'*	*Ja*
Hij wil duidelijk fuseren.	*Een 'mod' onder een 'vc' met vorm 'inf'*	*Ja*

6 Een indruk van de sterke nomina

6.1 Menselijke nomina

Veel van de sterke nomina hebben betrekking op mensen (zo'n 600, waarvan ongeveer 100 met voorvoegsels als *aarts-, bleek-, boos-, brokken-, dom-, door-, duivels-, dwars-, glad-, klere-, maf-, mis-, pracht-, mis-, ras-, rot-, smeer-, wereld-* en *zeik-*). Veel daarvan kunnen als scheldwoord betiteld worden, althans als woord waarin een persoon een sterk negatieve eigenschap wordt toegeschreven: *armoedzaaier* gaat verder dan *arm persoon*, *barbaar* gaat verder dan *onbeschaafd iemand,* een *blaaskaak* is in hoge mate een opschepper, een *oen* is meer dan onhandig, een *sadist* iemand met een hoge mate van leedvermaak.

Sommige nomina refereren overigens aan groepen, niet aan individuen: *gajes, geboefte*.

Scheldwoorden zijn soms afgeleid van adjectieven of werkwoorden die op zich niet sterk zijn. *Morsen* staat niet bij de sterke werkwoorden, maar *morspot* wel bij de sterke nomina. Dat komt waarschijnlijk doordat het nomen generaliseert (een constante dispositie toeschrijft), en het werkwoord dat niet doet. Zoiets geldt ook voor *flemen* en *flemer, profiteren* en *profiteur* en *slap en slappeling*.

Een woord dat het niet 'gehaald' heeft, is *betweter.* Er is geen neutrale uitdrukking te vinden waarvan dat woord een sterkere variant is. Je kunt bovendien zeggen: 'hij is een beetje een betweter'. Hetzelfde geldt voor woorden als *conformist, engerd* en *egotripper*. Een ander voorbeeld is *dilettant*. Dat heeft een iets negatievere bijklank dan amateur, maar het verschil is niet groot genoeg om van een intensiveerder te spreken.

Er zijn natuurlijk ook positieve sterke nomina die naar mensen verwijzend, maar dat zijn er een stuk minder. Voorbeelden zijn *reus, pionier* en *lieverd.*

Een bijzonder geval vormen menselijke nomina die met een bepaalde liefhebberij te maken hebben: woorden met *fanaat* erin (*filmfanaat)*.

6.2 Niet-menselijke nomina met voorvoegsels

Heel duidelijke voorbeelden zijn de 260 nomina met voorvoegsels als *buiten-, dood(s)-, giga-, heksen-, kut-, luizen-, lul-, mega-, monster-, nood-, over-, pokken-, reuze-, rot-, schijt-, super-, wan-, wereld-* en *zwijne-*. Toch zijn woorden met deze voorvoegsels niet per definitie sterk. Het voorvoegsel heeft in maten een zakelijke betekenis (*megahertz*), net zoals *wereldkampioen* een unieke referentie heeft, anders dan *wereldster*. Vergelijk ook *superster* met *superbenzine*. Daarom kunnen we niet alleen op voorvoegsels afgaan, en is een woordenlijst nodig.

6.3 Andere niet-menselijke nomina

Er zijn onder de niet-menselijke sterke nomina vinden we een aantal terugkerende thema's, vooral waar het de negatieve nomina betreft:

- Onzin (*kletskoek, gelul, gezever,* enz.)
- Gezeur (*gemekker, gewauwel, geteem*)
- Stemverheffing (*geschreeuw, gekrijs, gegil*)
- Fouten en mislukkingen (*blunder, echec, fiasco*, enz.); een *mislukking* of *fout* is echter geen intensiveerder.
- Grofheid (*brutaliteit*)
- Prutswerk (*gepruts, gekluns*)
- Ergernissen en problemen (*gedoe, trammelant, heisa*); maar niet *moeilijkheden*
- Frustratie (*chagrijn, pesthumeur*); maar niet bijvoorbeeld *teleurstelling*
- Harde klap (*dreun, (dood)smak, doodklap, oplawaai*); maar niet gewoon *klap*
- (Overmatig) enthousiasme (*passie, dweepzucht, fanatisme, hysterie*)
- Wanorde (*pandemonium, chaos*)
- Gekibbel (*gekissebis, kinnesinne*)
- Zwoegen en piekeren (*gezwoeg, getob, gemodder, gesappel*)
- Beroemdheid en succes (*glorie, topjaar, grootheid*)
- Prestaties (*hoogstandje, stunt, krachttoer*)
- Minachting (*hoon, verachting*)
- Ondergang (*ineenstorting, vernietiging, ontluistering*)
- Valse pretenties (*kapsones, fratsen*)
- Een grote verzameling van iets (*keur, scala*)
- Niet-functionerende artefacten (*kutauto, rothotel*)
- Kwaadaardige streken en uitingen (*laster, intrigant, leugenaar*)

- Grootschalig geweld (*lynchpartij, massamoord*)

7 Een indruk van de 'sterke' werkwoorden

Sterke werkwoorden geven een 'heftig' beeld van processen. Voorbeelden, samen met minder sterke werkwoorden die in bepaalde contexten naar hetzelfde proces kunnen verwijzend:

Sterk werkwoord	**Neutraal werkwoord**
Afbekken	Toespreken
Afbeulen	Laten werken
Afdruipen	Weggaan
Afgaan	Falen
Afknappen	Gefrustreerd zijn
Afkukelen	Afvallen
Afmatten	Moe maken
Afraggen	Gebruiken
Afranselen	Slaan
Afrukken	Aftrekken
Afslachten	Doden

Dit zijn allemaal werkwoorden met *af-*. Andere beginmorfemen die regelmatig voorkomen in sterke woorden volgen hieronder:

- Aan(bidden,gapen,stormen)
- Dood(ergeren,lachen)
- Door(douwen,zeven)
- Ineen(krimpen,storten)
- Los(barsten,slaan); maar in *losweken* wordt de los-toestand geleidelijk bereikt, dus *los* op zich versterkt niet.
- Neer(zijgen,smakken); maar *neerleggen* is niet sterk, dus *neer* op zich versterkt niet.
- Ont(luisteren,wrichten)
- Op(hitsen,jutten,duvelen)
- Opeen(stapelen,hopen)
- Over(weldigen,heersen)
- Overhoop(gooien,halen)
- Plat(gaan,spuiten)
- Rond(dolen,bazuinen)
- Uit(bazuinen,kotsen)
- Ver(afgoden,afschuwen)
- Voort(slepen,sukkelen)
- Vuil(bekken,spuiten)
- Weg(honen,kapen)

Naast de 240 werkwoorden met dit soort voorvoegsels vinden we ruim 400 ongelede werkwoorden als de volgende:

- Bazelen
- Bluffen
- Blindstaren
- Bonken
- Bonzen
- Brullen
- Bruuskeren
- Bunkeren

8 Bijwoorden

Heel wat bijwoorden hebben een versterkende betekenis. Het gaat regelmatig om de volgende semantische categorieën:

Afkorting	**Omschrijving**
A	Abrupte bewegingen (*halsoverkop, kriskras, rechtsomkeert*)
AFW	Woorden die de afwezigheid of vertrek van een entiteit aangeven (*ervandoor, foetsie*)
ALL	Woorden die 'alleen' betekenen (*louter*)
BO	Woorden die het bijna ontbreken van iets aanduiden (*amper, nauwelijks*), dan wel het bijna mislukken van iets (*ternauwernood*).
C	De continuïteit van processen (*aldoor*)
DIS	Markeringen van discourse-relaties met elementen van belang of graad (*bovenal, zelfs*)
G	Woorden die een intense graad van een eigenschap aangeven (*apert, mordicus*)
GR	Sterke woorden met de betekenis 'graag' (*dolgraag, grif, volgaarne*)
H	Zich herhalende processen (*achtereen, nogmaals, dikwijls, tienmaal*)
IGO	Woorden die intense graad van ontbreken van een kenmerk aangeven (*geenszins, überhaupt*)
N	Bijwoorden die extra grote noodzakelijkheid aangeven (*per se*)
O	Woorden die versterken door aan te geven dat iets openlijk gebeurt (*botweg, boudweg, ronduit*)
S	Snel beginnende of verlopende processen (*meteen, opeens, pardoes, stormenderhand, weldra*)
T	Bepalingen die aan continuïteit in de tijd refereren (*vanouds, voorgoed*)
UK	Universele kwantoren over hoeveelheden, tijden of plaatsen (*allemaal, telkens; nimmer, immer; alom, overal*)
W	Bijwoorden die grote waarschijnlijkheid aangeven (*allicht*)

Hieronder een lijst intensiverende bijwoorden, waar mogelijk van een categorie-aanduiding voorzien.

Bijwoord	**Type**
achtereen	H
achterelkaar	H
aldoor	C
allang	C
allejezus	G
allemaal	UK
allicht	W
almaar	C
alom	UK
alsmaar	C
amper	BO
andermaal	H
angstvallig	
apert	G
(tot) bloedens (toe)	G
botweg	O
boudweg	O
bovenal	DIS
breeduit	G
danig	G
deerlijk	G
dikwijls	H
duizendmaal	H
enkel	ALL
ervandoor	AFW
evenzeer	DIS
faliekant	G
foetsie	AFW
gaarne	GR
geenszins	IGO
graag	GR
grif	GR
halsoverkop	S
helemaal	UK
hoezeer	G
hogelijk	G
honderdmaal	H
honderduit	C

hoogst	G
immer	UK
integendeel	DIS
jewelste	G
kriskras	A
languit	
lichterlaaie	
louter	ALL
luidkeels	G
meteen	S
minstens	
moederziel	G
mordicus	G
naarstig	G
nauwelijks	BO
nimmer	UK
node	G
nogmaals	H
ondersteboven	
opeens	A
op-en-top	G
overal	UK
overhoop	
pal	G
pardoes	A
per se	N
plots	A
rakelings	
rechtsomkeert	A
reuze	G
rijkelijk	G
ronduit	O
sowieso	N
spoorslags	S
steeds	C
stierlijk	G
stormenderhand	S
straal	G
tekeer	

telkenmale	H
telkens	H
teloor	ON
temeer	
teniet	ON
ternauwernood	BO
tienmaal	H
tuurlijk	N
überhaupt	IGO
uitentreuren	C
uiteraard	N
uitermate	G
uiterst	G
uitsluitend	ALL
vanouds	T
verre	
verreweg	G
veruit	G
(tot) vervelens (toe)	G
voetstoots	
voorgoed	T
voorwaar	N
voorzeker	N
wederom	H
weldra	T
welletjes	
welste	G
wiedes	
zeer	G
zeerste	G
zelden	BO
zelfs	DIS
zielsveel	G
zienderogen	G
zondermeer	N

Al en *reeds* zijn niet opgenomen. Zij geven wel aan dat iets eerder dan verwacht gebeurt, maar hun kracht is te gering. Hetzelfde geldt voor *slechts*. Het bijwoord *vlak* (zoals in *vlak bij huis*) is ook niet sterk genoeg geacht.

9 Combinaties

Ten slotte bevat de lijst zo'n 190 vaste combinaties die een intense variant zijn van een enkelvoudige uitdrukking. Zo'n 50 daarvan zijn verdubbelingen van het type *geheel en al*. Andere frequente soorten combinaties zijn:

- Frases van de vorm *geen X* die 'niets' betekenen: *geen bal*, enz.
- Combinaties met *nog X-er* die een comparatief versterken: *nog beter(e)* enz.
- Combinaties beginnend met *tot* die uitputtendheid aangeven: *tot de nok, tot de tanden toe,* enz.

Meer voorbeelden volgen in de volgende tabel.

Combinatie (cursief)	**Neutrale uitdrukking**
Geheel en al	Geheel
Enkel en alleen	Alleen
Geen bal	Niets
Nog beter	Beter
Tot tranen toe geroerd	Geroerd
Voor geen cent	Niet
Ad libitum	Naar keuze
Als de beste	Goed
Bij bosjes	Veel
Brede grijns	Grijns
Dikke kans	Kans
Dolle pret	Pret
In ieder opzicht geslaagd	Geslaagd
Machtig mooi	Mooi
Meer dan ooit	Meer
Nergens voor nodig	Onnodig
Nogal wat	Wat
Stom geluk	Geluk
Stinkende best	Best
Ten enenmale onjuist	Onjuist
Volle kracht vooruit	Vooruit

Combinaties worden op string gezocht, dus "stinkende best" wordt alleen in deze vorm gezocht en aangekruist.

10 T-Scankenmerken wat betreft intensiveerders

Hieronder wordt verwezen naar de lijst 'intensiveringen.xlsx'.

Int_d	Dichtheid van alle intensiveerders uit de lijst bij elkaar
Int_bvnw_d	Dichtheid van de intensiveerders die in kolom B van de lijst 'bvnw' hebben
Int_bvbw_d	Dichtheid van de intensiveerders die in kolom B 'bvbw' hebben, mits zij in de Alpino-boom direct hangen onder een zinsdeel met ofwel: - de vorm AP, PPART, PPRES of INF (dus aan een adjectief of een niet-vervoegd werkwoord), ofwel - type SMAIN of SSUB (dus aan een vervoegd werkwoord).
Int_bw_d	Dichtheid van de intensiveerders die in kolom B 'bw' hebben
Int_combi_d	Dichtheid van de intensiveerders die in kolom B 'combi' hebben
Int_nw_d	Dichtheid van de intensiveerders die in kolom B 'nw' hebben
Int_tuss_d	Dichtheid van de intensiveerders die in kolom B 'tuss' hebben
Int_ww_d	Dichtheid van de intensiveerders die in kolom B 'ww' hebben

Bijlage J. Algemene nomina in T-Scan

1 Hoe zijn algemene nomina geïdentificeerd?

De algemene nomina, waarvan we er op dit moment een kleine 1200 hebben geïnventariseerd, vormen een deelverzameling van de abstracte nomina (dynamisch en niet-dynamisch). Als startpunt bij het identificeren deze nomina nemen we de omschrijving die Flowerdew & Forest (2015, 1) geven van 'signalling nouns':

> Signalling nouns (…) abstract nouns which are non-specific in their meaning when considered in isolation and which are made specific in their meaning by reference to their linguistic context.

Voorbeelden van dat soort nomina in het Nederlands zijn *idee, methode, resultaat, probleem* en *consensus*. Kenmerken van die woorden is dus dat ze vaak in de context worden gespecificeerd (Pander Maat 2002 sprak daarom van 'invulelementen'). In die specificatie wordt dus duidelijk om *welk* idee, resultaat enz. het gaat. Flowerdew en Forest (2015) presenteren een corpusstudie over 'signalling nouns', waarin zij zich beperken tot gevallen waarin de context daadwerkelijk zo'n specificatie levert. Hun interesse in deze woorden richt zich dan ook vooral op hun tekststructurerende rol.

In T-Scan vinden wij deze woorden primair om een andere reden interessant. De bedoeling van het tellen van deze nomina is dat veel abstracte woorden weliswaar abstract zijn, maar wel gebonden aan bepaalde inhouden, terwijl een kleinere groep abstract woorden vrijwel niet gebonden is aan het tekstthema, en dus overal kan voorkomen. Deze woorden een indruk geven van de mate waarin de tekst een *algemeen, niet-domeingebonden vocabulaire* hanteert, waarin de nadruk ligt op de analyse van het thema. Vandaar dat wij niet spreken van 'signalling nouns' maar van algemene nomina (AN).

Voortbouwend op Flowerdew & Forest (2015, 13ff.) bespreken we zinsframes waarin algemene nomina zich voegen. De algemene gedachte achter hun benadering is naar onze mening een zinvolle: zij verbinden de signalling nouns aan de categorieën die in Halliday's functionele grammatica gebruikt worden om 'logicosemantic' relaties tussen bijzinnen beschrijven. Aan de ene kant zijn dat de Expansion-relations (Elaboration, Extension, Enhancement), aan de andere kant de Projection-relaties tussen matrixzinnen en het idee (Idea) dan wel de uiting (Locution) die daarmee geïntroduceerd wordt (Flowerdew & Forest, 43-44). Dit perspectief leidt ertoe dat onze nomina in principe te gebruiken zijn als een korte aanduiding van een of meer non-finiete of finiete predicaties.

Dat blijkt ook uit de eerste twee zinsframes waarin ze gebruikt kunnen worden.

1. *AN + complementzin*, waarbij verschillende soorten complementen denkbaar zijn:
    a. AN + (zijn +) *dat*-complement: het probleem (is) dat hij lui is [477 woorden]. Deze De *dat*-complementen hierboven moeten niet worden verward met de bijzinnen die temporele termen als *moment* modificeren (*het moment dat ik hem zag*). Dat soort bijzinnen zijn geen specificaties; *dat* kan hier worden vervangen door *waarop*.
    b. AN + (zijn +) *of*-of vraagwoordcomplement: de vraag (is) of het nog goed komt, de kwesties is hoeveel het kost [9 woorden]
    c. AN + (zijn +) *(om) te*-complement: het doel (is) (om) de stad te ontzetten [83 woorden]
    d. Andere 'identifying clauses' waarin AN wordt gelijkgesteld met een propositie, niet met *zijn* maar met werkwoorden zoals *erin liggen (dat), inhouden (dat), erop neerkomen (dat),* e.d. [8 woorden]

2. *AN + nominalisatie*, een naamwoord dat in compacte vorm naar predicatie verwijst; het naamwoord speelt hier de rol van de complementzin uit variant 1 hierboven. Daarbij kan het zowel gaan om een werkwoordsnominalisatie (*vertrek*) als om een adjectiefnominalisatie (*afwezigheid*). [90 woorden]
    a. AN + zijn + nominalisatie: Het probleem is zijn vertrek / afwezigheid
    b. AN + andere identifying verbs + nominalisatie: het probleem ligt in zijn vertrek / afwezigheid

c. AN + van + nominalisatie: het probleem van zijn vertrek / afwezigheid; de reconstructie van de gebeurtenissen; de escalatie van de vijandelijkheden
d. AN + op + nominalisatie: de gerichtheid op het oplossen van het probleem
e. AN + tot + nominalisatie: een aanzet tot herstel

3. Een frame dat bij Flowerdew & Forest ontbreekt, is *AN + is / gaat / als volgt:* … In dit frame wordt wederom duidelijk dat het AN kan staan voor een propositie, of zelfs voor een reeks proposities: *het dilemma is als volgt: ik wil hem niet kritiseren en als medestander verliezen, maar ik kan ook niet alles van hem accepteren.* Bij AN als *verschil, contrast, onderscheid* ligt een meerdelige voortzetting voor de hand waarin beide polen van de vergelijking worden uitgewerkt; maar een meerdelige voortzetting is evenzeer aannemelijk als het gaat om een *theorie, visie, implementatie, geschiedenis, aanpak, casus* e.d. [43 woorden]

4. Een ander frame dat we toevoegen aan Flowerdew & Forest is *AN + voorzetsel + AN* [61 woorden]. Woorden als *afwijzing* en *stelligheid* kunnen niet gespecificeerd worden door een complementzin of nominalisatie, maar verwijzen naar activiteiten betrekking hebbend op, uitingen en cognities, of kenmerken daarvan. Zij combineren daarom bij voorkeur met andere algemene woorden (die op grond van andere frames identificeerbaar zijn). Omdat die woorden naar een predicatie verwijzen, handhaven we op deze manier de generalisatie dat algemene nomina gespecificeerd kunnen worden door verwijzingen naar predicaties.
   a. AN + van + AN: de afwijzing van zijn standpunt, de relevantie van de kwestie, de complexiteit van het probleem, de verwezenlijking van zijn doel, de falsificatie van zijn claim, de stelligheid van zijn uiting, de oorzakelijkheid van het verband, de bekrachtiging van het vonnis
   b. AN + voor + AN: de aansprakelijkheid voor zijn uitspraken, de aandacht voor het thema, zijn ontvankelijkheid voor beïnvloeding
   c. AN + aan + AN: de bijdrage aan de bestrijding van kanker

Tot zover zijn algemene nomina eraan te herkennen te herkennen dat zij verwijzen naar een of meer non-finiete of finiete predicaties, dan wel daar kenmerken van geven. In beide gevallen wordt de situatie, de handeling of het proces in kwestie niet ontvouwen (dat kan niet in een enkel woord, daarvoor is immers de predicatie nodig) maar in meer algemene zin getypeerd. En het is daarin dat het algemene van deze algemene nomina gelegen is.
Maar daarmee is de klasse nog niet afdoende omschreven. Kijken we in de woordenlijsten in de bijlagen van Flowerdew & Forest, dan valt op dat daarin ook woorden voorkomen als *process, form, part, variation, category* en *data;* en zelfs *period, activism, competition, stage, time, area* en *place.* Deze termen passen niet goed in de vier frames die hierboven zijn behandeld.
Nu is dat voor Flowerdew & Forest geen groot probleem; zij vinden een limitatieve definitie op basis van zinsframes onnodig, omdat zij zich open willen stellen voor alle termen die in hun corpus gepaard gaan met een of andere vorm van specificatie. Onze benadering is een andere. We menen wel dat er plaats is voor andere algemene termen, maar wensen wel een criterium op basis waarvan deze kunnen worden geïdentificeerd.
We zullen daarom een aantal extra frames behandelen die naar onze mening toegevoegd moeten worden aan de *predicatie*-refererende algemene nomina.

5. Het eerste frame waarmee dat soort woorden gevonden kan worden is *AN + wat betreft + X,* waarbij X staat voor het thema van overdenking of bespreking. Woorden die in dit frame passen zijn *afwegingsproces, beschouwing, denken, desoriëntatie, essay,* enzovoort. Omdat thema's cognitieve entiteiten zijn, lijkt deze klasse van algemene nomina semantisch verwant aan de predicatie-verwijzende woorden. Maar terwijl die woorden een gedachte specificeren, beperken de *wat betreft*-woorden zich tot een thema-aanduiding, of kiezen ze in plaats van een productperspectief eerder een procesperspectief op het denken of spreken [115 woorden].

Toch zijn er nog meer algemene nomina. Hoe kunnen we bijvoorbeeld rechtdoen aan de intuïtie dat *aanpassing, beschikbaarheid, component, mate, optimum, steekproef, soelaas, substituut* en *totstandkoming* algemene nomina zijn?

6. Wat deze woorden gemeenschappelijk hebben, is dat zij via een voorzetsel kunnen worden verbonden met naamwoorden uit heel verschillende klassen, zoals een blik op Google leert:

- aanpassing van: *het beleid / de spelregels / uw woning / de tarieven*;
- beschikbaarheid van: *water / digitale tv / groen in de stad / studio's*;
- component van: *de opleiding / de NAVO-flitsmacht / intimiteit / fijn stof*;
- mate van: arbeidsongeschiktheid / gehoorverlies / fijnheid
- optimum van: *het lichaamsgewicht / kosten en baten / verwijderingrendement*;
- steekproef uit: *een populatie / het ledenbestand / het handelsregister / doodsoorzaakverklaringen*;
- soelaas voor: *het bedrijfsleven / hoger opgeleide met handicap / woningmarkt / parkeerprobleem*;
- substituut voor: *tabak / acquisitie / erotiek / rundvlees / religie*;
- totstandkoming van: *voorstel / richtlijn / onderzoeksagenda / bekostiging*.

Er zijn 250 woorden die op deze manier een voorzetselgroep met een enkel argument toelaten. Daarnaast is er een groep van 42 woorden kan worden gevolgd door voorzetselgroepen met meerdere argumenten, zoals *combinatie (tussen A en B), opsomming (van A en B), verenigbaarheid (van A en B)* en *synergie (tussen/van A en B)*.

Terwijl de woorden uit het zesde frame omschreven kunnen worden als aanduidingen van *variabelen* waarop een variëteit aan nominale referenten gekarakteriseerd kan worden, gaat het bij de zevende groep om algemene concepten, waarin zeer uiteenlopende verschijnselen vallen.

7. Het gaat hier om 16 woorden die kunnen worden gespecificeerd door een bijstelling direct erachter: de *variabele V, de term T, het begrip B, de entiteit E*. Nu geldt dat specificeren middels een enkele term voor veel bijstellingen (vgl. *mijn vriend Peter*, in de betekenis 'de vriend waar het hier om gaat is Peter'). Maar terwijl een vriend altijd een mens is, kan achter een algemeen woord vrijwel alles als specificatie gelden. Iedere eigenschap kan worden beschouwd als een variabele, en vrijwel ieder woord kan worden beschouwd als een begrip, term of entiteit.

Laten we nu nog eens kijken naar een aantal woorden die bij Flowerdew & Forest als 'signalling noun' genoemd worden. Als we de Nederlandse equivalenten daarvan beschouwen, welke daarvan vallen dan onder onze criteria voor algemene nomina?

Engels	**Nederlands**	**Algemeen nomen in T-Scan?**
process	*proces*	Ja: combineert via voorzetsel met AN's, bv. *proces van transformatie.*
form	*vorm*	Nee: in sommige contexten laat *vorm* een complementzin toe en heeft dus een AN-lezing, maar vaak is de lezing ook simpelweg ruimtelijk.
part	*deel*	Nee: *deel* heeft veel concrete lezingen, in tegenstelling tot bijvoorbeeld *onderdeel*, dat in frame 6 past.
variation	*variatie*	Ja: past prima in frame 6.
category	*categorie*	Ja: past in frame 7 (specificatie door bijstelling).
data	*data*	Nee: *data* kan verwijzen naar meetgegevens, maar ook naar de kalender. De temporele lezing is aan de orde in meer dan 10 van de eerste 50 treffers van SoNaR. Woorden als *gegeven*, *informatie* en *bevinding* zijn wel AN.
period	*periode*	Nee: is een temporeel naamwoord en past in geen van de frames
activism	*activisme*	Nee: een abstract naamwoord, maar past in geen van de frames
competition	*competitie*	Nee: een abstract naamwoord, maar past in geen van de frames
stage	*fase*	Nee: is een temporeel naamwoord en past in geen van de frames
time	*tijd*	Nee: is een temporeel naamwoord en past in geen van de frames
area	*gebied*	Nee: heeft zowel abstracte als ruimtelijke lezingen
place	*plaats*	Nee: is een plaatswoord en past in geen van de frames

Tabel 1. Onze plaatsing van woorden die bij Flowerdew & Forest als signalling noun worden opgevoerd

We laten nu een aantal andere woorden volgen die de AN-lijst niet gehaald hebben, zodat duidelijk is dat we selectief te werk zijn gegaan.

Woord	Commentaar
aangelegenheid	Laat soms complementzinnen toe, maar staat ook vaak in de context van *dure*
aansluiting	Heeft zowel abstracte als concrete lezingen
achtergrond	Heeft zowel abstracte als concrete lezingen
antwoord	Heeft zowel abstracte als concrete lezingen, bv. in de weergave van een gesprek
basis	Heeft zowel abstracte als concrete lezingen
bewustzijn	Heeft zowel abstracte als concrete lezingen, bv. 'bij bewustzijn'
bloei	Heeft zowel abstracte als concrete lezingen
capaciteit	Heeft zowel abstracte als concrete (ruimtelijke) lezingen
congruentie	Is abstract maar niet algemeen; betreft vaak wiskundige of taalkundige zaken
democratie	Is abstract, maar niet algemeen, want beschrijft een politieke organisatievorm

Tabel 2. Woorden die niet in de lijst terecht zijn gekomen

2 Semantische klassen voor algemene nomina

Nu duidelijk is hoe we algemene nomina hebben proberen te identificeren, kunnen we kijken naar de semantische klassen die in onze verzameling te onderscheiden zijn. We presenteren hieronder een voorlopige omschrijving en indeling van die klassen en de subklassen daarbinnen.

Klasse	Nr	Subklasse	Voorbeelden
A. additie en alternatief	1.	Additie	Aaneenschakeling, clustering, surplus
	2.	Alternatief	Substituut, surrogaat
B. belang en interesse	3.	Belang	Hoofdzaak, relevantie
	4.	Belangstelling	Aandacht, fascinatie
	5.	Essentie	Clou, crux
C. beschrijving	6.	Kenmerk	Attribuut, kenmerk, facet, nuance, profiel
	7.	Kwantiteit	Hoeveelheid, omvang, zeldzaamheid
D. bestaan	8.	Bestaan	Aanwezigheid, afwezigheid, fenomeen
E. bewoording	9.	Bewoording	Betiteling, verwoording
F. concept (-systeem)	10.	Concept	Sleutelbegrip, definitie
	11.	Conceptsysteem	Denkwereld, visie, uitgangspunt
	12.	Classificatie	Type, indeling, soort
G. contrast en variatie	13.	Contrast	Anomalie, dichotomie, dualisme
	14.	Gelijkenis	Afspiegeling, analogie, parallel
	15.	Variatie	Afwisseling, diversificatie, uniformiteit
H. discussie	16.	(On)enigheid	Conflict, debat, frictie, repliek, scepsis
	17.	Bevestiging – ontkenning	Bevestiging, weerlegging
I. doel en het bereiken daarvan	18.	Realisatie	Belichaming, implementatie, praktijk
	19.	Streven	Ambitie, belofte, gerichtheid, hoofddoel
J. feitelijke juistheid	20.	(On)waarheid	Denkfout, feitelijkheid, misverstand, onfeilbaarheid
	21.	Weergave	Afschildering, nauwkeurigheid
K. gebeurtenis	22.	Gebeurtenis	Belevenis, incident
L. gedachte en standpunt	23.	Assumptie	Pretentie, vermoeden, aanname
	24.	Gedachte	Analyse, idee, overweging
	25.	Oordeel	Grief, nadeel, taxatie, voordeel

Klasse	Nr	Subklasse	Voorbeelden
	26.	Standpunt	Advies, commentaar, houding, stelling
	27.	Verwachting	Anticipatie, verrassing
M. gradatie	28.	Gradatie	Betrekkelijkheid, hevigheid, mate, sterkte
N. handelingen en keuzes	29.	(Her)overweging	Aanpassing, correctie
	30.	Handeling	Aarzeling, daad, ingreep, maatregel
	31.	Taak	Aansprakelijkheid, instructie, plicht
O. Informatie	32.	Communicatie	Bekendmaking, hint
	33.	Data	Bevinding, informatie, kerngegeven
P. interpretatie	34.	Interpreteerbaarheid	Ambiguïteit, codering, diagnose, strekking
Q. kennis-verwerving	35.	Kennis	Inzicht, herontdekking, onwetendheid
	36.	Kwestie	Dilemma, onderwerp, vraagstuk
	37.	Onderzoek	Meting, monitoring, steekproef
	38.	Perceptie	Aanblik, indruk, schijn
R. middel tot doel	39.	Gebruik	Aanwending, toepassing
	40.	Methode	Behandeling, manier, procedure
S. mogelijkheid	41.	Mogelijkheid handeling	Bedrevenheid, gelegenheid, mandaat, onmacht
	42.	Mogelijkheid gebeurtenis	Dreiging, eventualiteit, kans, toeval
T. ontwikkeling en stabiliteit	43.	Begin	Aanzet, intrede
	44.	Ontwikkeling	Climax, discontinuïteit, mijlpaal, opkomst
	45.	Ontwikkeling - beter / slechter	Ineenstorting, progressie, vervolmaking, verwording
	46.	Ontwikkeling – kwantiteit	Reductie, toename
	47.	Stabiliteit	Behoud, destabilisatie, handhaving
	48.	Verandering	Gedaanteverandering, keerpunt, transitie
U. probleem – oplossing	49.	Oplossing	Buitenkans, soelaas, uitweg
	50.	Probleem	Barrière, complicatie, impasse
V. redeneren en causaliteit	51.	Argumentatie	Betoog, bezwaar, evidentie, slotsom
	52.	Causaliteit	Aandrang, aanleiding, afhankelijkheid
	53.	Generaliseerbaarheid	Algemeenheid, casus, tegenvoorbeeld
	54.	Inperking	Beperking, restrictie, uitzondering
	55.	Teken	Aanwijzing, signaal
W. toestand	56.	Toestand	Omstandigheid, toestand, situatie
X. structuur	57.	Onderdeel	Bestanddeel, item, component
	58.	Rangorde	Selectie, volgorde, voorrang
	59.	Verband	Compatibiliteit, configuratie, ordening
Y. wenselijk-heid	60.	Eis – ideaal	Criterium, schrikbeeld, schending
	61.	Regel	Basisprincipe, beleid, inbreuk, traditie

Tabel 3. Groepen en subgroepen van algemene nomina

We onderscheiden dus 25 klassen, die in Tabel 3 alfabetisch gerangschikt zijn. Deze klassen overziend kunnen we stellen dat de nomina verwijzen naar over *algemene kenmerken van denken en spreken*. De 25 klassen vallen verder onder te verdelen in twee hoofdgroepen: enerzijds woorden die verwijzen naar kenmerken van *afzonderlijke* situaties, attitudes en uitingen, anderzijds en woorden die verwijzen naar *relaties* tussen deze eenheden, zie Tabel 4.

Denken en spreken over afzonderlijke situaties	**Denken en spreken over relaties tussen situaties**
Opvattingen, kennis en de vorming daarvan • belang en interesse • concept(systeem) • feitelijke juistheid • gedachte en standpunt • informatie • interpretatie • kennisverwerving • mogelijkheid • wenselijkheid *Kenmerken van taaluitingen* • bewoording *Kenmerken van entiteiten en situaties* • beschrijving • bestaan • gebeurtenis • toestand *Kenmerken van kenmerken* • gradatie	*Relaties tussen situaties, attitudes of uitingen* • additie en alternatief • contrast en variatie • discussie • doel en het bereiken daarvan • handelingen en keuzes • middel tot doel • ontwikkeling en stabiliteit • probleem - oplossing • redeneren en causaliteit • structuur

Tabel 4. Voorlopige indeling van groepen algemene nomina

3 T-Scankenmerken rond algemene nomina

We concluderen dat algemene nomina een indicatie zijn van de mate waarin de tekst *expliciet analytisch* is, dat wil zeggen expliciet aandacht besteedt aan de gebruikte concepten en redeneringen.

Voorlopig creëren we veertien kenmerken op basis van onze lijst, bestaande uit zeven tweetallen van dichtheid en proportie. De kenmerken zijn gebaseerd op de groepen die vermeld zijn in de tweede kolom van de lijst algemene woorden.

1.	Alg_nw_d	Dichtheid van algemene nomina (totaal)
2.	Alg_nw_p	Proportie van algemene nomina op alle nomina
3.	Alg_nw_afz_sit_d	Dichtheid van algemene nomina rond *afzonderlijke* situaties (zie linker kolom Tabel 4)
4.	Alg_nw_afz_sit_p	Proportie van algemene nomina rond *afzonderlijke* situaties op alle nomina (linker kolom Tabel 4)
5.	Alg_nw_rel_sit_d	Dichtheid van algemene nomina rond *relaties* tussen situaties (zie rechter kolom Tabel 4)
6.	Alg_nw_rel_sit_p	Proportie van algemene nomina rond *relaties* tussen situaties op alle nomina (rechter kolom Tabel 4)
7.	Alg_nw_hand_d	Dichtheid van nomina rond menselijk *handelen*
8.	Alg_nw_hand_p	Proportie van nomina rond menselijk *handelen* op alle nomina Het gaat bij 7 en 8 om nomina uit de volgende groepen: doel en het bereiken daarvan; handelingen en keuzes; middel tot doel; probleem – oplossing
9.	Alg_nw_kenn_d	Dichtheid van nomina rond *kennis* (incl. de juistheid en verwerving daarvan)
10.	Alg_nw_kenn_p	Proportie van nomina rond *kennis* (incl. de juistheid en verwerving daarvan) Het gaat bij 9 en 10 om nomina uit de volgende groepen: concept(systeem), feitelijke juistheid, gedachte en standpunt. informatie, interpretatie, kennisverwerving, discussie, redeneren en causaliteit
11.	Alg_nw_disc_caus_d	Dichtheid van nomina rond discussie, redeneren en causaliteit
12.	Alg_nw_disc_caus_p	Proportie van nomina rond discussie, redeneren en causaliteit Het gaat bij 11 en 12 om twee groepen: discussie, en redeneren / causaliteit
13.	Alg_nw_ontw_d	Dichtheid van nomina over ontwikkeling en stabiliteit
14.	Alg_nw_ontw_p	Proportie van nomina over ontwikkeling en stabiliteit

Tabel 5. Kenmerken rond algemene nomina

Bijlage K. Algemene werkwoorden in T-Scan

1 Hoe zijn algemene werkwoorden geïdentificeerd?

De algemene werkwoorden (op dit moment een kleine 800) zijn geïdentificeerd naar het conceptuele model van de algemene nomina (zie Bijlage J). Net als daar gaat het om een deelverzameling van de abstracte woorden, die de auteur een *algemeen, niet-domeingebonden vocabulaire* te beschikking stelt, waarin de nadruk ligt op de expliciete gedachtenvorming over een thema.

Eerst bespreken we de zinsframes waarin algemene werkwoorden (AW) kunnen voorkomen, daarna geven we enkele semantische overwegingen bij de identificatie.

1. *AW + complementzin*; daarbij zijn verschillende soorten complementen denkbaar:
    a. AW + *dat*-complement als objectzin (vgl. *bevestigen, beweren, denken, ontdekken* enz.)
    b. AW + *dat*-complement als subjectzin (vgl. *verontrusten, vaststaan*)
    c. AW + voorzetsel + *dat*-complement (vgl. *ernaar streven dat, erop doelen dat, erover inlichten dat, ervan overtuigen dat*)
    d. AW + *of*-complement (*controleren of, bezien of*)
    e. AW + vraagwoordcomplement (*onderzoeken wat …, inventariseren welke …, ontrafelen hoe …., uitdenken hoe …*)
    f. AW + *(om) te*-complement als objectzin (bv. *proberen, aarzelen*)
    g. AW + *(om) te*-complement als subjectzin (bv. *baten*)

2. *AW + objectconstituent verwijzend naar uitspraken, situaties of concept.* Meer in het bijzonder kan het gaan om:
    a. een actie (*voortzetten, coördineren*)
    b. een beschrijving (*uitwerken*)
    c. een beslissing (*heroverwegen*)
    d. een boodschap of uiting (*decoderen, interpreteren, natrekken*)
    e. een concept (*omschrijven*)
    f. een gebeurtenis (*verijdelen, bagatelliseren*)
    g. een gedachte (*concipiëren, comprimeren*)
    h. een vorm van gedrag (*dulden, normeren*)
    i. een gevolg (*ontketenen, toebrengen, aanrichten*)
    j. een kwaliteit (*ontberen, symboliseren*)
    k. een kwantiteit, dat wil zeggen de waarde van een bepaalde grootheid (*uitrekenen, schatten*)
    l. een handeling (*vergemakkelijken, bemoeilijken*)
    m. een kwestie (*bespreken, opwerpen*)
    n. een ontwikkeling (*initiëren, instigeren*)
    o. een plan of streven (*realiseren, effectueren*)
    p. een probleem (*lenigen, trotseren*)
    q. een regel (*eerbiedigen, voorschrijven*
    r. een situatie (*bestendigen, destabiliseren, tenietdoen*)
    s. een standpunt of uitspraak (*bijstellen, herzien*)

3. *AW + voorzetselobject verwijzend naar situatie, propositie of concept.* Bijvoorbeeld:
    a. een actie (*vermanen om …*)
    b. een desideratum (*tegemoetkomen aan …*)
    c. een oorzaak (*wijten aan ….*)
    d. een thema (*bijpraten over …*)
    e. een regel (*indruisen tegen …*)

4. *AW + subjectconstituent verwijzend naar situatie, propositie of concept.* Bijvoorbeeld:
    a. een gewenste eigenschap (*ontbreken*)
    b. een gebeurtenis (*plaatsvinden*)

c. een kwantiteit (*toenemen*)
d. een proces in het algemeen (*ontstaan, verlopen*)
e. een terrein van menselijke activiteit (*achteruitgaan, floreren, ineenstorten*)
f. een situatie in het algemeen (*ontaarden, tegenzitten*)
g. een samenhangend systeem (*desintegreren, fragmenteren*)
h. de verenigbaarheid van of verhouding tussen verschillende overwegingen (*prevaleren, conflicteren, harmoniëren, convergeren, divergeren*)

5. *AW + als volgt:* het gaat hier om processen die een uitwerking in de vorm van een zin toelaten, dat wil zeggen dat het gaat om gedachten of ordeningen (*citeren, formuleren, vereenvoudigen, systematiseren, sorteren, rangschikken, aanduiden, typeren*).

Tot zover zijn de werkwoorden eraan te herkennen dat ze uitgewerkt kunnen worden in een of meer non-finiete of finiete predicaties, of gekoppeld kunnen worden aan constituenten die verwijzen naar uitspraken, situaties of concepten.

Toch volstaat dit criterium niet als we op zoek zijn naar werkwoorden die verwijzen naar denk- en besluitvormingsprocessen. Er zijn vier extra frames nodig.

6. Het eerste nieuwe frame is *AW + wat betreft + X,* waarbij X staat voor het thema van overdenking of bespreking. Werkwoorden die in dit frame passen zijn *bijdraaien* en *weifelen*. Omdat thema's cognitieve entiteiten zijn, lijkt deze klasse van algemene werkwoorden semantisch verwant aan de predicatie-verwijzende woorden.

Maar ook werkwoorden als *samenvoegen, cumuleren, overwaarderen, fascineren, speuren* en *bezinnen* zijn algemeen van aard. Wat deze woorden gemeenschappelijk hebben, is dat zij via een voorzetsel kunnen worden verbonden met naamwoorden uit heel verschillende klassen. Daarbij treden die naamwoorden op als object (7), voorzetselobject (8), of subject (9).

7. Object:
- *samenvoegen* kan betrekking hebben op landen of op documenten
- *ontwaren* op allerlei waarneembare entiteiten
- *overwaarderen* op allerlei beoordeelde entiteiten
8. Voorzetselobject:
- *refereren aan X*
- *functioneren als X*
- *zich bezinnen op X*
- *speuren naar X*
- *discussiëren / polemiseren/ redetwisten over X*
- *opteren voor X*
9. Subject:
- *X bestaat/ fascineert/ imponeert / gaat teloor/ resteert*

De zinsbouwframes voldoen niet in alle opzichten als het gaat om het identificeren van algemene werkwoorden, want ze leveren wel voldoende maar nog geen noodzakelijke voorwaarden. Er zijn namelijk heel wat werkwoorden verwijzend naar taalactiviteiten (*roepen, schreeuwen, vragen*) die wel een predicatief complement toelaten, maar die we niet hebben opgenomen. De reden is dat dat soort werkwoorden op geen enkele manier de aard van de gedachte aankondigt.

Een tweede groep werkwoorden die we niet opnemen is de groep die naast algemene ook niet-algemene lezingen hebben. We laten nu een aantal woorden volgen die in deze zin te ruim zijn van betekenis.

Woord	**Waarom niet opgenomen**
aanbieden	Kan een actie betreffen maar ook een ding
bestrijden	Kan zich richten tegen ideeën of verschijnselen maar ook tegen personen
doceren	Wordt vaak niet met een gedachte maar met een vak gecombineerd
doorwerken	Heeft een causale lezing maar kan ook 'doorgaan met werken' betekenen
handelen	Kan ook betrekking hebben op kopen en verkopen
misleiden	Richt zich op een persoon en niet op een gedachte
ondersteunen	Kan ook verwijzen naar fysieke processen
ontluiken	Kan anders dan *ontstaan* ook betrekking hebben op bloemen
uitvoeren	Kan ook verwijzen naar export
verguizen	Richt zich op een persoon en niet op een gedachte
verhevigen	Kan worden gecombineerd met cognitieve maar ook met fysieke verschijnselen
vooruitdenken	Wordt niet gecombineerd met een gedachte

Tabel 2. Werkwoorden die niet in de lijst terecht zijn gekomen

2 Semantische klassen voor algemene werkwoorden

We laten nu onze indeling volgen van de algemene woorden, met voorbeelden van werkwoorden.

Uit tabel 3 blijkt dat de nominaklassen grotendeels bruikbaar zijn voor het indelen van werkwoorden, met uitzondering van de klasse 'toestand' en met de kanttekening dat sommige subklassen niet gevuld worden.

Klasse	Nr	Subklasse	Voorbeelden
A. Additie en alternatief	1.	Additie	aanvullen, combineren
	2.	Alternatief	substitueren
B. Belang en interesse	3.	Belang	accentueren, veronachtzamen
	4.	Belangstelling	interesseren, fascineren
	5.	Essentie	*geen werkwoorden*
C. Beschrijving	6.	Kenmerk	karakteriseren, rubriceren
	7.	Kwantiteit	becijferen, verdisconteren
D. Bestaan	8.	Bestaan	existeren, manifesteren
E. Bewoording	9.	Bewoording	betitelen, verwoorden
F. Concept (-systeem)	10.	Concept	conceptualiseren
	11.	Conceptsysteem	modelleren, theoretiseren
G. Contrast en variatie	12.	Contrast	afwijken, dichotomiseren
	13.	Gelijkenis	gelijkstellen, uniformeren
	14.	Variatie	diversificeren, fluctueren
H. Discussie	15.	(On)enigheid	debatteren, twisten
	16.	Bevestiging – ontkenning	aanvechten, beamen
I. Doel en het bereiken daarvan	17.	Realisatie	implementeren, vervolmaken
	18.	Streven	ijveren, trachten
J. Feitelijke juistheid	19.	(On)waarheid	aandikken, fingeren, onderschatten
	20.	Weergave	belichten, weergeven
K. Gebeurtenis	21.	Gebeurtenis	geschieden, plaatsvinden
L. Gedachte en standpunt	22.	Assumptie	gissen, postuleren
	23.	Gedachte	analyseren, beschouwen
	24.	Oordeel	aanprijzen, afkeuren
	25.	Standpunt	aansporen, afraden, poneren
	26.	Verwachting	anticiperen, verwonderen
M. Gradatie	27.	Gradatie	nuanceren, relativeren
N. Handelingen en keuzes	28.	(Her)overweging	bijdraaien, herzien
	29.	Handeling	besluiten, beslissen
	30.	Taak	fungeren, functioneren
O. Informatie	31.	Communicatie	bekendmaken, instrueren
	32.	Data	*geen werkwoorden*
P. Interpretatie	33.	Interpreteerbaarheid	concretiseren, doorgronden, duiden
Q. Kennis-verwerving	34.	Kennis	onderkennen, overzien
	35.	Kwestie	aankaarten, aanroeren
	36.	Onderzoek	experimenteren, nagaan
	37.	Perceptie	bespeuren, ondervinden
R. Middel tot doel	38.	Gebruik	verbruiken
	39.	Methode	hanteren
S. Mogelijkheid	40.	Mogelijkheid handeling	bekwamen, bijbrengen
	41.	Mogelijkheid gebeurtenis	*geen werkwoorden*
T. Ontwikkeling en	42.	Begin	initiëren, inluiden

Klasse	Nr	Subklasse	Voorbeelden
stabiliteit	43.	Ontwikkeling	evolueren, ontstaan
	44.	Ontwikkeling - beter / slechter	opbloeien, stagneren
	45.	Ontwikkeling – kwantiteit	overschrijden, reduceren
	46.	Stabiliteit	bestendigen, destabiliseren
	47.	Verandering	hervormen, wijzigen
U. Probleem – oplossing	48.	Oplossing	ondervangen, tegengaan
	49.	Probleem	beletten, bemoeilijken,
V. Redeneren en causaliteit	50.	Argumentatie	aantonen, baseren
	51.	Causaliteit	afhangen, bevorderen
	52.	Generaliseerbaarheid	generaliseren, illustreren
	53.	Inperking	inperken, uitzonderen
	54.	Teken	representeren, symboliseren
W. Toestand	55.	Toestand	*geen werkwoorden*
X. Structuur	56.	Onderdeel	behelzen, omvatten
	57.	Rangorde	prefereren, prevaleren
	58.	Verband	inpassen, relateren
Y. Wenselijkheid	59.	Eis – ideaal	behoren, tegemoetkomen
	60.	Regel	inachtnemen, schenden

Tabel 3. Overzicht van groepen en subgroepen van algemene werkwoorden

Zoals eerder getoond, vallen de klassen verder onder te verdelen in twee hoofdgroepen: enerzijds woorden die verwijzen naar kenmerken van *afzonderlijke* situaties, attitudes en uitingen, anderzijds en woorden die verwijzen naar *relaties* tussen deze eenheden, zie Tabel 4.

Afzonderlijke situaties	Relaties tussen situaties
Opvattingen, kennis en de vorming daarvan • belang en interesse • concept(systeem) • feitelijke juistheid • gedachte en standpunt • informatie • interpretatie • kennisverwerving • mogelijkheid • wenselijkheid *Kenmerken van taaluitingen* • bewoording *Kenmerken van entiteiten en situaties* • beschrijving • bestaan • gebeurtenis *Kenmerken van kenmerken* • gradatie	*Relaties tussen situaties, attitudes of uitingen* • additie en alternatief • contrast en variatie • discussie • doel en het bereiken daarvan • handelingen en keuzes • middel tot doel • ontwikkeling en stabiliteit • probleem - oplossing • redeneren en causaliteit • structuur

Tabel 4. Voorlopige indeling van groepen algemene werkwoorden

3 T-Scankenmerken rond algemene werkwoorden

We concluderen dat algemene nomina en werkwoorden een indicatie zijn van de mate waarin de tekst *expliciet analytisch* is, dat wil zeggen expliciet aandacht besteedt aan de gebruikte concepten, uitspraken en redeneringen. Dat is de functie van dit algemene vocabulaire.

Voorlopig creëren we veertien kenmerken op basis van onze werkwoordenlijst, bestaande uit zeven tweetallen van dichtheid en proportie. De kenmerken zijn gebaseerd op de groepen die vermeld zijn in Tabel 4.

1.	Alg_ww_d	Dichtheid van algemene werkwoorden (totaal)
2.	Alg_ww_p	Proportie van algemene werkwoorden op alle werkwoorden
3.	Alg_ww_afz_sit_d	Dichtheid van algemene werkwoorden rond *afzonderlijke* situaties (zie linker kolom Tabel 4)
4.	Alg_ww_afz_sit_p	Proportie van algemene werkwoorden rond *afzonderlijke* situaties op alle werkwoorden (linker kolom Tabel 4)
5.	Alg_ww_rel_sit_d	Dichtheid van algemene werkwoorden rond *relaties* tussen situaties (zie rechter kolom Tabel 4)
6.	Alg_ww_rel_sit_p	Proportie van algemene werkwoorden rond *relaties* tussen situaties op alle werkwoorden (rechter kolom Tabel 4)
7.	Alg_ww_hand_d	Dichtheid van werkwoorden rond menselijk *handelen*
8.	Alg_ww_hand_p	Proportie van werkwoorden rond menselijk *handelen* op alle werkwoorden Het gaat bij dit en het voorgaande kenmerk om de volgende groepen: doel en het bereiken daarvan; handelingen en keuzes; middel tot doel; probleem – oplossing
9.	Alg_ww_kenn_d	Dichtheid van werkwoorden rond *kennis* (incl. de juistheid en verwerving daarvan)
10.	Alg_ww_kenn_p	Proportie van werkwoorden rond *kennis* (incl. de juistheid en verwerving daarvan) Het gaat bij dit en het voorgaande kenmerk om de volgende groepen: concept(systeem), feitelijke juistheid, gedachte en standpunt. informatie, interpretatie, kennisverwerving, discussie, redeneren en causaliteit
11.	Alg_ww_disc_caus_d	Dichtheid van werkwoorden rond discussie, redeneren en causaliteit
12.	Alg_ww_disc_caus_p	Proportie van werkwoorden rond discussie, redeneren en causaliteit
13.	Alg_ww_ontw_d	Dichtheid van werkwoorden over ontwikkeling en stabiliteit
14.	Alg_ww_ontw_p	Proportie van werkwoorden over ontwikkeling en stabiliteit

Tabel 5. Kenmerken rond algemene werkwoorden

Bijlage L. Kenmerken rond samenstellingen

1 Wat verstaat T-Scan onder een compositionele samenstelling?

T-Scan maakt gebruik van een lijst waarin voor ongeveer 83.000 nomina is aangegeven of het samenstellingen zijn, en zo ja wat het hoofd ervan is. Dat doen we om erachter te komen waar de beperkingen liggen van de kenmerken woordlengte en woordfrequentie. We denken dat sommige lange en infrequente woorden eenvoudiger zijn dan ze lijken, omdat het gaat om samenstellingen. Daarom hebben we ons in de annotatie van de woordenlijst beperkt tot samenstellingen waarvan de onderdelen steun bieden bij de interpretatie. Het gaat dus om samenstellingen die ook wel 'transparant' genoemd worden, dan wel compositioneel interpreteerbaar. Die term hebben we gedefinieerd aan de hand van drie eisen, waarvan de eerste preliminair van aard is.

1 De samenstelling bevat meerdere vrije morfemen
Om te beginnen hanteren we een wat striktere definitie van *samenstelling* dan die van de Algemene Nederlandse Spraakkunst (http://ans.ruhosting.nl/e-ans/12/03/02/body.html): we accepteren als zodanig alleen woorden met meerdere vrije morfemen. Dat betekent dat beide woorden los moeten kunnen voorkomen in de vorm waarin ze in de samenstelling staan. Daardoor sneuvelen woorden als *bij-vakker, baby-boom, dorst-lesser, druk-doenerij, een-akter, erf-genaam* en *midden-velder*: het tweede deel van deze woorden kan niet los voorkomen.

2 Het hoofdwoord kan in deze betekenis op zichzelf staan
Het tweede en derde criterium betreffen de compositionaliteit van de interpretatie. We gaan uit van een samenstelling 'X-Y' waarin X het bepalende en Y het bepaalde lid is (Y noemen we verder het hoofdwoord). Voor het hoofdwoord eisen we dat *in deze betekenis op zichzelf moet kunnen staan*. Dat betekent dat de samenstelling X-Y *verwijst naar een subklasse binnen de klasse* Y; met andere woorden, X-Y is 'een soort Y'. Weiskopf (2007) laat zien dat het bijzonder lastig is om de betekenisrelatie tussen X en Y buiten contexten om te analyseren, maar betoogt ook dat dit niet afdoet aan de mogelijkheid dat een samenstelling compositioneel is. Voorbeelden van woorden die niet aan deze hoofdwoord-eis voldoen:

- *Bakboord* en *stuurboord* zijn geen *boorden*.
- Evenmin zijn *voorspoed* en *tegenspoed* vormen van *spoed.*
- Evenzo is een *asbak* een bak, maar een *bullebak* niet.
- Een *aandeelhouder* is een *houder* (vgl. *kaarthouder*), maar een *aanhouder* niet.
- Iets dergelijks geldt voor *drinke-broer, stoke-brand, waag-hals, zeur-kous, zuip-schuit, elle-boog, flap-drol, gang-maker, handje-klap* en *heem-raad*.
- Minder evidente gevallen zijn *ren-baan, honden-baan* en *loop-baan*. *Baan* als vrij morfeem heeft twee heldere betekenissen (de ruimtelijke en die als 'dienstverband'); dat maakt *renbaan* en *luizenbaan* tot transparante samenstellingen. Maar *loopbaan* is dat niet, omdat het 'baan' hier in geen van beide betekenisgroepen valt.
- Vergelijkbaar is het onderscheid tussen *vernieuwingsbeweging* en *armbeweging* (dat zijn soorten bewegingen) en *voortbeweging* (dat is geen soort beweging; het kan in de hier gebruikte betekenis niet los voorkomen).
- Sterk polyseme naamwoorden als *gang* hebben allerlei combinaties (*ingang, voortgang*). Omdat in zulke reeksen het bepaalde lid voortdurend van betekenis/interpretatie verandert, zijn deze combinaties niet als samenstelling gerekend (vgl. de talrijke woorden eindigend op *–punt*).
- Ons criterium betekent ook dat veel naamwoorden die zijn afgeleid van samengestelde werkwoorden vallen buiten de boot vallen: in *droog-legging, geheim-houding, gerust-stelling, mis-leiding* kan het tweede deel kan niet los voorkomen in de betekenis die het heeft in deze combinatie.

3 Het satellietwoord vertoont een consistente betekenis
De strenge eis van los kunnen voorkomen geldt alleen voor het hoofdwoord. Voor het satellietwoord zijn we rekkelijker. Neem het startmorfeem *mis*. In *mis-daad* en *mis-handeling* kan het eerste deel niet los voorkomen in dezelfde betekenis die het heeft in deze combinaties, maar kent het wel semantische consistentie: telkens is de betekenis 'afkeurenswaardig'. Vergelijk ook *kunst-arm* en *kunst-gebit*: het gaat hier niet om de betekenis van 'kunst' als zelfstandig woord, maar om het artificiële karakter van de

referent van het hoofdwoord. Aan de eis van een consistente startmorfeem-betekenis kan ook voldaan worden door figuurlijk gebruikte beginmorfemen. Zo is *honden-baan* toegelaten, omdat het eerste deel ervan in *honden-weer* in een soortgelijke betekenis optreedt. Om dezelfde reden wordt *ergotherapie* niet als samenstelling gezien en *psychotherapie* wel (vgl. *psychoanalyse, psychotrauma* enz.).

Onze drie eisen hebben geleid tot een (bewerkelijke) handmatige annotatie van de woordenlijst. Daarbij blijven zich natuurlijk dilemma's voordoen. Het meest complexe daarvan betreft de compositionaliteit van woorden die beginnen met voorzetsels. Dat zijn vaak naamwoorden gebaseerd op samengestelde werkwoorden, zoals de werkwoorden beginnen met voorzetsels (*aanvoer, toevoer*). Het werkwoord *aanvoeren* is, zo valt te betogen, een samenstelling in die zin dat *aan* en *voeren* zelfstandig in dezelfde (dan wel soortgelijke) betekenissen kunnen voorkomen. Maar voor de nominalisaties op basis van zulke werkwoorden geldt dat niet: *voer* als naamwoord komt niet in de werkwoordelijke betekenis van *voeren* voor (vgl. *aanhouder* en *aanhouding* versus *aangroei* en *aanroep*, waar de werkwoordelijke basis wel semantisch transparant is).
Maar het probleem is fundamenteler: bij veel woorden die beginnen met voorzetsels is het de vraag of het deel *aan-* semantisch transparant en consistent is. Vergelijk *aanbetaling, aangroei, aanroep, aankomst, aankoop* en *aanmelding*. Het valt ons hier moeilijk de betekenisbijdrage van het voorvoegsel te specificeren. Daarom gelden ze niet als samenstelling. Veel woorden beginnend met *uit-* kennen dezelfde problemen. En bij *af-* zijn alleen die woorden als samenstelling gezien waarin *af-* de betekenis draagt van 'voltooid' (*afsluiting, afbouw*) of van 'weg' (*aftocht, afmars*).
Veel transparanter zijn voorzetsels met een dominant ruimtelijke en/of temporele betekenis (*voor, achter, na, binnen, buiten, boven, onder, beneden*). Soms zijn voorzetsels met abstracte betekenissen ook opmerkelijk consistent, zoals *over-* (dat erg vaak verwijst naar een onwenselijk hoge graad van iets, vergelijk *overvoeding*), of *tegen-* (waarmee vrijwel altijd een vorm van 'retourneren' aangeduid wordt, vergelijk *tegenbericht* resp. *tegenaanval*).

Een aantal andere dilemma's zijn de volgende.

- Een dilemma wat betreft het zelfstandig voorkomen van het hoofdwoord treffen we bij 'samenstellende afleidingen' als *actievoerder, bewindvoerder* en *penvoerder*. *Voerder* kan niet zelfstandig voorkomen. Toch hebben we deze woorden als samenstelling opgevat, omdat het hoofdwoord als werkwoordsnominalisatie semantisch transparant is en productief is, dus voorkomt in andere samenstellingen. Vergelijk ook *bedevaarder (koopvaarder), opdrachtgever (subsidiegever), oomzegger (ja-zegger)* en *bankzitter*. Het laatste woord laat zien dat dit soort hoofden niet productief hoeft te zijn om tot een transparant resultaat te leiden.
- Bij namen van planten en dieren is de betekenis van het satellietwoord soms ondoorgrondelijk (vgl. *lapjes-kat, vingerhoeds-kruid*). Zolang het hoofdwoord los kan voorkomen in dezelfde betekenis, accepteren we deze termen als samenstelling. Het is voor de gemiddelde lezer namelijk minder belangrijk om bij dit soort termen de precieze inhoud van het satellietwoord te kunnen reconstrueren, en het niet aannemelijk is dat veel lezers dat kunnen: deze kennis is veelal voorbehouden aan experts.
- Aan elkaar geschreven uitdrukkingen zoals *sta-in-de-weg* zijn niet als samenstellingen beschouwd. Het is namelijk moeilijk hierin een hoofd aan te wijzen.
- Landen met meerdelige namen (Opper-Guinee, Zuid-Afrika) zijn niet als samenstelling beschouwd, delen van landen (Zuid-Duitsland) wel.

Een bijzondere groep vormen woorden die worden bijeengehouden met een koppelteken.

- In zogenaamde 'copulatieve samenstellingen' (zie ans.ruhosting.nl/e-ans/12/03/02/02/02/body.html) worden twee woorden 'nevengeschikt': *patholoog-anatoom, geneesheer-directeur, Oostenrijk-Hongarije*. In deze gevallen is het tweede lid als hoofd beschouwd, hoewel het dat strikt genomen niet is.
- Een andere subgroep daarbinnen zijn de samenstellingen met een eigennaam als tweede lid: *commissie-Cohen*. Daarbij is het eerste lid als hoofd beschouwd.

2 Het aantal onderdelen van een samenstelling

In de lijst is naast het hoofdwoord en het bepalende lid ook het aantal onderdelen van de samenstelling opgenomen. Meestal zijn dat er twee, heel soms meer (in zo'n 2% van de samenstellingen). Het uitgangspunt achter onze onderverdeling is wederom dat we drie- en meerdelige samenstellingen willen identificeren die extra complex zijn qua interpretatie.
Daarom is bij het tellen van de onderdelen is weer gekeken naar de compositionaliteit. Zo heeft *hypotheekrenteaftrek* drie onderdelen, want er is sprake van *aftrek,* meer in het bijzonder van *renteaftrek*, en nog meer in het bijzonder van *hypotheekrenteaftrek*. Iets dergelijks geldt voor *vrijhandelsakkoord* (*akkoord > handelsakkoord > vrijhandelsakkoord*, waarin verbijzondering als '>' is aangegeven). Maar allerlei woorden zijn niet op deze manier driedelig compositioneel. Zo heeft *mensenrechtenactivist* heeft slechts twee onderdelen, want er is niet zoiets als een *rechtenactivist*. Met andere woorden, in het kader van *mensenrechtenactivist* functioneert *mensenrechten* als een eenheid. Iets dergelijks geldt voor *deeltijdarbeid*, *binnenhuisarchitectuur, hogedrukreiniger, mijnbouwbedrijf* en *landbouwbedrijf*. Weliswaar is *bouwbedrijf* los te gebruiken, maar een mijnbouwbedrijf is niet een soort bouwbedrijf. Daarentegen is een *scheepstimmerbedrijf* wel een soort *timmerbedrijf*, en dat is weer een soort *bedrijf*; dit wordt is dus drievoudig compositioneel.

Een lastig geval is *vrouwenvoetbal*. Het woord *voetbal* is in de betekenis van 'een soort bal' semantisch transparant; als verwijzing naar de sport als geheel is dat al niet meer ('het verwijst niet naar een sport van het type *bal*; zo'n algemene sport is er niet, althans niet gelexicaliseerd). Omdat een van de betekenissen compositioneel is, rekenen we *voetbal* als samenstelling. Maar omdat het bij *vrouwenvoetbal* gaat om een verwijzing naar de sport als geheel, en het woord in die lezing niet transparant is, gaat het hier om twee delen, niet om drie.
Vergelijk in dit kader ook *terugspeelbal* en *doorkopbal*. Dat laatste woord heeft drie delen, want het is een *bal* (in de zin van 'een actie rond een bal', zoals je kunt zeggen *mooie bal*). Een *kopbal* is een soort bal-actie; en *doorkopbal* is een soort kopbal. Maar bij *terugspeelbal* werkt dit niet, want *speelbal* bestaat wel, maar niet in de betekenis van bal-actie. In het kader van *terugspeelbal* functioneert het woordeel *terugspeel-* dus als eenheid. Daarom zijn er geen drie, maar twee delen.

Samenstellingen met meer dan twee delen worden in de lijst toch in twee delen weergegeven. Daarbij streven we naar delen die als woord zo bekend mogelijk zijn. Zo wordt *kunstroofzaak* gesplitst als kunstroof/zaak en niet als kunst/roofzaak; daarentegen wordt *kinderspeelplaats* wordt kinder/speelplaats.

3 Kenmerken rond samenstellingen

De lijst nomina met samenstellingsinformatie kent de volgende kolommen:
- het woord;
- de semantische klasse van het woord;
- een waarde voor de variabele wel / geen samenstelling (1=ja; 0=nee);
- het hoofdwoord (in een enkel geval is dat een samenstelling);
- het satellietwoord (idem);
- het satellietwoord, geschoond (dus zonder bv. –s) en gelemmatiseerd;
- het aantal delen van de samenstelling.

In Tabel 2 staan de kenmerken die we uit deze lijst destilleren op tekst-, alinea- en zinsniveau. Bij het kiezen van de 'samengestelde frequentiemaat' *Wrd_freq_log_(hfd_sat)* is gekozen voor het berekenen van de gemiddelde frequentielogaritme. Tabel 1 (laatste kolom) laat zien dat in die berekeningswijze zowel het hoofdwoord als het satellietwoord invloed hebben. De andere varianten, waarin de frequenties eerst worden opgeteld alvorens de logaritme te nemen, zijn erg gevoelig voor de frequentie van meest frequente woorddeel (vaak zal dat het hoofdwoord zijn, maar soms ook niet).

Freq hfdw	**Freq satw**	**Log_freq_ hfdw**	**Log_freq_ satw**	**Log(freq_hfdw + freq_satw)**	**Log(freq_hfdw + freq_satw)/2**	**(Log_freq_hfdw + log_freq_satw) / 2**
1000	100	3	2	Log(1100) = 3.04	Log(550) = 2.74	2.5
1000	200	3	2.30	Log(1200) = 3.08	Log(600) = 2.78	2.65
2000	100	3.30	2	Log(2100) = 3.32	Log(1050) = 3.02	2.65
2000	200	3.30	2.30	Log(2200) = 3.34	Log(1100) = 3.04	2.80

Tabel 1. Varianten voor samengestelde frequentiematen voor samenstellingen
(freq = frequentie; hfdw = hoofdwoord; satw = satellietwoord; log = logaritme met grondtal 10)

	Maat	**Omschrijving**
1.	Samenst_d	Dichtheid compositionele nominale samenstellingen
2.	Samenst_p	Proportie samenstellingen op de naamwoorden
3.	Samenst3_d	Dichtheid drie- en meerdelige samenstellingen
4.	Samenst3_p	Proportie drie- en meerdelige samenstellingen op naamwoorden
5.	Let_per_wrd_nw	Woordlengte in letters voor de naamwoorden in de tekst
6.	Let_per_wrd_nsam	Woordlengte in letters voor de niet-samenstellingen
7.	Let_per_wrd_sam	Woordlengte in letters voor de nominale samenstellingen
8.	Let_per_wrd_hfdwrd	Woordlengte in letters voor de hoofdwoorden
9.	Let_per_wrd_satwrd	Woordlengte in letters voor de satellietwoorden
10.	Let_per_wrd_nw_corr	Gecorrigeerde naamwoordlengte (voor samenstellingen geldt de hoofdwoordlengte in plaats van de woordlengte)
11.	Let_per_wrd_corr	Gecorrigeerde woordlengte (voor samenstellingen geldt de hoofdwoordlengte in plaats van de woordlengte)
12.	Wrd_freq_log_nw	Woordfrequentie (logaritme) van de naamwoorden in de tekst
13.	Wrd_freq_log_ong_nw	Woordfrequentie (logaritme) van de niet-samenstellingen
14.	Wrd_freq_log_sam_nw	Woordfrequentie (logaritme) van de nominale samenstellingen
15.	Wrd_freq_log_hfdwrd	Woordfrequentie (logaritme) van de hoofdwoorden in samenstellingen
16.	Wrd_freq_log_satwrd	Woordfrequentie (logaritme) van de satellietwoorden in samenstellingen
17.	Wrd_freq_log_(hfd_sat)	Gemiddelde van de logaritmen van de woordfrequentie van hoofdwoorden en satellietwoorden in de samenstellingen
18.	Wrd_freq_log_nw_corr	Gecorrigeerde naamwoordfrequentie (voor samenstellingen geldt de hoofdwoordfrequentie in plaats van de woordfrequentie)
19.	Wrd_freq_log_corr	Gecorrigeerde woordfrequentie (voor samenstellingen geldt de hoofdwoordfrequentie in plaats van de woordfrequentie)
20.	Freq1000_nw	Proportie naamwoorden horend bij de meest frequente 1000 woorden
21.	Freq5000_nw	Idem voor de meest frequente 5000 woorden
22.	Freq20000_nw	Idem voor de meest frequente 20000 woorden
23.	Freq1000_nsam_nw	Proportie van de niet-samenstellingen die hoort bij de meest frequente 1000 woorden
24.	Freq5000_nsam_nw	Idem voor de meest frequente 5000 woorden
25.	Freq20000_nsam_nw	Idem voor de meest frequente 20000 woorden
26.	Freq1000_sam_nw	Proportie samenstellingen horend bij de meest frequente 1000 woorden
27.	Freq5000_sam_nw	Idem voor de meest frequente 5000 woorden
28.	Freq20000_sam_nw	Idem voor de meest frequente 20000 woorden
29.	Freq1000_hfdwrd_nw	Proportie hoofdwoorden van nominale samenstellingen horend bij de meest frequente 1000 woorden
30.	Freq5000_hfdwrd_nw	Idem voor de meest frequente 5000 woorden
31.	Freq20000_hfdwrd_nw	Idem voor de meest frequente 20000 woorden
32.	Freq1000_hfdwrd_nw	Proportie satellietwoorden van nominale samenstellingen horend bij de meest frequente 1000 woorden
33.	Freq5000_hfdwrd_nw	Idem voor de meest frequente 5000 woorden
34.	Freq20000_hfdwrd_nw	Idem voor de meest frequente 20000 woorden
35.	Freq1000_nw_corr	Gecorrigeerde proportie naamwoorden horend bij de meest frequente 1000 woorden (voor samenstellingen geldt de hoofdwoordfrequentie)
36.	Freq5000_nw_corr	Idem voor de meest frequente 5000 woorden
37.	Freq20000_nw_corr	Idem voor de meest frequente 20000 woorden
38.	Freq1000_corr	Gecorrigeerde proportie woorden horend bij de meest frequente 1000 woorden (voor samenstellingen geldt de hoofdwoordfrequentie)
39.	Freq5000_corr	Idem voor de meest frequente 5000 woorden
40.	Freq20000_corr	Idem voor de meest frequente 20000 woorden

Tabel 2. Kenmerken rond samenstellingen: zins-, paragraaf- en tekstniveau

Op woordniveau geeft T-Scan de volgende samenstellingskenmerken:

	Maat	**Omschrijving**

1	Samenst	Gaat het om een samenstelling (1=ja, 0=nee)
2	Samenst_delen	Aantal delen van de samenstelling (als het geen samenstelling betreft, zijn deze en de volgende kenmerken 'NA')
3	Let_per_wrd_hfdwrd	Woordlengte in letters voor het hoofdwoord
4	Let_per_wrd_satwrd	Woordlengte in letters voor het satellietwoord
5	Wrd_freq_log_hfdwrd	Woordfrequentie (logaritme) van het hoofdwoord
6	Wrd_freq_log_satwrd	Woordfrequentie (logaritme) van het satellietwoord
7	Wrd_freq_log_(hfd_sat)	Gemiddelde van de logaritmen van de woordfrequentie van hoofdwoorden en satellietwoorden in de nominale samenstellingen
8	Freq1000_hfdwrd	Hoort het hoofdwoord bij de meest frequente 1000 woorden
9	Freq5000_hfdwrd	Hoort het hoofdwoord bij de meest frequente 5000 woorden
10	Freq20000_hfdwrd	Hoort het hoofdwoord bij de meest frequente 20000 woorden
11	Freq1000_satwrd	Hoort het satellietwoord bij de meest frequente 1000 woorden
12	Freq5000_satwrd	Hoort het satellietwoord bij de meest frequente 5000 woorden
13	Freq20000_satwrd	Hoort het satellietwoord bij de meest frequente 20000 woorden

Tabel 3. Kenmerken rond samenstellingen: woordniveau

Bijlage M. De eerste duizend woorden uit het Subtlex-corpus

Hulpww
gehad
had
hadden
heb
hebben
hebt
heeft
kan
kon
konden
kun
kunnen
kunt
mag
mocht
moest
moesten
moet
moeten
mogen
werd
werden
wil
wilde
willen
wilt
word
worden
wordt
wou
zal
zou
zouden
zul
zullen
zult

Koppelww
ben
bent
blijf
blijft
blijven
geweest
is
leek
lijken
lijkt
waren
was
zijn

Lidwoord
de
een
het

Tussen-werpsel
ach
alsjeblieft
alstublieft
hallo
he
hé
ja
nee
o
oh
sorry
welterusten

Telwoord
1
2
3
5
10
20
acht
derde
drie
eén
één
eentje
eerste
hoeveel
tien
twee
tweede
vier
vijf
zes
zeven
zoveel

Voegwoord
als
alsof
behalve
en
nadat
of
omdat
tenzij
terwijl
totdat
voordat
want
zoals
zodat
zodra

Voornaam-woord
alle
allebei
alles
daar
dat
degene
deze
die
dit
elk
elkaar
elke
ene
enkele
er
ergens
geen
haar
hem
hen
hier
hij
hun
ie
ieder
iedere
iedereen
iemand
iets
ik
je
jezelf
jij
jou
jouw
jullie
me
meer
men
mezelf
mij
mijn
mijne
minder
m'n
nergens
niemand
niets
niks
ons
onze
overal
'r
sommige
't
u
uw
veel
waarvoor
wat
we
weinig
welk
welke
wie
wij
ze
zich
zichzelf
zij
z'n
zoiets
zo'n
zulke

Voorzetsel
aan
achter
af
beneden
bij
binnen
boven
buiten
door
heen
in
langs
mee
met
na
naar
naast
om
onder
op
over
per
rond
sinds

te
tegen
tijdens
toe
tot
tussen
uit
van
vanaf
vandaan
vanuit
vanwege
via
volgens
voor
zonder

Adjectief
aardig
absoluut
ander
andere
anderen
arme
bang
bekend
belangrijk
best
beste
beter
bezig
blij
boos
dezelfde
dicht
direct
doden
dom
dronken
druk
duidelijk
echt
echte
eerder
eerlijk
eigen
eigenlijk
eindelijk
enige
erg
erger
fantastisch
fijn
fout
geboren
gek
gelijk
gelukkig
gevaarlijk
geweldig
geweldige
gewoon
goed
goede
goeie
grappig
groot
grootste
grote
half
hard
heel
heerlijk
hele
hetzelfde
idioot
jammer
jong
jonge
juist
juiste
kapot
klaar
klein
kleine
koud
kwaad
kwalijk
kwijt
laat
laatste
lang
lange
langer
later
lekker
leuk
leuke
lief
lieve
los
makkelijk
mis
moe
moeilijk
mogelijk
mooi
mooie
natuurlijk
nieuw
nieuwe
nodig
normaal
oké
onmogelijk
open
oud
oude
ouwe
perfect
prachtig
precies
prima
raak
raar
rijk
rot
rustig
schuldig
serieus
slecht
slechte
slim
snel
sneller
sterk
stil
stom
stomme
trots
vaak
vast
veilig
ver
verder
verdomde
verkeerd
verkeerde
verliefd
vol
voorbij
voorzichtig
vorige
vreemd
vreselijk
vrij
vroeg
vroeger
waar
waard
waarschijnlijk
wakker
warm
welkom
zeker
ziek
zwaar
zwarte

Bijwoord
al
alleen
allemaal
altijd
anders
bijna
daarmee
daarna
daarom
daarvoor
dan
dus
eens
eerst
eraan
erbij
erin
ermee
erop
erover
eruit
ervan
ervoor
even
geleden
genoeg
gisteravond
gisteren
graag
helemaal
hierheen
hoe
inderdaad
liever
maar
meteen
misschien
morgen
naartoe
neer
net
niet
nog
nogal
nooit
nou
nu
OK
omhoog
ongeveer
ooit
ook
opnieuw
pas
samen
slechts
soms
steeds

tenminste
ter
terug
thuis
toch
toen
trouwens
tuurlijk
vanavond
vandaag
vannacht
vooral
vooruit
waarom
wanneer
weer
weg
wel
zeer
zelf
zelfs
ziens
zo
zolang
zomaar
zover

Nomen
aarde
advocaat
afspraak
agent
antwoord
arm
auto
avond
baan
baas
baby
bal
band
bank
bed
bedrijf
beetje
begin
bewijs
bloed
boek
boot
brief
broer
buurt
contact
controle
dag
dagen
dame
dames
deel
deur
dienst
ding
dingen
dochter
doel
dokter
dollar
dood
dr.
droom
drugs
eer
eind
einde
enkel
familie
feest
film
foto
foto's
gang
gebruik
gedachten
geest
geheim
geld
geloof
geluk
generaal
gevaar
geval
gevangenis
gevoel
gezicht
gezin
god
goedemorgen
grapje
groep
grond
haast
haat
hand
handen
hart
hè
heer
hel
hemel
heren
hoezo
hoi
hond
honger
hoofd
hoop
hotel
huis
hulp
huwelijk
idee
informatie
jaar
jaren
jawel
jongen
jongens
kaart
kamer
kans
kant
kantoor
kapitein
keer
kerel
kerk
kind
kinderen
kleren
klootzak
koffie
kolonel
komaan
koning
kont
kop
kracht
land
leger
leven
lichaam
licht
liefde
liefje
lieverd
lijk
loop
lucht
lul
maand
maanden
macht
mam
mama
man
manier
mannen
meid
meisje
meisjes
meneer
mens
mensen
meter
mevrouw
miljoen
minuten
moeder
moment
mond
moord
moordenaar
muziek
naam
nacht
namen
neus
nietwaar
nieuws
nummer
ogen
oma
onderzoek
ongeluk
onzin
oom
oorlog
ouders
pa
paar
paard
pak
pap
papa
pardon
persoon
pijn
pistool
plaats
plan
plek
plezier
politie
president
prijs
probleem
problemen
punt
raad
raam
recht
reden
regel
regels
reis
relatie
respect

rest
rij
rug
ruimte
rust
schat
schatje
schip
schoenen
school
schuld
seconden
seks
sheriff
shit
situatie
sla
slaap
sleutel
soort
spel
spijt
spullen
stad
stap
stel
stem
stop
straat
stuk
stuur
succes
tafel
team
teken
telefoon
tijd
toekomst
trein
tv
uur
vader
val
vent
verdomme
verhaal
vertrouwen
vliegtuig
volk
vraag
vriend
vrienden
vriendin
vrouw
vrouwen
vuur
waarheid
wagen
wapen
wapens
water
wedstrijd
week
weken
wereld
werk
wet
woord
woorden
zaak
zak
zaken
ziekenhuis
ziel
zin
zoek
zoon
zorg
zus
zuster

Werkwoord
afgelopen
bedankt
bedoel
bedoelt
beginnen
begint
begon
begonnen
begrepen
begrijp
begrijpen
begrijpt
bel
belde
bellen
beloof
beloofd
beschermen
bestaat
betaald
betalen
betekent
bewijzen
breng
brengen
brengt
dacht
dank
dansen
deden
deed
denk
denken
denkt
doe
doen
doet
Dragen
drink
drinken
dromen
duurt
eet
eten
excuseer
ga
gaan
gaat
gaf
gebeld
gebeurd
gebeurde
gebeuren
gebeurt
gebracht
gebruiken
gebruikt
gedaan
gedacht
gedood
geef
geeft
gefeliciteerd
gegaan
gegeven
gehoord
gekomen
gekregen
geleerd
gelezen
gelooft
geloven
gemaakt
gemist
genomen
geprobeerd
geraakt
gered
gesproken
gestolen
gestuurd
getrouwd
geven
gevonden
gevraagd
gewerkt
gewonnen
geworden
gezegd
gezien
ging
gingen
gooi
haal
halen
heet
help
helpen
helpt
herinner
herinneren
hield
hoef
hoeft
hoor
hoorde
hoort
horen
hou
houd
houden
houdt
keek
ken
kende
kennen
kent
kijk
kijken
kijkt
klinkt
klopt
kom
komen
komt
kopen
kost
kreeg
krijg
krijgen
krijgt
kwam
kwamen
lachen
lag
laten
leeft
leg
leren
let
lezen
liep
liet
liggen
ligt
loopt

lopen
luister
luisteren
lukt
maak
maakt
maakte
maken
meen
missen
nam
neem
neemt
nemen
noem
noemen
ontmoet
ontmoeten
pakken
praat
praten
probeer
probeerde
probeert
proberen
raken
red
redden
regelen
rennen
rijden
schelen
schiet
schieten
schrijven
slaan
slapen
snap
speel
speelt
spelen
spreek
spreekt
spreken
sta
staan
staat
stellen
sterven
stierf
stond
stoppen
sturen
trek
trekken
trouwen
vallen
valt
vechten
veranderd
veranderen
vergeet
vergeten
verkopen
verlaten
verliezen
verloren
vermoord
vermoorden
vertel
verteld
vertelde
vertellen
vertelt
vertrekken
vertrouw
verwacht
viel
vind
vinden
vindt
vliegen
voel
voelde
voelen
voelt
volg
volgen
volgende
vond
voorstellen
vraagt
vragen
wacht
wachten
wees
weet
wegwezen
werken
werkt
werkte
weten
wilden
winnen
wist
wonen
woont
zag
zat
zeg
zeggen
zegt
zei
zeiden
zet
zetten
zie
zien
ziet
zit
zitten
zoeken
zoekt
zorgen
zweer

Bijlage N. Soorten bijzinnen zoals onderscheiden door T-Scan

Inleiding

De bijzinskenmerken zijn gedefinieerd in termen van Alpinoknopen; elke knoop heeft twee labels: een dependentielabel en een categorielabel. Als hieronder sprake is van een knop van het type *mod-rel* wil dat zeggen: dependentielabel is *mod*, categorielabel is *rel*. Het gaat om de volgende kenmerken:

1. Betr_bijzin_per_zin = het aantal knopen met categorielabel *ssub*:
    a. dat direct of indirect wordt gedomineerd door een knoop van het type *mod-rel* of *mod-whrel*.
    b. of indirect wordt gedomineerd door *mod-conj* en direct door *cnj-rel* of *cnj-whrel*.

    Dat wil in gewoon Nederlands zeggen: het aantal deelzinnen met vervoegd werkwoord hangend onder een relatieve zin of een relatieve zin met ingesloten antecedent. Meestal gaat het om één deelzin, maar er kan nevenschikking optreden.
2. Bijw_bijzin_per_zin = de som van:
    a. het aantal knopen met categorielabel *ssub* of *sv1*:
        i. dat direct of indirect wordt gedomineerd door een knoop van het type *mod-cp*;
        ii. of indirect wordt gedomineerd door *mod-conj* of *sat-conj* en direct door *cnj-cp*.

        Dat wil zeggen, het aantal bijzinnen met vervoegd werkwoord dat hangt onder een bijwoordelijke bepaling gevormd door een 'complementizer phrase'. Meestal gaat het om één deelzin, maar er kan nevenschikking optreden.
    b. het aantal knopen met categorielabel *sv1* of *cp* dat naast een knoop met dependentielabel *nucl* hangt, tenzij direct of indirect onder de *cp*-knoop nog knopen voorkomen van het type *cnj-ssub* (want dan is 2d van toepassing);
    c. het aantal knopen met *cnj-sv1* dat valt onder een knoop met dependentielabel *sat* die naast een knoop met dependentielabel *nucl* hangt;
    d. het aantal knopen met *cnj-ssub* dat direct of indirect valt onder een knoop met dependentielabel *sat* die naast een knoop met dependentielabel *nucl* hangt.

    De toevoeging onder b. is nodig om licht afwijkende bijzinnen te vatten zoals 1. *ben je moe, ga dan naar huis,* 2. *als je moe bent dan ga je naar huis* en 3. *al is hij klein, hij is sterk*. De toevoeging onder c. is nodig voor zinnen als *zie je hem niet lopen en haar niet fietsen, dan ga je naar huis*. Toevoeging d. is nodig voor zinnen als *als je hem niet ziet lopen en haar niet ziet fietsen, dan ga je naar huis.* Zie de tabel hieronder voor de Alpino-labeling van deze zinnen.
3. Compl_bijzin_per_zin = het aantal knopen met categorielabel *ssub* dat direct of indirect wordt gedomineerd door een knoop met als categorie:
    a. *whsub* (d.w.z. constituentvraag in ondergeschikte zin);
    b. *whrel* (relatieve bijzin met ingesloten antecedent) met uitzondering van *mod-whrel* gevallen (bepalingen) en gevallen van *cnj-whrel* die hangen onder een *mod-conj*; want bij deze twee typen gaat het om betrekkelijke bijzinnen;
    c. *cp* (complementizer phrase), met uitzondering van:
        i. *mod-cp* gevallen (bepalingen) en gevallen van *cnj-cp* die hangen onder een *mod-conj*; want bij deze twee typen gaat het om bijwoordelijke bijzinnen;
        ii. *sat-cp* gevallen.

    Extra eis: de knoop hangt niet direct onder een *top*-knoop.
4. Fin_bijzin_per_zin = het totaal aantal finiete bijzinnen per zin, dat wil zeggen de som van 1-3 hierboven.
5. Mv_fin_inbed_per_zin = het aantal meervoudige finiete inbeddingen per zin, dat wil zeggen het aantal bijzinsknopen dat valt onder een andere bijzinsknoop. 'Bijzinsknoop' is gedefinieerd als alle knopen van het type 1-3 hierboven.
6. Infin_compl_per_zin = het aantal infinitiefcomplementen in de zin, dat wil zeggen het aantal knopen met de categorieën *ti* (*te*-infinitief) of *oti* (*om te*-infinitief).
7. Bijzin_per_zin = het totaal aantal bijzinnen (finiet dan wel infiniet) per zin, dat wil zeggen de som van 1, 2, 3, en 6 hierboven.
8. Mv_inbed_per_zin = het totaal aantal meervoudige inbeddingen per zin, dat wil zeggen het totaal aantal bijzinsknopen dat valt onder een andere bijzinsknoop. 'Bijzinsknoop' is nu gedefinieerd als alle knopen van het type 1, 2, 3 en 6 hierboven.

9. Betr_bijzin_los = het aantal losgekoppelde betrekkelijke bijzinnen, gedefinieerd als het aantal knopen met categorielabel *rel* of *whrel* dat direct hangt onder de knoop met categorielabel *top*.
10. Bijw_compl_bijzin_los = het aantal losgekoppelde bijwoordelijke en complementsbijzinnen, gedefinieerd als het aantal knopen met categorielabel *cp* of *whsub* dat direct hangt onder de knoop met categorielabel *top*.

Voorbeelden

In de voorbeelden hieronder wordt eerst de Alpino-benoeming gegeven van de gecursiveerde bijzin. Daarbij wordt voor de slash soms de naast-hogere knoop weergegeven. In de rechterkolom volgt de T-Scancategorisatie. Het zal blijken dat enkele minder frequente bijzinstypen foutief worden ingedeeld (zie voetnoten, aangegeven met sterretjes).

Type bijzin	**Alpino-ontleding** *(legenda volgt tot slot)*	**Label T-Scan**
1. Betrekkelijke / bijvoeglijke bijzinnen		
Mijn broer, *die in Leuven woont*, is morgen jarig.	Mod-rel /body-ssub	Betr.
Alle foto's *waarop hij te zien was*, zijn verdwenen.	Mod-rel / body-ssub	Betr.
Dit is het dorp *waar hij is opgegroeid*.	Mod-rel / body-ssub	Betr.
Dit is de man *aan wie ik geld geef*.	Mod-rel / body-ssub	Betr.
Het boek *waarover ik schreef* is uitgekomen.	Mod-rel / body-ssub	Betr.
Ik kookte, *waarna ik wegging*.	Mod-whrel / body-ssub	Betr.
Hij hoestte, *wat mij teleurstelde en ergerde.*	Mod-whrel/body-conj/cnj-ssub	Betr. (2)
Hij hoestte, *wat mij teleurstelde en wat mij ergerde.*	Mod-conj /cnj-whrel /body-ssub	Betr. (2)
Ik ging weg, *wat/hetgeen een rare indruk maakte*.	Mod-whrel / body-ssub	Betr.
Hoe meer hij eet, *hoe dikker hij wordt*.	Mod-whrel / body-ssub	Betr.*
Boeken zijn dingen *die hij leest en daarna weggooit.*	Mod-rel / body-conj/cnj-ssub	Betr. (2)
Daar ligt het boek *dat hij schreef en dat goed verkoopt*.	Mod-conj /cnj-rel /body-ssub	Betr. (2)
Daar ligt het boek *wat hij schreef en wat goed verkoopt.*	Mod-rel/body-conj/cnj-ssub	
We behouden de volwaardige zorg, *waaronder verloskunde.*	Mod-whrel zonder body-ssub eronder	Geen bijzin
2. Bijwoordelijke bijzinnen		
Hij kwam *omdat ik hem gevraagd had*.	Mod-cp / body-ssub	Bijw.
Hij kwam *toen ik hem uitgenodigd had*.	Mod-cp / body-ssub	Bijw.
Hij kwam *hoewel ik hem gevraagd had weg te blijven*.	Mod-cp / body-ssub	Bijw.
Hij kwam *doordat ik hem gevraagd had*.	Mod-cp / body-ssub	Bijw.
Het liep *zoals ik voorspeld had*.	Mod-cp / body-ssub	Bijw.
Ik ga weg *omdat ik moe ben en naar bed wil.*	Mod-cp / body-conj/cnj-ssub	Bijw. (2)
Ik ga weg *omdat ik moe ben en omdat ik naar bed wil.*	Mod-conj /cnj-cp /body-ssub	Bijw. (2)
Zie je hem niet staan, ga dan meteen naar huis.	(Du) /tag-sv1 Ernaast: nucl/sv1	Bijw.
Zie je hem niet staan, dan ga je meteen naar huis.	(Du) /Sat-sv1 Ernaast: nucl-smain	Bijw.
Zie je hem niet staan dan ga je naar huis	(Du) /dp-sv1 Ernaast: dp-smain	
Als je hem niet ziet staan, ga dan meteen naar huis	(Du) /Sat-cp Ernaast: nucl-sv1	Bijw.
Als je hem niet ziet staan, dan ga je meteen naar huis	(Du) /Sat-cp Ernaast: nucl-smain	Bijw.
Als je hem niet ziet staan en als je haar niet ziet fietsen, dan ga je naar huis.	Sat-conj/cnj-cp/ssub Ernaast: nucl-smain	Bijw. (2)
Als je hem niet ziet staan en haar niet ziet lopen, dan ga je naar huis	Sat-cp/body-conj/cnj-ssub	Bijw. (2)
Ook al is hij klein, hij is beresterk	(Du) /Sat-cp	Bijw.

Type bijzin	**Alpino-ontleding** *(legenda volgt tot slot)*	**Label T-Scan**
	Ernaast: nucl-smain	
Al is de leugen nog zo snel, de waarheid achterhaalt hem wel	(Du) /Sat-cp Ernaast: nucl-smain	Bijw.
Al gaat hij op zijn kop staan, geef niet toe	(Du) /Sat-cp Ernaast: nucl-sv1	Bijw.
Ik heb alles gedaan, *inclusief het werk voor morgen.*	Mod-cp zonder body-ssub	Geen bijzin
3. Finiete complementszinnen		
a. Onderwerpszin		
Wat u doet, is onaanvaardbaar.	Su-whsub /body-ssub	Compl.
Wie niet weg is, is gezien.	Su-whrel /body-ssub	Compl.
Dat hij komt, is onaannemelijk.	Su-cp /body-ssub	Compl.
b. Lijdendvoorwerpszin		Compl.
Wie te laat komt, laten we niet meer binnen.	Ob1-whrel /body-ssub	Compl.
Ik denk *dat hij komt* .	Vc-cp /body-ssub	Compl.
Ik weet niet *wat ik moet denken.*	Vc-whsub /body-ssub	Compl.
Ik wil weten *of hij komt*.	Vc-cp /body-ssub	Compl.
c. Meewerkend-voorwerpszin		Compl.
Wie het niet begrijpt, zal ik het nog eens uitleggen.	(foutief) Vc-whsub	Compl.
Ik twijfel eraan *dat/of hij komt.*	Vc-cp /body-ssub	Compl.
Zo belangrijk is kinderopvang voor wie gaat werken	Mod-pp /ob1-whrel /body-ssub	Compl.
d. Voorzetselvoorwerpszin		Compl.
Ik twijfel aan *wat je zegt* .	Ob1-whrel /body-ssub	Compl.
De conditie van *wie hier woont* is niet goed.	Ob1-whrel /body-ssub	Compl.
e. Predicaatszin / gezegdezin		Compl.
Hij is geworden *wie hij altijd wilde zijn*.	Predc-whrel /body-ssub	Compl.
Hij is niet *wie je denkt*.	Predc-whrel /body-ssub	Compl.
Er is een bepaling *volgens welke dit mag*.	Predc-whrel / body-ssub (foutief)	Compl.**
Mijn mening is *dat hij liegt* .	Vc-cp /body-ssub	Compl.
f. Bijwoordelijke bepaling in bijzinsvorm		
Waar ik vandaan kom, houden ze juist van dat soort humor	Predc-whrel /body-ssub (foutief)	Compl.
Hij zat *waar ik hem eerder had gezien.*	Ld-whrel /body-ssub	Compl.
Hij heeft over het touw gelopen *zonder dat hij viel*.	Mod-pp/vc-cp/body-ssub	Compl.
Ondanks dat ik er niet in geloofde is het gelukt	Mod-pp/vc-cp/body-ssub	Compl.***
g. Bijvoeglijke bepaling in bijzinsvorm		
De verwachting *dat er een einde aan zou komen*, werd niet bewaarheid	(Su-np) /vc-cp /body-ssub	Compl.
De vraag *of hij zou komen*, bleef in de lucht hangen	(Su-np) /vc-cp /body-ssub	Compl.
De tijd *dat iemand zelf zijn kleren maakt*, is voorbij	Su-np /vc-cp /body-ssub	Compl.
Het bericht *als zou de koning ziek zijn*, wekte verwarring	Mod-cp /body-ssub (foutief)	Bijw.****
Ik had een gevoel *alsof ik zweefde*	Mod-cp /body-ssub (foutief)	Bijw.****
Dit is een boek *zoals ik zelf zou willen schrijven*	Predc-cp /body-ssub (foutief)	Bijw.****
Er zijn allerlei verhalen *over hoe hij won*.	(Su-np) /mod-pp /vc-whsub /body-ssub	Compl.
Er zijn rapporten *van wat hij gezegd heeft*.	(Su-np) /mod-pp /ob1-whrel /body-ssub	Compl.
h. Complement bij zo + adjectief *of* zo'n + N		
Hij is zo blind *dat hij dat niet ziet*	(Predc-ap) /mod-ap /obcomp-cp /body-ssub	Compl.

Type bijzin	**Alpino-ontleding** *(legenda volgt tot slot)*	**Label T-Scan**
Hij is zo gek *dat hij dat doet*.	(Predc-ap) /mod-ap /obcomp-cp /body-ssub	Compl.
Ik doe dat op zo'n manier *dat hij niet boos wordt*.	(Mod-pp/ob1-np/det-detp) /obcomp-cp/body-ssub	Compl.
4. Infinitiefcomplementen		
Het valt niet mee *om rustig te blijven*.	Su-oti	Inf.-compl.
Ik vroeg hem *om rustig te blijven*.	Vc-oti	Inf.-compl.
Het valt niet mee *hem te begrijpen*.	Su-ti	Inf.-compl.
Ik heb vertrouw erop *u voldoende te hebben ingelicht*	Vc-ti	Inf.-compl.
Ik heb zin *om de strijd op te geven*.	Vc-ti	Inf.-compl.
Het is een jongen *om te zoenen*.	Mod-oti	Inf.-compl.
Om kiezers te winnen ging hij de straat op	Mod-oti	Inf.-compl.
5. Meervoudige inbeddingen		
Dit is de man [1] die doet [2] *wat ik wil*	1. Mod-rel /body-ssub; *2. Vc-whsub* /body-ssub	1 fin. mv. inbedding
Dit is de man [1] die de vrouw trouwde [2] *waarmee ik werk*	1. Mod-rel /body-ssub; *2. Mod-rel* /body-ssub	1 fin. mv. inbedding
Ik zie mensen [1] die de verwachting hebben [2] *dat er een ramp komt*	1. Mod-rel /body-ssub; *2. Vc-cp* /body-ssub	1 fin. mv. inbedding
Hij kwam [1] hoewel ik gevraagd had [2] *om weg te blijven*	1. Mod-cp /body-ssub; *2. Vc-oti*	1 infin. mv. inbedding
Ik denk [1] dat je niet weet [2] *wat er gebeurd is* [3] *toen jij weg was*	1. Vc-cp /body-ssub; *2. Vc_whsub* /body-ssub; *3. Mod-cp* /body-ssub	2 fin. mv. inbedding

* Over de classificatie van dit soort zinnen valt te twisten.
** Wanneer de hoofdzin *zijn* als koppelwerkwoord bevat, wordt een betrekkelijke bijzin soms abusievelijk als predicaatscomplementzin gezien.
*** Bijzinnen met *ondanks dat* worden door Alpino gezien als bijwoordelijke bepaling bij het werkwoord, terwijl *ondanks dat* eerder als voegwoord functioneert; eigenlijk zou het dus beter zijn hier van een bijwoordelijke bijzin te spreken.
**** De Algemene Nederlandse Spraakkunst spreekt in dit soort gevallen van een verbale complementszin bij een naamwoordgroep; Alpino maakt er een *mod-cp* bij het zinshoofd. Het rechtzetten van deze fout zou de definities aanzienlijk complexer maken; gezien de lage frequentie van deze constructies zien we daarvan af.

Legenda Alpino-afkortingen voor dependentierelaties:
Body = romp bij complementizer
Cnj = lid van nevenschikking
Ld = locatief of directioneel complement
Mod = bijwoordelijke bepaling
Nucl = kernzin
Obcomp = vergelijkingscomplement
Ob1 = direct object
Pc = voorzetselvoorwerp
Predc = predicatief complement
Sat = satelliet; aan- of uitloop
Su = subject
Tag = aanhangsel, tussenvoegsel
Vc = verbaal complement

Legenda Alpino-afkortingen voor categorielabels:
Ap = bijvoeglijk-naamwoordgroep
Conj = nevenschikking
Cp = frase ingeleid voor onderschikkend voegword (complementizer phrase)

Du = discourse unit
Inf = kale infinitiefgroep
Np = naamwoordelijke constituent
Oti = *om te*-infinitief
Pp = voorzetselconstituent
Rel = relatieve zin
Smain = declaratieve zin
Ssub = bijzin (werkwoord finaal)
Sv1 = werkwoordsinitiële zin
Ti = *te*-infinitief
Whrel = relatieve zin met ingesloten antecedent
Whsub = constituentvraag in ondergeschikte zin