Handleiding T-Scan

Versie 1 mei 2015

Auteurs:

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

Andere leden van het team dat T-Scan bouwde:

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

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

INHOUD *(KLIK OP EEN TITEL OM NAAR HET DESBETREFFENDE HOOFDSTUK TE GAAN)*

1. Wat is T-Scan?

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

De eerste versie van T-Scan is ontwikkeld door Rogier Kraf en Henk Pander Maat, waarbij de code is geschreven door Rogier Kraf in Python. Met deze versie is bijvoorbeeld een heranalyse van de CLIB-leesbaarheidsdata uitgevoerd (Kraf en Pander Maat, 2009). In de periode 2009-2012 is T-Scan onderhouden door Rogier Kraf. Met behulp van 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.

Andere mensen die belangrijk zijn geweest voor T-Scan zijn Antal van den Bosch (veel software onder de motorkap van T-Scan is onder zijn leiding ontwikkeld), Maarten van Gompel (CLAM-interface), Suzanne Kleijn (testen van kenmerken) en Nick Dekker (testen van kenmerken; woordenlijsten; handleiding).

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): semantisch geannoteerde woordenlijsten. Deze lijsten zijn handmatig gecorrigeerd en deels opnieuw geannoteerd door H. Pander Maat;
- Wopr[6] (Berck & Van den Bosch, 2009): maten voor trigramprobabiliteit, entropie en perplexiteit.

De huidige versie van T-Scan is een voorlopige; er wordt nog gewerkt aan een semantische ruimte voor het Nederlands, die te zijner tijd nieuwe kenmerken voor T-Scan zal gaan opleveren. Andere uitbreidingen zullen zijn een betere herkenning van samenstellingen en het corrigeren van woordfrequenties voor samenstellingen.

T-Scan is op dit moment voor onderzoeksdoeleinden toegankelijk via de CLAM-interface (http://webservices-lst.science.ru.nl/). Aanvragen voor gebruikersrechten kunnen worden gedaan bij h.l.w.pandermaat@uu.nl.

---

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

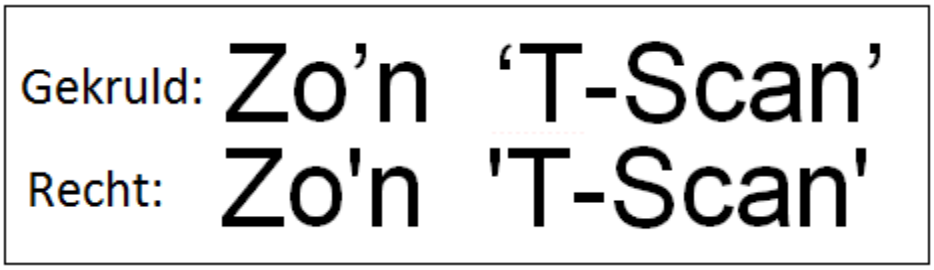
2. Werken met T-Scan

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. Gebruik onderstaande checklist om na te gaan of je teksten daaraan voldoen:

- Het bestand is gecodeerd in UTF8.
- De bestandsnaam bevat geen spaties (je kunt wel lage streepjes gebruiken: bestand_nieuwsbericht)
- Het bestand bevat geen typ- of spelfouten.
- Elke zin wordt afgesloten met een leesteken.
- Een nieuwe paragraaf wordt aangegeven door een witregel.
- Titels en kopjes zijn verwijderd.
- Speciale karakters worden juist weergegeven (aanhalingstekens, accentstreepjes, etc.). Zorg dat aanhalingstekens en apostroffen zijn gecodeerd met de gestandaardiseerde symbolen (U0022, U0027).

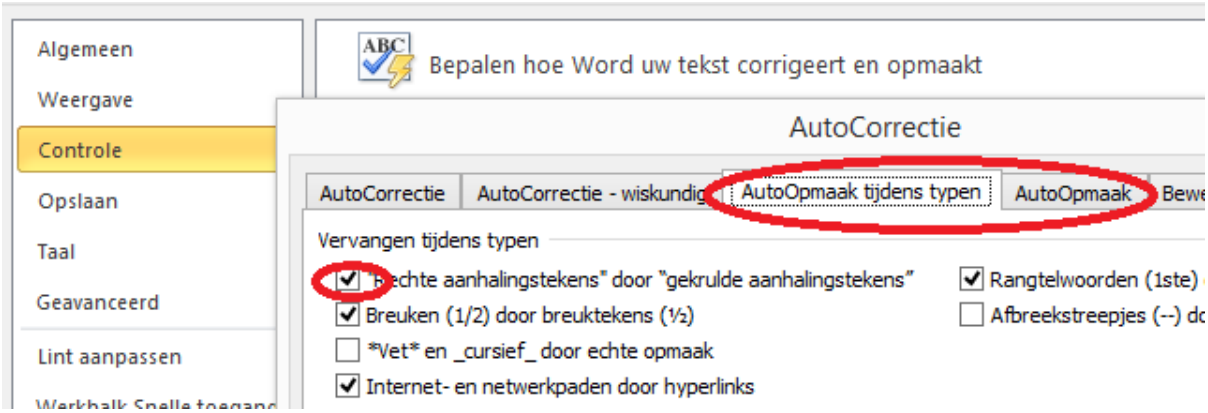
We gaan nu verder in op het laatste punt, omdat dit in de praktijk het meest weerbarstig is. Vooral met aanhalingstekens ontstaan vaak problemen, omdat Word deze automatisch verandert. Voor T-Scan moet de apostrof (bijv. in *zo'n* of *'s nachts*) en het aanhalingsteken (bijv. 'T-Scan') gelijk aan elkaar zijn. Wanneer een apostrof als aanhalingsteken gebruikt wordt, maakt Windows er echter automatisch een 'gekromd aanhalingsteken' van:



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*).



5. 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.]

6. 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.

2 Oplossing: speciale karakters vervangen in Notepad++

Zoals hierboven besproken heeft T-Scan moeite met bepaalde tekens, zoals gekromde aanhalingstekens en apostrofs. 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.

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), beschouwt T-Scan wat er op deze regel staat als commentaar. Deze optie is geschikt voor koppen die niet langer zijn dan een regel.
2. Als je een groter element wilt markeren, kan dat als volgt. Je laat de eerste regel beginnen met <<<; je sluit het commentaar af door de laatste regel te laten beginnen met >>>. Ook de rest van die regel wordt nog genegeerd.
3.

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 tot verwerkingsproblemen kunnen leiden die we nog niet hebben kunnen oplossen. Dus als je merkt dat een tekst met <<<-markeringen niet goed verwerkt wordt, is het verstandig het nog eens te proberen na deze markeringen te hebben vervangen door ###-markeringen.

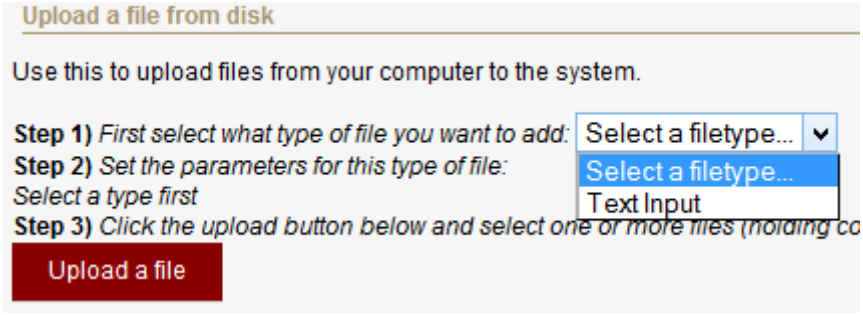
2.3 T-Scan starten en teksten invoeren

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 vanaf je harde schijf (a) of je kunt je teksten invoeren in de browser (b).

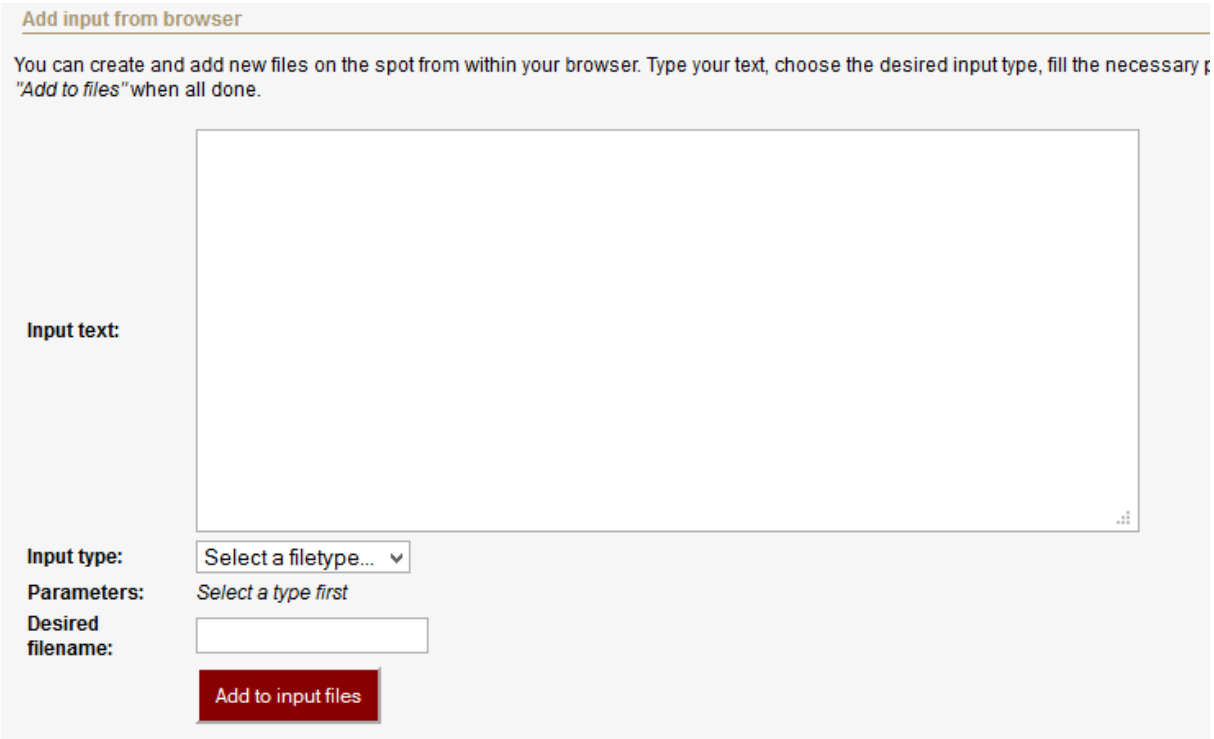
a. **Vanaf harde schijf**. Als je teksten van je harde schijf wilt uploaden gebruik je het menu onder '*Upload a file from disk'*. Vergeet niet om je teksten in UTF-8 te coderen (Zie 2.2)! Klik op het pijltje en kies voor 'Text Input'. Vervolgens klik je op 'Upload a file'. Nu kun je de bestanden kiezen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

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

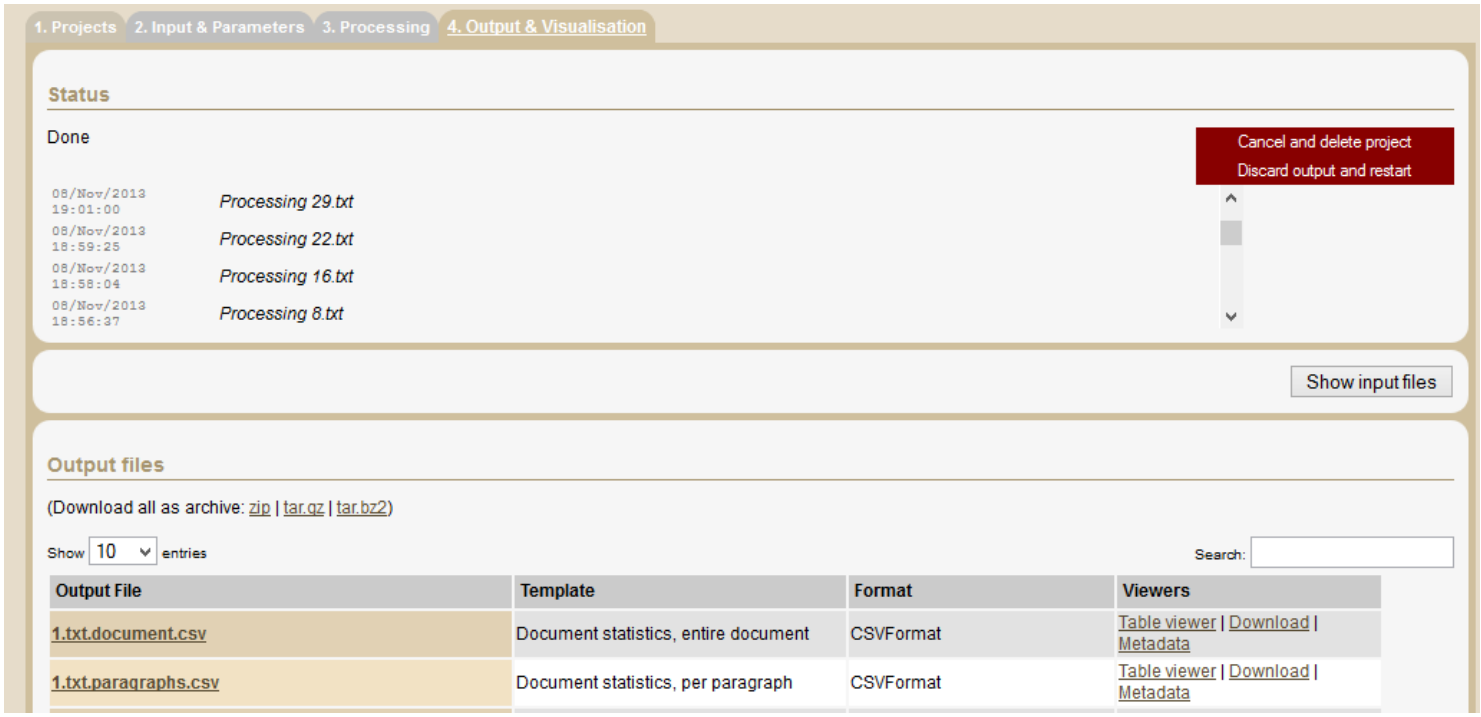


Bij 'Input type' kies je voor 'Text input'. In het grote scherm (*Input Text)* kun je teksten typen of plakken. Je kan direct vanuit Microsoft Word kopiëren, T-Scan codeert de teksten automatisch in UTF-8. Geef elke tekst en naam, en kies voor 'Add to input files'.

4. Stel nu de parameters (zeg maar: de instellingen) in voor je analyse.
    a. Rarity level; zie 3.5 bij 'zeldzaamheidsindex' voor een toelichting.
    b. Overlap size; zie 3.5 bij 'bufferoverlap voor een toelichting.
    c. Frequency clipping; het gaat hier om het inkorten van de gebruikte woordfrequentielijsten om tijd te winnen; zie verder 3.3 bij woordfrequenties.
    d. MTLD factor size; zie 3.5 bij MTLD voor een toelichting.
    e. 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.
    f. Use Wopr; het gaat hier om het al of niet gebruiken van Wopr om probabiliteitskenmerken te berekenen, zie verder 3.10. Standaard wordt Wopr gebruikt.
    g. Use LSA analyzer; omdat LSA nog niet werkt, gaan we hieraan voorbij.
    h. Word frequency list; er is een keuze tussen verschillende frequentielijsten voor woordfrequenties, zie 3.3
    i. Lemma frequency list; er is een keuze tussen verschillende frequentielijsten voor lemmafrequenties, zie 3.3
    j. Top frequency list; er is een keuze tussen verschillende frequentielijsten om de meest frequente 1000, 2000, 3000, 5000, 10000 en 20000 woorden uit te gebruiken voor het bepalen van de proportie frequente woorden, zie 3.3
5. *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 3).

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* en kies *Read text data.*
3. Zoek je CSV-file en open dit.

4. In stap 1 van de wizard antwoord je bij *predefined format?* ontkennend (*no*) en ga je naar *next.*
5. In stap 2 laat je het bolletje bij 'variable arrangement' staan op *delimited* en geef je aan dat er bovenaan de file variabelennamen staan (*yes*). Ga naar *next.*
6. In stap 3 laat je alles zoals het is: elke regel geeft een case, en je wilt al je cases invoeren.
7. In stap 4 laat je als afscheidingsteken (delimiter) alleen de komma staan. Als je op *next* klikt, vraagt SPSS of het variabelennamen mag veranderen. Je klikt op *OK*.
8. In stap 5 krijg je een indruk van hoe je file eruit gaat zien. Je kunt daar al dingen veranderen aan kolombreedte en dataformat, maar het is handiger om dat in de file zelf te doen.
9. In stap 6 krijg je de kans om de invoerprocedure weer te geven als een lijst SPSS-commando's.
    a. Doe je dat, dan kom je in een Syntax-scherm. Je kunt de commando's bewaren voor een volgende keer. Als je de commando's laat uitvoeren, krijg je de file op het scherm.
    b. Doe je dat niet, dan druk je gewoon op *Finish* en krijg je je file op het scherm.

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. Soms geven ongebruikelijke volgordes van punten en aanhalingstekens problemen bij het scheiden van zinnen.
- Kijk of alle variabelen als Numeric gedefinieerd zijn. SPSS maakt variabelenkolommen waarin letters staan tot 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.

Tip: zet de zinnen naast de getallen

In een resultatenfile op woordniveau staan de woorden automatisch vermeld in de kolom *woord*. Bij resultaten op zinsniveau is dat niet zo, maar zou je toch graag willen zien op welke zin de kenmerken betrekking hebben. Dan is het handig om de zinnen in de eerste kolom te zetten.

- In SPSS doe je dat door de kolom *inputfile* voldoende breed hebt gemaakt (doe dat door *variable view* te openen en dan bij *width* ruimte te maken voor een paar honderd tekens).
- In Excel kunnen de zinnen ook eenvoudig in de eerste kolom worden toegevoegd; dat biedt het extra voordeel dat je die kolom 'vast kunt zetten' (menu View; Freeze Panes), en er telkens nieuwe kenmerken naast zetten.

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 drie *tekstregio's*:

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

Er zijn zo'n 300 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

Op zins-, paragraaf- en tekstniveau treffen we 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	Status van de Alpino-parser

Inputfile (1a) spreekt voor zich. Bij kenmerk 1b 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 kenmerk 1b. De kenmerken 2, 3 en 4 geven het aantal alinea's, zinnen en woorden van het document.

Kenmerk 5 geeft de status van Alpino aan. Alpino is de ontleedmachine die basisinformatie levert voor een behoorlijk aantal kenmerken. Alpino kan echter ook 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 ontleed;
'1' = Alpino is er niet in geslaagd de zin te ontleden.

De laatste waarde is het meest interessant. 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 als bij zinnen waarin genummerde opsommingen voorkomen.

Wanneer Alpino_status de waarde '1' heeft, wordt de zin genegeerd bij de berekeningen van gemiddeldes op paragrafen tekstniveau. Op woord- en zinsniveau worden dan geen waarde gegeven voor kenmerken die een beroep doen op Alpino.

Op woordniveau treffen we wat extra informatie bij de algemene kenmerken, te weten het woord, het lemma daarvan, en de morfemen. Zie verder hoofdstuk 4.

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
15.	Namen_d	Dichtheid van namen
16.	Sam_delen_per_wrd	Samenstellingsdelen per woord
17.	Sam_d	Samenstellingsdichtheid
18.	Freq50_Staph	De proportie woorden die in de Staphorsius-frequentielijst 50% van de meest frequente woordtokens uitmaken
19.	Freq65_Staph	Idem maar nu gaat het om 65% van de woordtokens
20.	Freq77_Staph	Idem maar nu gaat het om 77% van de woordtokens
21.	Freq80_Staph	Idem maar nu gaat het om 80% van de woordtokens
22.	Wrd_freq_log	Woordfrequentie, logaritme
23.	Wrd_freq_zn_log	Woordfrequentie zonder namen, logaritme
24.	Lem_freq_log	Lemmafrequentie, logaritme
25.	Lem_freq_zn_log	Lemmafrequentie zonder namen, logaritme
26.	Freq1000	De proportie woorden horend bij de meest frequente 1000 woorden
27.	Freq2000	Idem voor de meest frequente 2000 woorden
28.	Freq3000	Idem voor de meest frequente 3000 woorden
29.	Freq5000	Idem voor de meest frequente 5000 woorden
30.	Freq10000	Idem voor de meest frequente 10000 woorden
31.	Freq20000	Idem voor de meest frequente 20000 woorden

Woordlengtes, namen en samenstellingen

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.

Een specifiek type van langere woorden is de samenstelling. T-Scan geeft op dit moment op basis van een lijst samenstellingen aan hoeveel woorden een samenstelling zijn en is en hoeveel samenstellingsdelen de tekstwoorden gemiddeld tellen. Op woordniveau is het aantal samenstellingsdelen meestal 2 (bv. 'school-directeur'), maar het kan ook 3 zijn ('school-directeuren-overleg'). Maar de lijst kent zijn beperkingen. Nieuwe of ongebruikelijke samenstellingen ('dak-dekkers-overleg') worden niet herkend als samenstelling, ook al worden delen daarvan wel herkend ('dak-dekker'). We werken aan een nieuwe methode om samenstellingen te herkennen, op basis van de woorddelen die FROG uit de woorden haalt.

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.

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.

Het derde type woordfrequentiegegevens geeft net als het eerst type een proportie van tekstwoorden die als frequent 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. Daarbij zijn er drie keuzes te maken voor het onderliggende corpus:

- SoNaR totaal (Oostdijk et al. 2013);
- SoNaR kranten;
- Subtlex (Keuleers et al. 2010).

Bij SoNaR gaat het vooral om schriftelijk taalgebruik, waarbij informele genres in de minderheid zijn. Subtlex daarentegen is een corpus met Nederlandse ondertitels voor films, en bevat met name alledaagse (zij het niet-spontane) conversatie. Zie Pander Maat et al. 2014 voor enkele verschillen tussen de frequentieprofielen op basis van SoNaR en Subtlex.

Het heeft zin om woordlengtes en woordfrequenties ook zonder namen te geven, omdat uit onderzoek van Camblin et al. 2007 blijkt dat namen anders worden verwerkt dan 'gewone woorden'. Daarom zijn een aantal varianten gedefinieerd waarbij namen worden overgeslagen.

Omdat de lijsten met woordfrequenties erg lang zijn en ingelezen moeten worden, valt er 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.

3.4 Zinscomplexiteit

32.	Wrd_per_zin	Woorden per zin
33.	Wrd_per_dz	Woorden per deelzin
34.	Zin_per_wrd	Zinnen per woord
35.	Dzin_per_wrd	Deelzinnen per woord
36.	Wrd_per_nwg	Woorden per naamwoordgroep
37.	Ov_bijzin_d	Dichtheid van overige bijzinnen
38.	Ov_bijzin_per_zin	Overige bijzinnen per zin
39.	Betr_bijzin_d	Dichtheid van betrekkelijke bijzinnen
40.	Betr_bijzin_dz	Betrekkelijke bijzinnen per deelzin
41.	Pv_d	Dichtheid van persoonsvormen
42.	Pv_per_zin	Persoonsvormen per zin
43.	D_level	D-level
44.	D_level_gd4_p	Proportie zinnen met een D-level hoger dan 4
45.	Nom_d	Nominalisatiedichtheid
46.	Lijdv_d	Dichtheid van lijdende vormen
47.	Lijdv_dz	Aantal lijdende vormen per deelzin
48.	Ontk_zin_d	Dichtheid van zinsontkenningen
49.	Ontk_zin_dz	Zinsontkenningen per deelzin
50.	Ontk_morf_d	Dichtheid van morfologische ontkenningen
51.	Ontk_morf_dz	Morfologische ontkenningen per deelzin
52.	Ontk_tot_d	Dichtheid van ontkenningen totaal
53.	Ontk_tot_dz	Ontkenningen totaal per deelzin
54.	Meerv_ontk_d	Dichtheid van meervoudige ontkenningen
55.	Meerv_ontk_dz	Meervoudige ontkenningen per deelzin
56.	AL_sub_ww	Afhankelijkheidslengte (AL) tussen werkwoord en bijbehorend subject, in woorden
57.	AL_ob_ww	AL werkwoord – bijbehorend direct object
58.	AL_indirob_ww	AL werkwoord – bijbehorend indirect object
59.	AL_ww_vzg	AL werkwoord – bijbehorende bijwoordelijke voorzetselgroep
60.	AL_lidw_znw	AL zelfstandig naamwoord – bijbehorend lidwoord
61.	AL_vz_znw	AL voorzetsel – bijbehorend naamwoord
62.	AL_pv_hww	AL persoonsvorm – bijbehorend hoofdwerkwoord
63.	AL_vg_wwbijzin	AL voegwoord – persoonsvorm van de bijbehorende bijzin
64.	AL_vg_conj	AL voegwoord – hoofd van het bijbehorende conjunct
65.	AL_vg_wwhoofdzin	AL voegwoord – persoonsvorm van bijbehorende hoofdzin
66.	AL_znw_bijzin	AL naamwoord – hoofd van de bijbehorende betrekkelijke bijzin
67.	AL_ww_schdw	AL werkwoord – scheidbaar deel van dit werkwoord
68.	AL_ww_znwpred	AL koppelwerkwoord – zelfstandig-naamwoordpredicaat
69.	AL_ww_bnwpred	AL koppelwerkwoord – bijvoeglijk-naamwoordpredicaat
70.	AL_ww_bnwbwp	AL werkwoord – bijwoordelijke bepaling met een bijvoeglijk naamwoord
71.	AL_ww_bwbwp	AL werkwoord – bijwoordelijke bepaling met een bijwoord
72.	AL_ww_znwbwp	AL werkwoord – bijwoordelijke bepaling met een zelfstandig naamwoord
73.	AL_gem	Het gemiddelde van alle afhankelijkheidslengtes per zin
74.	AL_max	Maximale AL per zin
75.	Bijw_bep_d	Dichtheid van bijwoordelijke bepalingen
76.	Bijw_bep_dz	Bijwoordelijke bepalingen per deelzin
77.	Attr_bijv_nw_d	Dichtheid van attributieve bijvoeglijke naamwoorden
78.	Attr_bijv_nw_dz	Aantal attributieve bijvoeglijke naamwoorden per deelzin
79.	Bijv_bep_d	Dichtheid van bijvoeglijke bepalingen
80.	Bijv_bep_dz	Aantal bijvoeglijke bepalingen per deelzin
81.	Ov_bijv_bep_d	Dichtheid van bijvoeglijke bepalingen zonder bijvoeglijke naamwoorden
82.	Ov_bijv_bep_dz	Aantal van bijvoeglijke bepalingen zonder adjectieven per deelzin

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. Dat betekent dat zogenaamde beknopte bijzinnen niet als deelzin worden gezien. De eerstvolgende zin bevat dus volgens T-Scan slechts twee deelzinnen, de zin erna slechts één:

Nadat we afscheid genomen hadden, vertrokken we.
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.

Bijzinnen
Een aantal andere kenmerken gaat over bijzinnen, een categorie die dus net wat ruimer is dan deelzinnen. Daarbij wordt onderscheid gemaakt tussen betrekkelijke bijzinnen en overige bijzinnen. De betrekkelijke bijzinnen worden onderscheiden op basis van het betrekkelijk voornaamwoord dat hen inleidt.

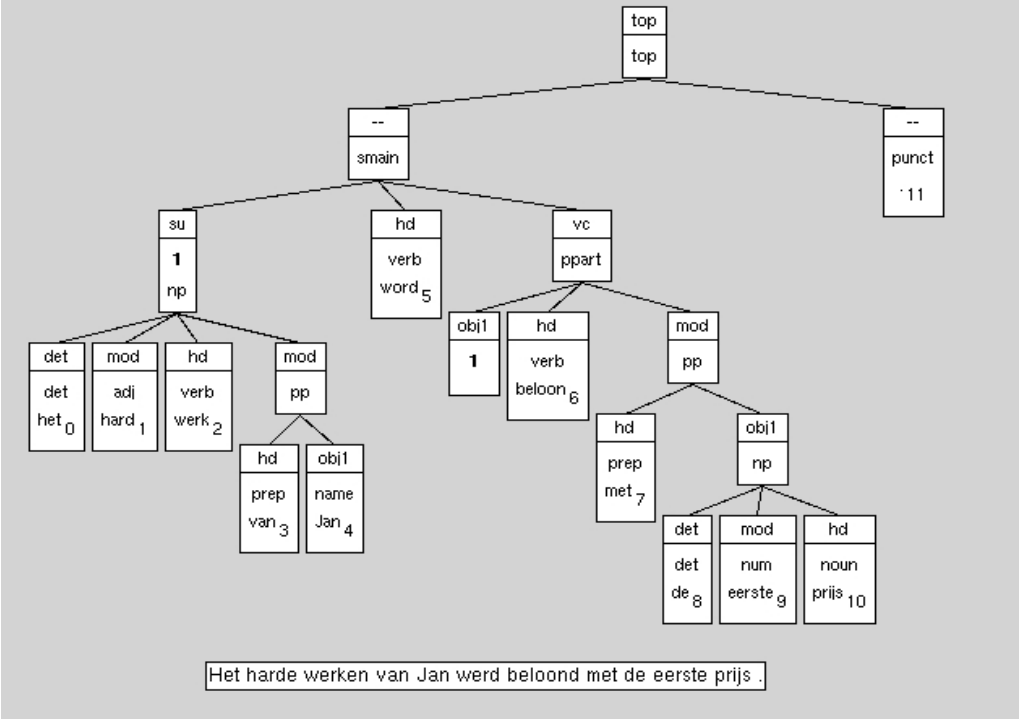
D-level
Twee kenmerken behandelen het D-level van de zin. D-level is een afkorting van 'Development Level'; het gaat om een classificatie en rangordening van zinstypen naar moeilijkheid, oorspronkelijk afkomstig uit Rosenberg & Abbeduto (1987). Zinsconstructies worden geordend op de D_levelschaal op 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. Zie verder *appendix X*.

T-Scan geeft het D-level, alsmede de proportie zinnen met een D-level hoger dan 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.

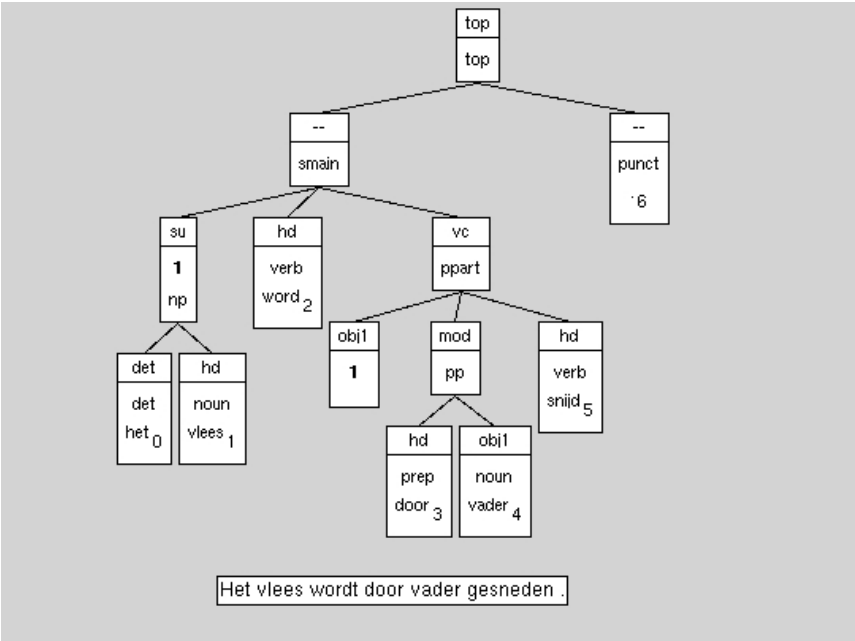
Daarnaast kennen we in het Nederlands nog genominaliseerde infinitieven ('Het harde werken van Jan werd beloond met de eerste prijs.'). Deze zijn redelijk makkelijk te herkennen aan de hand van de Alpino ontleding: het gaat in deze gevallen altijd om een werkwoordsknoop die onder een NP valt. Zie figuur 1.1.



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

Lijdende vormen

Ook lijdende vormen worden geïdentificeerd op basis van Alpino. In de Alpino-output wordt gekeken of het een hulpwerkwoord van passieve vorm betreft: een vorm van 'worden' of 'zijn' met een VC (verbal complement) dependent met als hoofd het hoofdwerkwoord dat een 'lege' object dependent heeft die een index deelt met het subject. Alpino laat dus zien dat het subject in de passieve zin in feite het object is van het hoofdwerkwoord in de actieve zin. Zie figuur 1.2.



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

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.

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. Het hoofd van een zin is het vervoegde werkwoord, met als dependents 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 dependents, hoe lastiger dat lijkt te zijn 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 1 geven we per type een voorbeeld. Daarbij wordt ook duidelijk dat nogal was 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.

De laatste twee kenmerken rondom afhankelijkheidslengtes proberen een samenvatting te bieden. Allereerst wordt het gemiddelde gegeven van alle soorten lengtes voor een bepaalde zin. Die soorten lengtes die 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 score die is aangetroffen onder de kenmerken 54-70. Een hoge score bij AL_max is een handige indicatie dat er iets aan de hand is met een zin. Het ontbreken van een hoge score bij AL_max is helaas geen garantie dat er niets aan de hand is, omdat onder een laag gemiddelde nog steeds een hoge afzonderlijke lengte schuil kan gaan.

Soort lengte	**Voorbeeld**
AL werkwoord ─ subject	**Peter**, mijn neef uit Canada, **schrijft** regelmatig. **Ik ben** gisteren naar de bioscoop **gegaan**. Voor deze zin wordt het gemiddelde genomen van de afstanden tussen *ik* en *ben* (0) en die tussen *ik* en *gegaan* (4).
AL werkwoord ─ direct object	Karel **gaf** mij een geweldige **roman**.
AL werkwoord ─ bijbehorend 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 persoonsvorm ─ hoofdwerkwoord	Ik **ben** gisteren naar de bioscoop **gegaan**.
AL voegwoord ─ persoonsvorm bijzin	Ik ging naar huis **omdat** ik moe geworden **was**.
AL voegwoord ─ hoofd conjunct	Ik heb hem twee **boeken** gegeven **en** drie hele kleine **plantjes**. Ik **gaf** hem een boek **en** hij **was** niet eens blij. (De aard van het 'hoofd' hangt af van de aard van de door het voegwoord verbonden conjuncten.)
AL voegwoord ─ persoonsvorm hoofdzin	**Omdat** ik moe geworden was, **ging** ik naar huis.
AL naamwoord ─ hoofd betrekkelijke bijzin	Hij heeft die **man** gezien die jij gisteren **sprak**.
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 1. Voorbeelden van afhankelijkheidslengtes in T-Scan

Als laatste moeten we opmerken dat Alpino soms fouten maakt, met name in lange zinnen. In Pander Maat et al. (2014) geven we een voorbeeld van een voorzetselgroep aan het eind van een lange zin die ten onrechte aan het hoofdwerkwoord wordt opgehangen, waarmee een erg lange lengte ontstaat. In de betreffende zin stonden nogal veel voorzetselgroepen, waarmee kennelijk de kans op incorrecte aanhechting toeneemt.

Bijwoordelijke en bijvoeglijke bepalingen

Bijwoordelijke bepalingen worden door T-Scan alleen globaal benoemd, en geteld per deelzin en per 1000 woorden. Daarbij moeten we bedenken dat een bijwoordelijke bepaling zowel kan bestaan uit een woord (… *eens*) als uit grote woordgroepen (… *op een feestje in Boston, waar ik toen woonde*).

Ook voor bijvoeglijke bepalingen worden globale maten gegeven, maar hier is het ook mogelijk om verder te onderscheiden tussen attributieve adjectieven en andersoortige bijvoeglijke bepalingen. Daarbij heeft T-Scan op basis van Alpino oog voor bijvoeglijke voorbepalingen bestaande uit:

- 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*)

Daarnaast worden nabepalingen geteld bestaande uit:

- Bijstellingen (*de schipper, een voorzichtig man, bleef thuis*)
- Naamwoorden of naamwoordgroepen (*de wedstrijd Ajax-Anderlecht vorige week*: twee bepalingen)
- 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 (*de kans dat hij weer opknapt; de groep waartoe de herten behoren*).

3.5 Referentiële coherentie en lexicale diversiteit

83.	TTR_wrd	Type-token-ratio voor woorden
84.	MTLD_wrd	Measure of textual lexical diversity voor woorden
85.	TTR_lem	Type-token-ratio voor lemma's
86.	MTLD_lem	Measure of textual lexical diversity voor lemma's
87.	TTR_namen	Type-token-ratio voor namen
88.	MTLD_namen	Measure of textual lexical diversity voor namen
89.	TTR_inhwrd	Type-token-ratio voor inhoudswoorden
90.	MTLD_inhwrd	Measure of textual lexical diversity voor inhoudswoorden
91.	Inhwrd_d	Dichtheid van inhoudswoorden
92.	Inhwrd_dz	Aantal inhoudswoorden per deelzin
93.	Zeldz_index	De proportie lemma's die minder dan vijf keer voorkomen
94.	Vnw_ref_d	Dichtheid van terugverwijzende voornaamwoorden
95.	Vnw_ref_dz	Terugverwijzende voornaamwoorden per deelzin
96.	Arg_over_vzin_d	Dichtheid van argumenten die voorkomen in de vorige zin
97.	Arg_over_vzin_dz	Aantal argumenten die voorkomen in de vorige zin per deelzin
98.	Lem_over_vzin_d	Dichtheid van lemma's die voorkomen in met de vorige zin
99.	Lem_over_vzin_dz	Aantal lemma's die voorkomen in de vorige zin per deelzin
100.	Arg_over_buf_d	Dichtheid van argumenten die voorkomen in de voorgaande X woorden; X is de bufferomvang die in te stellen is
101.	Arg_over_buf_dz	Aantal argumenten die voorkomen in de voorgaande X woorden per deelzin
102.	Lem_over_buf_d	Dichtheid van lemma's die voorkomen in de voorgaande X woorden
103.	Lem_over_buf_dz	Aantal lemma's die voorkomen in de voorgaande X woorden per deelzin
104.	Onbep_nwg_p	Proportie onbepaalde naamwoordgroepen op naamwoordgroepen
105.	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.

Een andere woordherhalingsmaat is de zeldzaamheidsindex, die staat voor de proportie lemma's in de tekst die minder bepaald aantal keren voorkomen in de tekst. De drempelwaarde is in te stellen bij 'Rarity level' (zie hoofdstuk 2.4 hierboven). Het ligt in de rede om de drempelwaarde voor de zeldzaamheidsindex te laten afhangen van de tekstlengte.

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 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 defaultwaarde die aangeraden wordt in de literatuur:

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

De MTLD werk zowel 'heen' (voorwaarts door de tekst) als 'terug' (achterwaarts). De voorwaartse berekening staat in Tabel 2 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 2. 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 wanneer je 50, 100, 200 of 300 woorden uit een tekst neemt maar dat de MLTD-waarden veel stabieler zijn. Dat kan een voordeel zijn, maar wat de meest interessante maat is, wordt bepaald door de onderzoeksvraag.

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 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 zinsoverlapmaten 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 bufferoverlapmaten wordt de buffer als een 'venster' over de tekst heen geschoven, en wordt telkens voor het woord na de buffer overlap bekeken. In een tekst van 100 woorden

met een bufferinstelling van 50 wordt dus eerst gekeken of woord 51 overlap vertoont met een woord uit de woorden 1-50, vervolgens wordt woord 52 vergeleken 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.

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 zelfstandig naamwoorden, werkwoorden, adjectieven en bijwoorden per 1000 woorden, of genomen per deelzin.

Terugverwijzende voornaamwoorden

Een tekstkenmerk dat vrij rechtstreeks teruggaat op referentiële coherentie is het aantal terugverwijzende voornaamwoorden. Daaronder worden voornaamwoorden gerekend 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

106.	Conn_temp_d	Dichtheid van temporele verbindingswoorden
107.	Conn_temp_dz	Temporele verbindingswoorden per deelzin
108.	Conn_temp_TTR	Type-token-ratio voor temporele verbindingswoorden
109.	Conn_temp_MTLD	Measure of Lexical Diversity in Text voor temporele verbindingswoorden
110.	Conn_reeks_wg_d	Dichtheid van reeksaanduiders voor woordgroepen
111.	Conn_reeks_wg_dz	Reeksaanduiders per deelzin voor woordgroepen
112.	Conn_reeks_wg_TTR	Type-token-ratio voor reeksaanduiders voor woordgroepen
113.	Conn_reeks_wg_MTLD	Measure of textual lexical diversity voor reeksaanduiders voor woordgroepen
114.	Conn_reeks_zin_d	Dichtheid van reeksaanduiders voor (deel)zinnen
115.	Conn_reeks_zin_dz	Reeksaanduiders per deelzin voor (deel)zinnen
116.	Conn_reeks_zin_TTR	Type-token-ratio voor reeksaanduiders voor (deel)zinnen
117.	Conn_reeks_zin_MTLD	Measure of textual lexical diversity voor reeksaanduiders voor (deelzinnen)
118.	Conn_contr_d	Dichtheid van contrastieve verbindingswoorden
119.	Conn_contr_dz	Contrastieve verbindingswoorden per deelzin
120.	Conn_contr_TTR	Type-token-ratio voor contrastieve verbindingswoorden
121.	Conn_contr_MTLD	Measure of Lexical Diversity in Text voor contrastieve verbindingswoorden
122.	Conn_comp_d	Dichtheid van comparatieve verbindingswoorden
123.	Conn_comp_dz	Comparatieve verbindingswoorden per deelzin
124.	Conn_comp_TTR	Type-token-ratio voor comparatieve verbindingswoorden
125.	Conn_comp_MTLD	Measure of Lexical Diversity in Text voor comparatieve verbindingswoorden
126.	Conn_caus_d	Dichtheid van causale verbindingswoorden
127.	Conn_caus_dz	Causale verbindingswoorden per deelzin
128.	Conn_caus_TTR	Type-token-ratio voor causale verbindingswoorden
129.	Conn_caus_MTLD	Measure of Lexical Diversity in Text voor causale verbindingswoorden
130.	Causaal_d	Dichtheid van causale inhoudswoorden
131.	Ruimte_d	Dichtheid van ruimtewoorden
132.	Tijd_d	Dichtheid van tijdwoorden
133.	Emotie_d	Dichtheid van emotiewoorden
134.	Causaal_TTR	Type-token-ratio voor causale inhoudswoorden
135.	Causaal_MTLD	Measure of textual lexical diversity voor causale inhoudswoorden
136.	Ruimte_TTR	Type-token-ratio voor ruimtewoorden
137.	Ruimte_MTLD	Measure of Lexical Diversity in Text voor ruimtewoorden
138.	Tijd_TTR	Type-token-ratio voor tijdwoorden
139.	Tijd_MTLD	Measure of Lexical Diversity in Text voor tijdwoorden
140.	Emotie_TTR	Type-token-ratio voor emotiewoorden
141.	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 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.

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 en woordconcreetheid

Om vast te stellen hoe concreet de woorden in een tekst zijn, gebruikt T-Scan semantisch geannoteerde woordenlijsten die oorspronkelijk afkomstig zijn uit het Referentie Bestand Nederlands (Martin & Maks 2005). Deze lijsten zijn echter handmatig gecorrigeerd, uitgebreid en gehergroepeerd. Er is een lijst met zelfstandige naamwoorden, een met adjectieven en een met werkwoorden. Omdat het aantal kenmerken groot is, geven we overzichten per woordsoort.

3.7.1 Zelfstandige naamwoorden

142.	Conc_nw_strikt_p	Proportie van strikt-concrete naamwoorden
143.	Conc_nw_strikt_d	Dichtheid van strikt-concrete naamwoorden
144.	Conc_nw_ruim_p	Proportie van ruim-concrete naamwoorden
145.	Conc_nw_ruim_d	Dichtheid van ruim-concrete naamwoorden
146.	Pers_nw_p	Proportie van naamwoorden verwijzend naar personen
147.	Pers_nw_d	Dichtheid van naamwoorden verwijzend naar personen
148.	PlantDier_nw_p	Proportie van naamwoorden verwijzend naar planten en dieren
149.	PlantDier_nw_d	Dichtheid naamwoorden verwijzend naar planten en dieren
150.	Gebr_vw_nw_p	Proportie van naamwoorden verwijzend naar gebruiksvoorwerpen
151.	Gebr_vw_nw_d	Dichtheid van naamwoorden verwijzend naar gebruiksvoorwerpen
152.	Subst_conc_nw_p	Proportie van naamwoorden verwijzend naar concrete substanties
153.	Subst_conc_nw_d	Dichtheid van naamwoorden verwijzend naar concrete substanties
154.	Voed_verz_nw_p	Proportie van naamwoorden verwijzend naar voeding en verzorging
155.	Voed_verz_nw_d	Dichtheid van naamwoorden verwijzend naar voeding en verzorging
156.	Concr_ov_nw_p	Proportie van overige concrete naamwoorden
157.	Concr_ov_nw_d	Dichtheid van overige concrete naamwoorden
158.	Gebeuren_conc_nw_p	Proportie naamwoorden verwijzend naar concrete gebeurtenissen
159.	Gebeuren_conc_nw_d	Dichtheid naamwoorden verwijzend naar concrete gebeurtenissen
160.	Plaats_nw_p	Proportie van naamwoorden verwijzend naar plaatsen en ruimtes
161.	Plaats_nw_d	Dichtheid van naamwoorden verwijzend naar plaatsen en ruimtes
162.	Tijd_nw_p	Proportie van naamwoorden verwijzend naar tijd
163.	Tijd_nw_d	Dichtheid van naamwoorden verwijzend naar tijd
164.	Maat_nw_p	Proportie van naamwoorden verwijzend naar maten
165.	Maat_nw_d	Dichtheid van naamwoorden verwijzend naar maten
166.	Subst_abstr_nw_p	Proportie van naamwoorden verwijzend naar abstracte substanties
167.	Subst_abstr_nw_d	Dichtheid van naamwoorden verwijzend naar abstracte substanties
168.	Gebeuren_abstr_nw_p	Proportie naamwoorden verwijzend naar abstracte gebeurtenissen
169.	Gebeuren_abstr_nw_d	Dichtheid naamwoorden verwijzend naar abstracte gebeurtenissen
170.	Organisatie_nw_p	Proportie van naamwoorden verwijzend naar organisaties
171.	Organisatie_nw_d	Dichtheid van naamwoorden verwijzend naar organisaties
172.	Ov_abstr_nw_p	Proportie overige abstracte naamwoorden
173.	Ov_abstr_nw_d	Dichtheid overige abstracte naamwoorden
174.	Undefined_nw_p	Proportie van naamwoorden die ongedefinieerd blijven in de lijst
175.	Gedekte_nw_p	Proportie van naamwoorden die in de lijst staan

Zelfstandige naamwoorden zijn opgedeeld in veertien klassen, zie Tabel 3. De RBN-lijst is handmatig gecorrigeerd, uitgebreid en gehergroepeerd door H. Pander Maat en telt nu zo'n 46000 woorden. 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 woorden ongedefinieerd gelaten, wat een vijftiende groep oplevert.

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	*IJswater, 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. Abstract niet-dynamisch	*Christendom, motto*	Abstract
14. Organisatie	*Werkgeversorganisatie*	Abstract
15. Undefined	*Kant, poot*	

Tabel 3. 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 alleen lemma's met de woordsoort 'noun' opzoekt in de lijst. Bij verkeerde herkenning van de woordsoort wordt er dus geen semantisch label toegekend.

3.7.2 Bijvoeglijke naamwoorden

176.	Waarn_mens_bvnw_p	Proportie van bijvoeglijke naamwoorden over waarneembare kenmerken van mensen
177.	Waarn_mens_bvnw_d	Dichtheid van bijv. naamwoorden over waarneembare kenmerken van mensen
178.	Emosoc_bvnw_p	Proportie van bijvoeglijke naamwoorden over emoties en sociaal gedrag
179.	Emosoc_bvnw_d	Dichtheid van bijv. naamwoorden over emoties en sociaal gedrag
180.	Waarn_nmens_bvnw_p	Proportie van bijv. naamwoorden die verwijzend naar waarneembare kenmerken van dingen
181.	Waarn_nmens_bvnw_d	Dichtheid van bijv. naamwoorden die verwijzend naar waarneembare kenmerken van dingen
182.	Vorm_omvang_bvnw_p	Proportie van bijv. naamwoorden die verwijzend naar vorm en omvang
183.	Vorm_omvang_bvnw_d	Dichtheid van bijv. naamwoorden die verwijzend naar vorm en omvang
184.	Kleur_bvnw_p	Proportie van bijv. naamwoorden die verwijzend naar kleur en licht
185.	Kleur_bvnw_d	Dichtheid van bijv. naamwoorden die verwijzend naar kleur en licht
186.	Stof_bvnw_p	Proportie van bijv. naamwoorden die verwijzend naar stoffen
187.	Stof_bvnw_d	Dichtheid van bijv. naamwoorden die verwijzend naar stoffen

188.	Geluid_bvnw_p	Proportie van bijv. naamwoorden die verwijzend naar geluid
189.	Geluid_bvnw_d	Dichtheid van bijv. naamwoorden die verwijzend naar geluid
190.	Waarn_nmens_ov_bvnw_p	Proportie van bijv. naamwoorden die verwijzend naar andere waarneembare kenmerken van dingen
191.	Waarn_nmens_ov_bvnw_d	Dichtheid van bijv. naamwoorden die verwijzend naar andere waarneembare kenmerken van dingen
192.	Technisch_bvnw_p	Proportie van bijv. naamwoorden die verwijzend naar technisch waarneembare kenmerken van dingen
193.	Technisch_bvnw_d	Dichtheid van bijv. naamwoorden die verwijzend naar technisch waarneembare kenmerken van dingen
194.	Tijd_bvnw_p	Proportie van bijv. naamwoorden die verwijzend naar tijd
195.	Tijd_bvnw_d	Dichtheid van bijv. naamwoorden die verwijzend naar tijd
196.	Plaats_bvnw_p	Proportie van bijv. naamwoorden die verwijzend naar plaats en ruimte
197.	Plaats_bvnw_d	Dichtheid van bijv. naamwoorden die verwijzend naar plaats en ruimte
198.	Spec_positief_bvnw_p	Proportie van bijv. naamwoorden die specifiek positief evalueren
199.	Spec_positief_bvnw_d	Dichtheid van bijv. naamwoorden die specifiek positief evalueren
200.	Spec_negatief_bvnw_p	Proportie van bijv. naamwoorden die specifiek negatief evalueren
201.	Spec_negatief_bvnw_d	Dichtheid van bijv. naamwoorden die specifiek negatief evalueren
202.	Alg_positief_bvnw_p	Proportie van bijv. naamwoorden die algemeen positief evalueren
203.	Alg_positief_bvnw_d	Dichtheid van bijv. naamwoorden die algemeen positief evalueren
204.	Alg_negatief_bvnw_p	Proportie van bijv. naamwoorden die algemeen negatief evalueren
205.	Alg_negatief_bvnw_d	Dichtheid van bijv. naamwoorden die algemeen negatief evalueren
206.	Alg_ev_zr_bvnw_p	Proportie van bijv. naamwoorden die evalueren zonder richting
207.	Alg_ev_zr_bvnw_d	Dichtheid van bijv. naamwoorden die evalueren zonder richting
208.	Ep_positief_bvnw_p	Proportie van bijv. naamwoorden die epistemisch positief evalueren
209.	Ep_positief_bvnw_d	Dichtheid van bijv. naamwoorden die epistemisch positief evalueren
210.	Ep_negatief_bvnw_p	Proportie van bijv. naamwoorden die epistemisch negatief evalueren
211.	Ep_negatief_bvnw_d	Dichtheid van bijv. naamwoorden die epistemisch negatief evalueren
212.	Abstract_ov_bvnw_p	Proportie van de overige abstracte bijv. naamwoorden
213.	Abstract_ov _bvnw_d	Dichtheid van de overige abstracte bijv. naamwoorden
214.	Spec_ev_bvnw_p	Proportie van bijv. naamwoorden die specifiek evalueren
215.	Spec_ev_bvnw_d	Dichtheid van bijv. naamwoorden die specifiek evalueren
216.	Alg_ev_bvnw_p	Proportie van bijv. naamwoorden die algemeen evalueren
217.	Alg_ev_bvnw_d	Dichtheid van bijv. naamwoorden die algemeen evalueren
218.	Ep_ev_bvnw_p	Proportie van bijv. naamwoorden die epistemisch evalueren
219.	Ep_ev_bvnw_d	Dichtheid van bijv. naamwoorden die epistemisch evalueren
220.	Conc_bvnw_strikt_p	Proportie van strikt-concrete bijv. naamwoorden
221.	Conc_bvnw_strikt_d	Dichtheid van strikt-concrete bijv. naamwoorden
222.	Conc_bvnw_ruim_p	Proportie van ruim-concrete bijv. naamwoorden
223.	Conc_bvnw_ruim_d	Dichtheid van ruim-concrete bijv. naamwoorden
224.	Subj_bvnw_p	Proportie van subjectieve bijv. naamwoorden
225.	Subj_bvnw_d	Dichtheid van subjectieve bijv. naamwoorden die
226.	Undefined_bvnw_p	Proportie van bijv. naamwoorden die in de lijst ongedefinieerd zijn
227.	Gelabeld_bvnw_p	Proportie van bijv. naamwoorden die in de lijst een label krijgen
228.	Gedekte_bvnw_p	Proportie van bijv. naamwoorden die in de lijst staan

De adjectieven uit de oorspronkelijke RBN-lijst zijn opnieuw geclassificeerd, en er zijn woorden aan toegevoegd. De indeling is te vinden in Tabel 4. De nieuwe indeling is wat gedetailleerder dan de eerdere: 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. Neutrale abstracte term	*aanverwant, aandachtig*
11. Ongedefinieerd	*belastbaar, druk, small*

Table 4. 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 onderscheidene 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 die klassen 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 globale concreetheidskenmerken

229.	Conc_ww_p	Proportie van concrete werkwoorden
230.	Conc_ww_d	Dichtheid van concrete werkwoorden op werkwoorden
231.	Abstr_ww_p	Proportie van abstracte werkwoorden
232.	Abstr_ww_d	Dichtheid van abstracte werkwoorden op werkwoorden
233.	Undefined_ww_p	Proportie van werkwoorden die in de lijst ongedefinieerd blijven
234.	Gedekte_ww_p	Proportie van werkwoorden die in de lijst staan
235.	Conc_tot_p	Proportie van (strikt-)concrete naamwoorden, bijv. naamwoorden en werkwoorden
236.	Conc_tot_d	Dichtheid van (strikt-)concrete naamwoorden, bijv. naamwoorden 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 5. 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 5. Werkwoorden gecodeerd naar concreetheid en 'state of affairs'

Globale concreetheidskenmerken

Ten slotte is over zelfstandige naamwoorden, bijvoeglijke naamwoorden en werkwoorden het totaal aantal concrete woorden gesommeerd, en uitgedrukt als een proportie van het totaal aantal nomina, adjectieven en verba, en ook als dichtheid.

3.8 Persoonlijke elementen

237.	Pers_ref_d	Dichtheid van verwijzingen naar personen
238.	Pers_vnw1_d	Dichtheid van persoonlijke en bezittelijke voornaamwoorden van de eerste persoon
239.	Pers_vnw2_d	Dichtheid van persoonlijke en bezittelijke voornaamwoorden van de tweede persoon
240.	Pers_vnw3_d	Dichtheid van persoonlijke en bezittelijke voornaamwoorden van de derde persoon
241.	Pers_vnw_d	Dichtheid van alle persoonlijke en bezittelijke voornaamwoorden

T-Scan geeft dichtheden voor verschillende soorten verwijzingen naar personen. Om te beginnen gaan kenmerk 227-230 over zowel persoonlijke als bezittelijke voornaamwoorden verwijzendd naar personen. 'Het' is dus niet meegeteld. Omdat het gaat om indicaties van het persoonlijke karakter van een tekst, is ook een eerder onpersoonlijke verwijzing als 'men' hier buiten beschouwing gelaten.

Als verwijzingen naar een persoon worden hier opgevat:

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

3.9 Andere lexicale informatie

3.9.1 Namen

242.	Pers_namen_p	Proportie van persoonsnamen op alle namen
243.	Pers_namen_p2	Proportie van persoonsnamen op alle namen en naamwoorden
244.	Pers_namen_d	Dichtheid van persoonsnamen
245.	Plaatsnamen_d	Dichtheid van plaatsnamen
246.	Org_namen_d	Dichtheid van organisatienamen
247.	Prod_namen_d	Dichtheid van productnamen
248.	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).

3.9.2 Werkwoorden

249.	Actieww_p	Proportie van actiewerkwoorden op werkwoorden
250.	Actieww_d	Dichtheid van actiewerkwoorden
251.	Toestww_p	Proportie van toestandswerkwoorden op werkwoorden
252.	Toestww_d	Dichtheid van toestandswerkwoorden
253.	Procesww_p	Proportie van proceswerkwoorden op werkwoorden
254.	Procesww_d	Dichtheid van proceswerkwoorden
255.	Undefined_ATP_ww_p	Proportie van werkwoorden die ongedefinieerd blijven voor het kenmerk Actie/Proces/Toestand (ATP)
250.	Ww_tt_p	Proportie van tegenwoordige tijden op alle persoonsvormen
251.	Ww_tt_dz	Aantal werkwoorden in tegenwoordige tijd per deelzin
252.	Ww_mod_d	Dichtheid van modale werkwoorden
253.	Ww_mod_dz	Aantal modale werkwoorden per deelzin
254.	Huww_tijd_d	Dichtheid van hulpwerkwoorden van tijd
255.	Huww_tijd_dz	Aantal van hulpwerkwoorden van tijd per deelzin
256.	Koppelww_d	Dichtheid van koppelwerkwoorden
257.	Koppelww_dz	Gemiddeld aantal koppelwerkwoorden per deelzin
258.	Infin_bv_d	Dichtheid van bijvoeglijke infinitieven
259.	Infin_bv_dz	Bijvoeglijke infinitieven per deelzin
260.	Infin_nw_d	Dichtheid van naamwoordelijke infinitieven
261.	Infin_nw_dz	Naamwoordelijke infinitieven per deelzin
262.	Infin_vrij_d	Dichtheid van vrijstaande infinitieven
263.	Infin_vrij_dz	Vrijstaande infinitieven per deelzin
264.	Vd_bv_d	Dichtheid van bijvoeglijke voltooid deelwoorden
265.	Vd_bv_dz	Bijvoeglijke voltooid deelwoorden per deelzin
266.	Vd_nw_d	Dichtheid van naamwoordelijke voltooid deelwoorden
267.	Vd_nw_dz	Naamwoordelijke voltooid deelwoorden per deelzin
268.	Vd_vrij_d	Dichtheid van vrijstaande voltooid deelwoorden
269.	Vd_vrij_dz	Vrijstaande voltooid deelwoorden per deelzin
270.	Ovd_bv_d	Dichtheid van bijvoeglijke onvoltooid deelwoorden
271.	Ovd_bv_dz	Bijvoeglijke onvoltooid deelwoorden per deelzin
272.	Ovd_nw_d	Dichtheid van naamwoordelijke onvoltooid deelwoorden
273.	Ovd_nw_dz	Naamwoordelijke onvoltooid deelwoorden per deelzin
274.	Ovd_vrij_d	Dichtheid van vrijstaande onvoltooid deelwoorden
275.	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 (dichtheid en aantal tt-vormen per deelzin) berekend.

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. Daarbij worden alle vormen van *zullen* opgevat als hulpwerkwoorden van tijd, ook in zinnen als 'ik zou het niet doen'.

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 lezen 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

276.	Imp_ellips_p	Proportie van gebiedende wijzen op persoonsvormen/deelzinnen
277.	Imp_ellips_d	Dichtheid van gebiedende wijzen
278.	Vragen_p	Proportie van vraagzinnen op zinnen
279.	Vragen_d	Dichtheid van vraagzinnen

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

280.	Bvnw_d	Dichtheid van bijvoeglijke naamwoorden
281.	Vg_d	Dichtheid van voegwoorden
282.	Vnw_d	Dichtheid van voornaamwoorden
283.	Lidw_d	Dichtheid van lidwoorden
284.	Vz_d	Dichtheid van voorzetsels
285.	Bijw_d	Dichtheid van bijwoorden
286.	Tw_d	Dichtheid van telwoorden
287.	Nw_d	Dichtheid van zelfstandige naamwoorden
288.	Ww_d	Dichtheid van werkwoorden
289.	Tuss_d	Dichtheid van tussenwerpsels
290.	Spec_d	Dichtheid van speciale typen woord (afkorting of naam)
291.	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.

3.9.5 Afkortingen

292.	Afk_d	Dichtheid van alle afkortingen tezamen
293.	Afk_gen_d	Dichtheid van generieke afkortingen
294.	Afk_int_d	Dichtheid van internationale afkortingen
295.	Afk_jur_d	Dichtheid van juridische afkortingen
296.	Afk_med_d	Dichtheid van medische afkortingen
297.	Afk_ond_d	Dichtheid van onderwijsafkortingen
298.	Afk_pol_d	Dichtheid van politieke afkortingen
299.	Afk_ov_d	Dichtheid van afkortingen overig
300.	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 verwijzendd naar nationaliteiten (BE, UK) of internationale organisaties (OPEC, IMF)
- Juridisch: afkortingen verwijzendd naar wetten, regelingen en juridische organisaties (AAW, Anw)
- Medisch: afkortingen verwijzendd naar aandoeningen (add, ALS)
- Onderwijs (HEAO, DUO)
- Politiek: afkortingen verwijzendd naar overheids- of politieke organisaties (AIVD, DS'70)
- Zorg: afkortingen verwijzendd 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

301.	Vzu_d	Dichtheid van voorzetseluitdrukkingen
302.	Vzu_dz	Aantal voorzetseluitdrukkingen per deelzin
303.	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*).

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

304.	Int_d	Dichtheid van alle intensiveerders uit de lijst bij elkaar
305.	Int_bvnw_d	Dichtheid van de intensiverende adjectieven
306.	Int_bvbw_d	Dichtheid van de intensiverende adjectieven die bijwoordelijk worden gebruikt
307.	Int_bw_d	Dichtheid van de intensiverende bijwoorden
308.	Int_combi_d	Dichtheid van de intensiverende woordcombinaties
309.	Int_nw_d	Dichtheid van de intensiverende naamwoorden
310.	Int_tuss_d	Dichtheid van de intensiverende tussenwerpsels
311.	Int_ww_d	Dichtheid van de intensiverende werkwoorden

3.10 Probabiliteitsmaten

312.	Log_prob	Logaritme van de trigram-probabiliteit
313.	Entropie	Entropie
314.	Perplexiteit	Perplexiteit

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).

Allereerst de voorwaartse trigram-probabliteit: 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 polo*. De probabiliteiten van woorden worden gemiddeld op hogere tekstniveaus.

Entropie is een maat voor onzekerheid in een taal. Hoe onverwachter een taaluiting is, hoe hoger de entropie ervan. Omdat de waarschijnlijkheid van een sequentie van gebeurtenissen veel kleiner wordt naarmate deze langer wordt, is de entropie van een lange zin hoger dan die van een korte zin. Die wie zinnen wil vergelijken op entropie, doet er goed aan de entropie te delen door het aantal woorden in de zin. Hetzelfde geldt voor entropie in teksten.

Perplexiteit is sterk gerelateerd aan entropie (het gaat om de entropie gegeven een bepaald model; voor details, zie Manning & Schütze 1999, 60-78). Bij een hoge entropie is de voorspelbaarheid laag en zijn er veel keuzemogelijkheden: de perplexiteit is dan hoog. Net als entropie stijgt de perplexiteit van een passage naarmate die langer wordt.

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 meer kwalitatief van aard: het gaat om nominale variabelen.

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. POS-tags
    d. Afkortingen
    e. Voorzetseluitdrukkingen
    f. Overig
8. Probabiliteitsmaten

Nr	Naam	Korte 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	Wrdsoort	De woordsoort (Part-Of-Speech): • ADJ = bijvoeglijk naamwoord (adjectief) • BW = bijwoord • LET = interpunctieteken • LID = lidwoord • N = zelfstandig naamwoord (noun) • SPE = speciale eenheden, zoals namen, tijden en URLs • TW = telwoord • VG = voegwoord • VNW = voornaamwoord • VZ = voorzetsel • WW = werkwoord	7c
8	Afk	Afkorting (1=ja; 0=nee)	7d
9	Let_per_wrd	Letters per woord	1
10	Wrd_per_let	Woorden per letter	1
11	Let_per_wrd_zn	Letters per woord, zonder namen	1

Nr	Naam	Korte toelichting	Groep
12	Wrd_per_let_zn	Woorden per letter, zonder namen	1
13	Morf_per_wrd	Morfemen per woord	1
14	Wrd_per_morf	Woorden per morfeem	1
15	Morf_per_wrd_zn	Morfemen per woord, zonder namen	1
16	Wrd_per_morf_zn	Woorden per morfeem, zonder namen	1
17	Sam_delen_per_wrd	Samenstellingsdelen per woord	1
18	Sam_d	Samenstellingsdichtheid	1
19	Freq50_staph	Hoort het woord bij de meest frequente woordtypes die in het Staphorsius-corpus 50% van de woordtokens uitmaken (1=ja; 0=nee)	1
20	Freq65_Staph	Idem maar nu gaat het om 65% van de woordtokens	1
21	Freq77_Staph	Idem maar nu gaat het om 77% van de woordtokens	1
22	Freq80_Staph	Idem maar nu gaat het om 80% van de woordtokens	1
23	Wrd_freq_log	Woordfrequentie, logaritme	1
24	Wrd_freq_zn_log	Woordfrequentie zonder namen, logaritme	1
25	Lem_freq_log	Lemmafrequentie, logaritme	1
26	Lem_freq_zn_log	Lemmafrequentie zonder namen, logaritme	1
27	Freq1000	Hoort het woord bij de meest frequente 1000 woorden	1
28	Freq2000	Idem voor de meest frequente 2000 woorden	1
29	Freq3000	Idem voor de meest frequente 3000 woorden	1
30	Freq5000	Idem voor de meest frequente 5000 woorden	1
31	Freq10000	Idem voor de meest frequente 10000 woorden	1
32	Freq20000	Idem voor de meest frequente 20000 woorden	1
33	connector_type	Type verbindingswoord (als het om een verbindingswoord gaat) • Causaal • Comparatief • Contrastief • Opsommend • Temporeel	4
34	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
35	Vnw_ref	Terugverwijzendd voornaamwoord (1=ja; 0=nee)	3
36	Semtype_nw	Het semantische type van een zelfstandig naamwoord	5
37	Conc_nw_strikt	Is het zelfstandig naamwoord concreet in strikte zin? (1=ja; 0=nee)	5
38	Conc_nw_ruim	Is het zelfstandig naamwoord concreet in ruime zin? (1=ja; 0=nee)	5
39	Semtype_bvnw	Het semantische type van een bijvoeglijk naamwoord	5
40	Conc_bvnw_strikt	Is het bijvoeglijk naamwoord concreet in strikte zin? (1=ja; 0=nee)	5
41	Conc_bvnw_ruim	Is het bijvoeglijk naamwoord concreet in ruime zin? (1=ja; 0=nee)	5
42	Semtype_ww	Het semantische type van een werkwoord; voor waarden	5
43	Pers_ref	Verwijzing naar personen	6
44	Pers_vnw1	Eerste-persoonsvoornaamwoord	6
45	Pers_vnw2	Tweede-persoonsvoornaamwoord	6
46	Pers_vnw3	Derde-persoonsvoornaamwoord	6
47	Pers_vnw	Persoonlijk of bezittelijk voornaamwoord	6
48	Naam_POS	Is het woord een naam? (1=ja; 0=nee) Het gaat hier om de Frog-woordsoort (POS-tag) *spec, eigen*	7a
49	Naam_NER	Welk soort naam is het woord volgens de Named Entity Recognition (NER) module? LOC = plaatsnaam ORG = organisatienaam PER = persoonsnaam PRO = productnaam MISC = andersoortige naam	7a

Nr	Naam	Korte toelichting	Groep
50	Pers_nw	Verwijzing naar een persoon (1=ja; 0=nee)	6
51	Imp	Gaat het om een werkwoord in een zin zonder subject?	7b
52	Ww_vorm	Om welk soort werkwoord gaat het? • Hoofdwerkwoord • Koppelwerkwoord • Modaal werkwoord • Hulpwerkwoord van de lijdende vorm • Hulpwerkwoord van tijd	7b
53	Ww_tt	Gaat het om een vorm in de tegenwoordige tijd? (1=ja; 0=nee)	7b
54	Vol_dw	Gaat het om een voltooid deelwoord? (1=ja; 0=nee)	7b
55	Onvol_dw	Gaat het om een onvoltooid deelwoord? (1=ja; 0=nee)	7b
56	Infin	Gaat het om een infinitief? (1=ja; 0=nee)	7b
57	Archaisch	Gaat het om een archaïsche naamvalsvorm? (1=ja; 0=nee)	1
58	Log_prob	De waarschijnlijkheid van dit woord gegeven de 2 woorden die eraan voorafgaan (hiervan: de logaritme)	8

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.

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.

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.

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.

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.

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. Manuscript.

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.

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

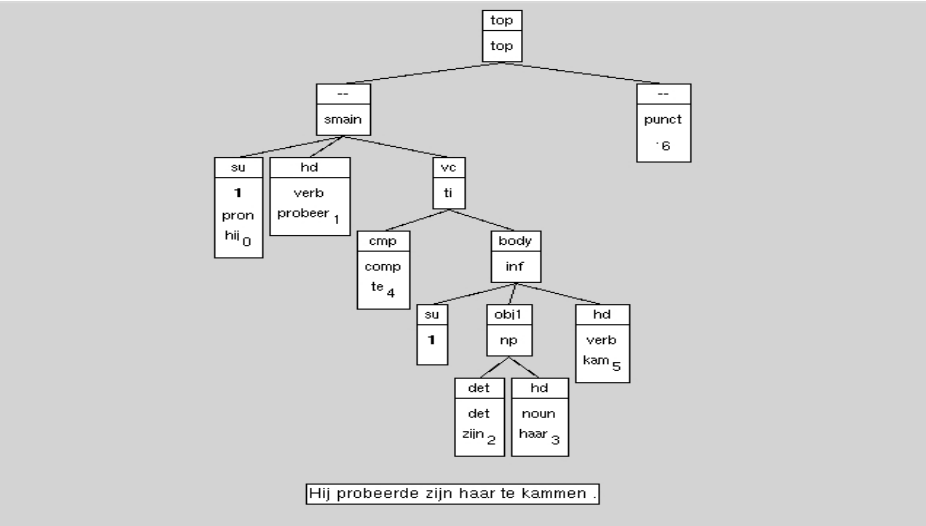
Bijlage A. De implementatie van D-level in T-Scan

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."*

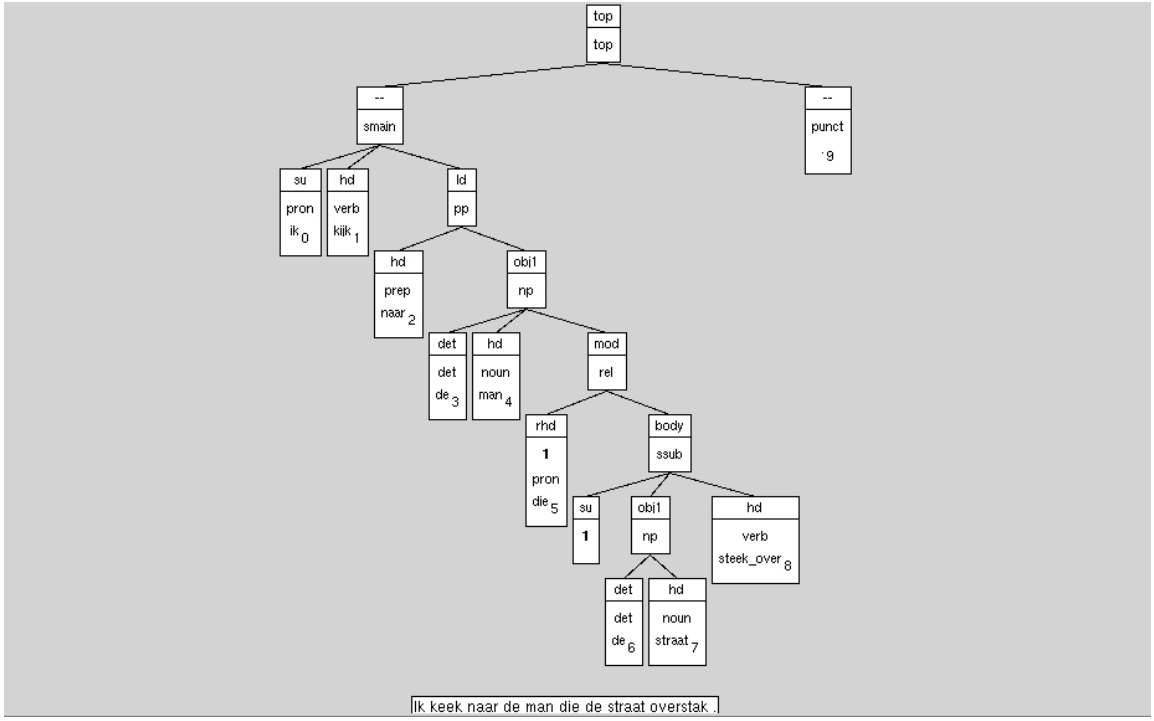
Implementatie: T-Scan controleert zinnen op de aanwezigheid van nevenschikkende voegwoorden, die door Frog aangeduid worden.

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

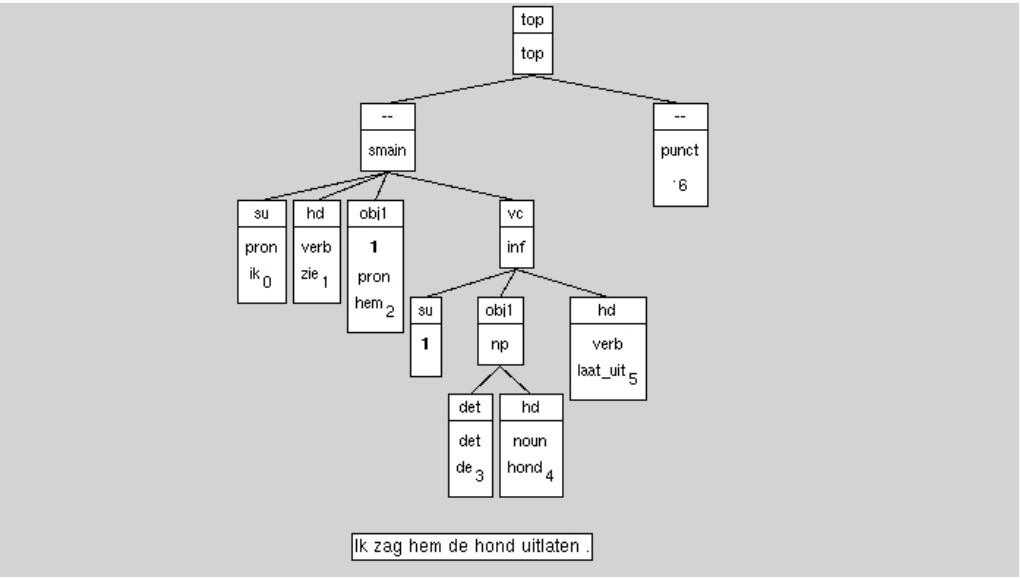
Level 4

– 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.

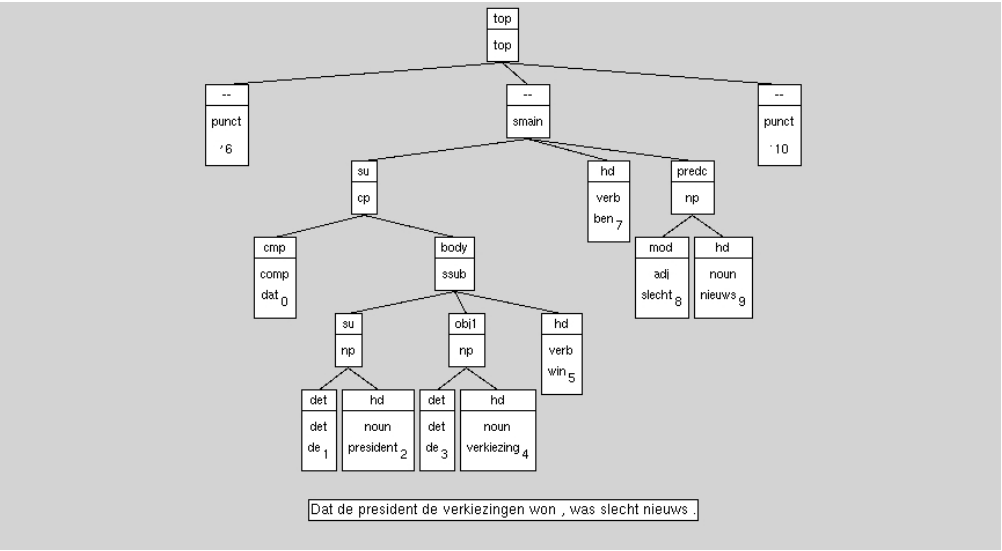
Implementatie: T-Scan controleert zinnen op de aanwezigheid van onderschikkende voegwoorden, die door Frog aangeduid worden.

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 [7].

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 (eerste kolom: woordgroep; kolom 2-4: zin)

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 -*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 in principe meervouden en verkleinwoorden worden gecodeerd op basis van het basiswoord.

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 'gewone' spelling, die met een kleine 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 verwijzend. 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.

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 (incl. eetbare planten: groenten). Organismen die alleen zichtbaar zijn onder de microscoop (*amoebe,* virus) zijn hier niet opgenomen; zij zijn geplaatst bij de abstracte substanties.

Een aantal woorden is zowel human als non-human, zoals *baardaap, beest* en *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 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 woorden die vooral naar de uiterlijke vorm van de tekst verwijzend (*flyer*).

Sommige woorden kunnen algemeen verwijzend 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 die 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*);
- 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 verwijzend 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, voorstelling* of *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.

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, tehuis* en *centrum*. Er zijn wel uitzonderingen: *webwinkels* zijn geen plaatsen, en *workshops* ook niet.

Als Instituut worden gezien:

- afleidingen eindigend op Z-*erij* met de betekenis 'plaats waar ge-X-t wordt': *drukkerij, bakkerij* enz. Dat geldt ook voor –*theek* (maar niet *hypotheek*).
- woorden met uitgang: -*wezen*, -*ziekenhuis*, -*museum*, -*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.

Leeg gemaakt zijn ten slotte de woorden die ambigu zijn tussen Place en Abstract (*splitsing*) of tussen Place en Dynamic (*disco*).

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, fabrieken (incl. bijvoorbeeld energiecentrales), 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 woorden 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.

Daarentegen zijn georganiseerde groepen zoals *bewonersgroep* of *belangengroep* als instituties gezien. Hetzelfde geldt voor maatschappelijke groepen zoals *klootjesvolk* 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*).

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, pijn, 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. Abstract_overig: de abstracte woorden die niet bij 7 tot en met 9 horen
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. Dus 'clownesk' is niet emosoc, want het wordt ook gebruikt voor performances en niet alleen voor personen. Hetzelfde geldt voor 'dominant', 'spannend' en 'diplomatiek'.
Onder 'emosoc' vallen niet:

- opvattingen (*conservatief*) en liefhebberijen (*cinefiel*);
- objectieve kenmerken van personen als *Franssprekend, 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*)
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 en medische eigenschappen (*afbreekbaar, radioactief, bloeddrukverhogend, brachiaal*, *ongewerveld, brandgevaarlijk*)
- 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 verwijzend 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. Concrete en niet-concrete werkwoorden

1 Features

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'.

Op basis van de lijst worden zes features gevormd. Daarnaast noem ik twee features gevormd op basis van het combineren van naamwoord-, adjectief- en werkwoordkenmerken.

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

In de herziene werkwoordenlijst is het feature 'concreetheid' te vinden in kolom D.

2 Hoe zijn de groepen onderscheiden?

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 verwijzend 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 adjectiva 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 adjectiva 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 adjectiva 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

Bijage 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 volgt.

	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 verwijzend 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 tests. 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. *Sterker nog* is dan wel bruikbaar:
    - *N. sterker nog, I.*
    - Hij is ongemanierd. Sterker nog, hij is een *barbaar*.
3. Een ander patroon dat past bij intensiveerders is de metalinguïstische negatie, die ook een versterking geeft:
    - Hij is niet (gewoon) ongemanierd, hij is een barbaar.
    - Hij praat niet (gewoon) veel, hij is een *blaaskaak*.
4. Ten slotte kunnen intensiveerders, zo ze gradeerbaar zijn, veelal niet gecombineerd worden met verzwakkers, maar wel met versterkers. Voor niet-intensiverende woorden zijn zowel verzwakkers als versterkers mogelijk:
    - ? Hij is een beetje een lul / Hij is een enorme lul
    - Hij is een beetje een onaangenaam mens / Kij is een enorm onaangenaam mens
    - ? Ik ben een beetje uitgeteld / Ik ben helemaal uitgeteld
    - Ik ben een beetje moe / Ik ben erg moe
5. De *I is erg N*-test. Je kunt sterke woorden definiëren met behulp van een minder sterke term (adjectief of adjectief + nomen) voorafgegaan door *erg*:
    - Een 'lul' is een *erg onaangename* man
    - Een 'genie' is een *erg intelligent* mens
    - 'Uitgeteld' is *erg moe*.
    - 'Fenomenaal' is *erg goed*.
    - 'Heerlijk is *erg lekker.*

3 Hoe zijn de intensiveerders verzameld?

Er zijn drie lijsten doorgenomen.

1. De RBN-lijsten met nomina, adjectiva 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 adjectiva in overweging genomen (*overheersing > overheersen, overheersend*);
    b. hetzelfde geldt voor sterke adjectiva (*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.

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'
- *Gepronconceerd*: je kunt goed zeggen 'enigszins gepronconceerd'
- *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

Woord	**Gebruikscontexten (*= niet versterkend)**
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.

Voorbeeld	**Alpino-label voor 'bijzonder'**	**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'*	
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'*	
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 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.); geen sterk woord is echter *mislukking* of *fout*.
- 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 (*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

Enzovoort.

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, contrast of graad (*bovenal, zelfs*)
G	Woorden die een intense graad aangeven (*apert, mordicus*)
GR	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 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*)
T	Bepalingen die aan tijd of tempo refereren (*vanouds, weldra*)
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:

- Combinaties met *geen* die 'niets' betekenen: *geen bal*, enz.
- Combinaties met *nog* 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