

Cursus systematisch terminologiebeheer

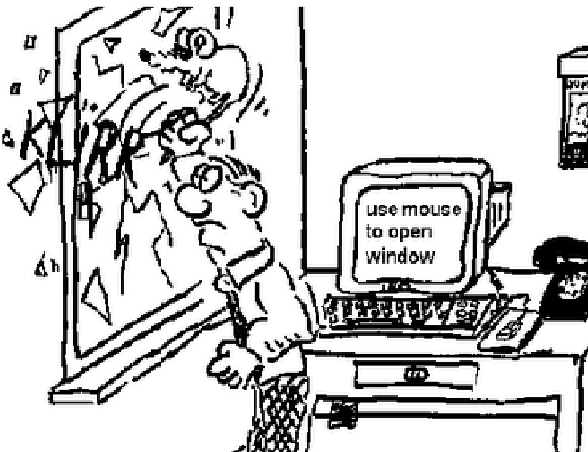
(door Attila Görög & Dr. Hennie van der Vliet, medewerkers van het Steunpunt Nederlandstalige Terminologie)



"Mouse, virus, firewall, why can't you computer people come up with your own words, rather than stealing ours?"

1. INTRODUCTIE

In deze webcursus gaan we in op systematisch terminologiebeheer. Systematisch terminologiebeheer staat hier in contrast met ad-hoeterminologie. Elders op NedTerm vindt u de webcursus ad-hoeterminologie, bedoeld voor wie éénmalig en snel wil weten wat een term betekent en hoe die vertaald kan worden. Systematisch terminologiebeheer is geschikt voor professionals, vertalers die niet eenmalig naar een term zoeken, maar die beseffen dat die term later weer zal opduiken. Zij voelen de noodzaak om de zaken systematischer aan te pakken.



Vertalers die aan systematisch terminologiebeheer doen, houden daarvoor meestal een terminologische databank bij. Een terminologische databank bevat belangrijke informatie over elke term (het vakgebied waarbinnen de term gebruikt wordt, definitie, relaties met andere termen, vertaling in verschillende talen, context, combinatoriek, bronnen etc). Deze informatie kan worden gebruikt bij het expliciteren van de keuze voor een bepaalde vertaling. De betrouwbaarheid

van de gebruikte bronnen en de definities van de bron- en doeltaaltermen spelen hierin een belangrijke rol. Als een vertaler zich d.m.v. systematisch terminologie-onderzoek heeft verdiept in een bepaald vakgebied en een termenbank heeft bijgehouden, kan hij zijn keuze voor een bepaalde term verantwoorden. Bovendien komt zo'n vertaler geloofwaardiger over dan een vertaler die puur op ad-hocbasis met terminologie omgaat.

Dit neemt natuurlijk niet weg dat ad-hoeterminologie een manier is en blijft om dagelijks met terminologie om te gaan. Zelfs in de meest recente termenbanken kunnen termen ontbreken. In dat geval kan ad-hoeterminologie een oplossing bieden. Maar ad-hoeterminologie leidt slechts tot voorlopige resultaten. Daarom is een systematische aanpak op de lange duur efficiënter en betrouwbaarder. De vraag blijft: hoe kunnen vertalers hier tijd voor vrijmaken? En waar kunnen vertalers hier meer informatie over krijgen?

De eerste vraag ligt bij de vertalers. Wie regelmatig teksten uit een specifiek vakgebied vertaalt, zal de investering in tijd snel terugverdienen in de vorm van hogere efficiëntie en kwaliteit. Aan de tweede vraag probeert het Steunpunt invulling te geven met deze cursus systematisch terminologiebeheer.

2. WAT IS TERMINOLOGIE?

Terminologie beschrijft de taal die specifiek is voor een bepaald vakgebied. In zo'n vakgebied bestaan vele begrippen, entiteiten, acties, handelingen enzovoort, en die worden met een woord, een groep woorden, eigennamen of een afkorting aangeduid. Dat zijn de termen.



2.1 Termen en concepten

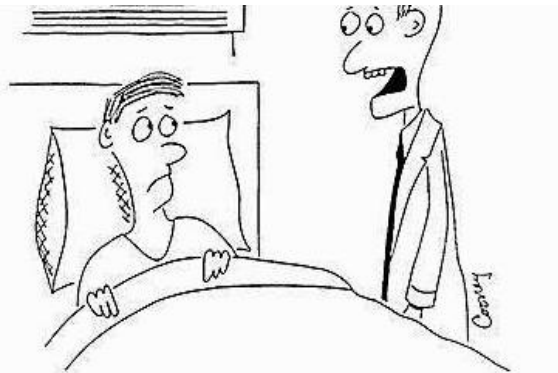
Termen

Termen (woorden, groepen van woorden, eigennamen of afkortingen) worden gebruikt om concepten van een specifiek vakgebied mee uit te drukken. Een term wordt daarom gedefinieerd als **linguïstische expressie van een concept uit een specifiek domein**. Termen worden per definitie gebruikt door kenners, door vakspecialisten. Een term is de verwoording van een brokje vakkennis. Een woord is géén term als het:

- niet door vakmensen, maar alleen door leken wordt gebruikt

(zoals *kleuterjuf*) en
 - verwijst naar een begrip dat niet specifiek is voor het domein
 (zoals *schoolplein*).

In beide gevallen ontbreekt natuurlijk de noodzaak om de woorden op te nemen in een terminologisch bestand, ze staan in elk goed vertaalwoordenboek.



"I've never been very big on medical terminology so just let me say you're really sick."

Concepten

De brokjes vakkennis, de begrippen, entiteiten, acties etc., zijn de concepten. Een concept moet inhoudelijk gefixeerd zijn en onderdeel uitmaken van de conceptuele structuur van het te beschrijven domein. Met andere woorden, concepten moeten inhoudelijk duidelijk afgebakend zijn, specifiek zijn voor een vakgebied en tezamen in onderlinge relatie een inhoudelijke beschrijving van dat vakgebied vormen.

2.2 Enkele voorbeelden

Krachtens is een woord dat typisch voorkomt in overheidsteksten en niet zomaar in alledaags taalgebruik. Toch is het geen term, want het is geen uitdrukking van een linguïstisch gefixeerd concept. Het is een functiewoord (met grammaticale functie) en vormt géén vaste verbinding tussen een concept en een talige vorm.

Voorwaarts kan een gewoon woord, uit het gewone taalgebruik zijn ("*moedig ging hij voorwaarts*") maar binnen het domein van de paardensport is het een term, die gebruikt wordt voor een bepaalde 'houding' van het paard. Het woord "*domein*" is hier trouwens ook een goed voorbeeld van. In het dagelijks taalgebruik heeft "*domein*" een heel brede, algemene betekenis en is het zeker geen term. In het vakgebied van de taalwetenschap functioneert het woord echter als term.

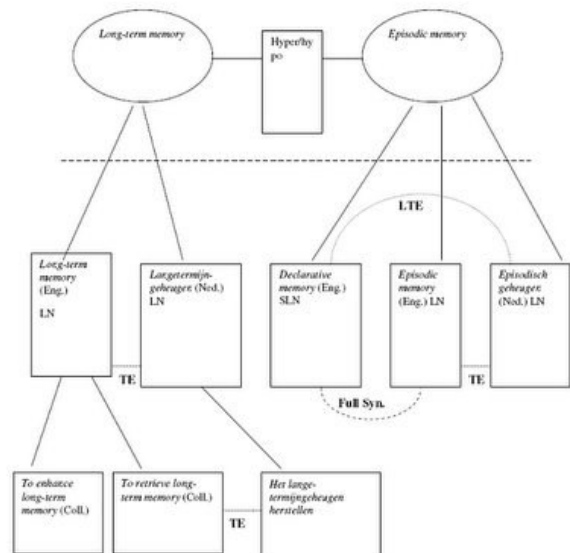
Kleuterjuf is zoals gezegd géén term, want het is weliswaar de linguïstische expressie van een concept uit een specifiek domein, maar deze benaming wordt alleen gebruikt door leken, niet door de vakspecialisten zelf (die zullen "*kleuterleidster*" zeggen). Hetzelfde geldt voor *links* en *rechts* in maritiem taalgebruik; de termen *stuurboord* en *bakboord* verwoorden verwante, zij het specifiekere uitgewerkte concepten.



2.3 Relatie tussen termen en concepten

Onderstaand voorbeeld is een schematische weergave van een fragment van een datamodel voor een terminologische databank. De concepten en termen zijn afkomstig uit het vakgebied van cognitieve psychologie en daarbinnen uit het specialisme dat zich bezighoudt met geheugenprocessen.

Deze figuur laat zien hoe concepten (bovenste niveau, taalonafhankelijk) gekoppeld zijn aan termen (in het midden, taalafhankelijk) en collocaties (onderste niveau, sterk taalafhankelijk). Collocaties zijn veel voorkomende woordcombinaties. Ze laten zien hoe de termen daadwerkelijk in levende vaktaal gebruikt worden.



- LN = Gelexicaliseerd Zelfstandig Naamwoord
- SLN = Half-gelexicaliseerd Zelfstandig Naamwoord
- Coll. = Collocatie
- Full Syn. = Volledige Synoniem (Variant)
- TE = Vertaalequivalent
- LTE = Bepaalde Vertaalequivalent

3. WERKMETHODE

De werkmethode in terminologiebeheer wordt bepaald op basis van de doelen en de beschikbare bronnen. Men moet een keuze maken tussen ad hoc- en systematisch onderzoek. Ad-hoconderzoek is van toepassing wanneer er een probleem snel moet worden opgelost tijdens de vertaling. Systematisch onderzoek geniet dan weer de voorkeur wanneer alle terminologie van een vakgebied moet worden opgesteld. Terminologische verkenning van een tekst situeert zich tussen beiden.



Ad-hoconderzoek leidt niet altijd tot bevredigende resultaten. De kans op fouten is relatief hoog omdat er meestal geen tijd is voor een conceptuele benadering. Een voorbeeld: juridische stukken van overheidsinstanties kunt u niet goed vertalen zonder dat u het juridische systeem in de brontaal en in de doeltaal bestudeert. (Voor een uitgebreide beschrijving van de ad-hoconderzoek zie onze Webcursus ad-hoconterminologie: http://taalunieversum.org/taal/terminologie/webcursus_ad-hoconterminologie/)

Systematisch onderzoek geeft de meest bevredigende resultaten omdat volgens deze methode de terminologie van een heel vakgebied of subvakgebied kan worden beschreven. Door ook expliciet de relaties tussen de verschillende concepten van het vakgebied te beschrijven, ontstaat een samenhangend beeld van de gehele terminologie van het vakgebied. Volgens de CEOV Aanbevelingen voor terminologie (http://taalunieversum.org/taal/terminologie/docs/CEOV_aanbevelingen.pdf) "in vergelijking met ad-hoconderzoek is thematisch onderzoek van betere kwaliteit zonder dat het meer werk vraagt."

In wat volgt beschrijven wij de vier stappen van onze werkmethode inclusief voorbeelden en informatie over tools.

3.1 VAKGEBIEDEN AFBAKENEN

3.1.1 Voorbeeld: afbakenen van een vakgebied

3.1.2 Informatiebronnen

3.1.3 CEOV aanbevelingen

3.2 DATAMODEL

3.3 CORPUS & TERMEXTRACTIE

3.3.1 Ééntalig

3.3.1.1 Verzameling teksten (ééntalig)

3.3.1.2 Tekstanalyse & Termextractie (ééntalig)

3.3.2 Meertalig

3.3.2.1 Verzameling teksten (meertalig)

3.3.2.2 Tekstanalyse & Termextractie (meertalig)

3.4 APPLICATIE & BEHEER

3.4.1 Definities

3.4.2 Termenbanken inrichten en bijhouden

3.1 VAKGEBIEDEN AFBAKENEN

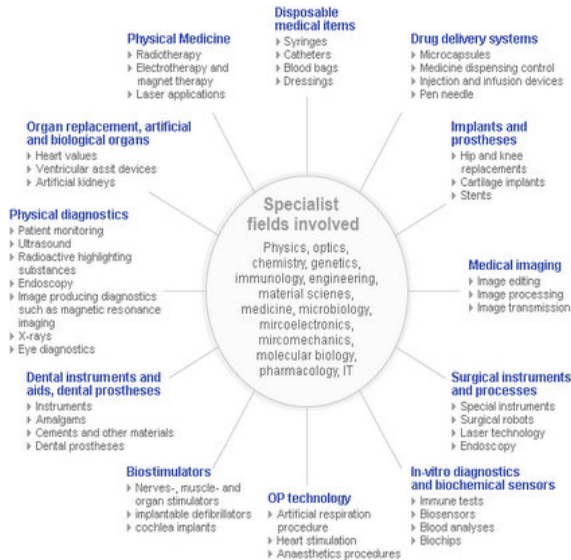
De eerste stap naar systematisch terminologiebeheer is het afbakenen van het te beschrijven vakgebied. Dat houdt de volgende taken in: 1.) het verkennen van het vakgebied d.m.v. inlezen 2.) het bepalen van basisconcepten en de structuur van het vakgebied.

Het is vooral belangrijk dat u d.m.v. vakliteratuur een goed inzicht krijgt in de structuur van het vakgebied. Voor het bepalen van de structuur en positie van het vakgebied in vergelijking met andere vakgebieden is het noodzakelijk dat u zich in het vakgebied inleest.

Om veelvoorkomende termen te vinden, te begrijpen en informatie te verzamelen kunt u gebruik maken van encyclopedieën, studieboeken, algemene werken over het onderwerp, vakbladen, thesauri, Internetbronnen ([Wikipedia](#), [online fora](#), [informatieve webpagina's](#), [webcursussen](#), [DMOZ](#)), [website van uw klant](#) etc.

Daarnaast kunt u het classificatiesysteem UDC (Universele Decimale Classificatie) raadplegen. De UDC is een internationaal indelings- en classificatieschema met honderdduizenden begrippen. Voor meer informatie over dit systeem, ga naar: http://nl.wikipedia.org/wiki/UDC_%28classificatiesysteem%29 Om het systeem als drietalige lijst (Nederlands, Engels en Frans) in pdf te downloaden klik hier: <http://www.vub.ac.be/BIBLIO/pdf/udc.pdf>

Bij het afbakenen van een vakgebied is het handig om de inhoudelijke structuur in de vorm van een **boomdiagram** of op andere overzichtelijke manier in kaart te brengen. De namen van concepten (vaak basistermen) kunnen worden gebruikt voor het classificeren van de termen en het structureren van de termenbank.



3.1.1 Voorbeeld: afbakenen van een vakgebied

Stelt u zich voor dat u zich wilt specialiseren in het vakgebied van de *nieuwe media* (of *interactieve media*). In [Wikipedia](http://nl.wikipedia.org/wiki/Categorie:Media) vinden we de volgende informatie over dit vakgebied:

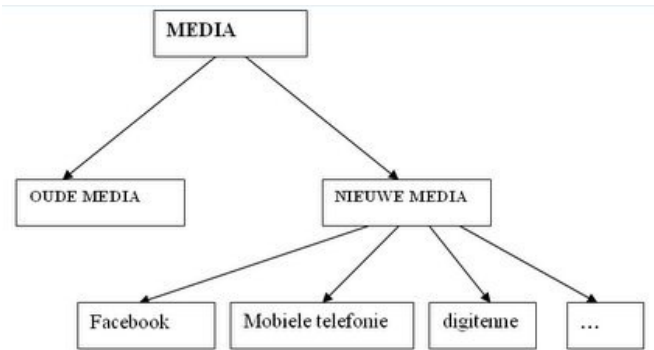
Nieuwe **media** is een manier om media in te delen. Nieuwe media komen dan tegenover oude media te staan. Het gebruik van de term varieert. Tegenwoordig worden vooral de **digitale media** bedoeld met de term nieuwe media. In die zin hoort een **mobiele telefoon** bij de nieuwe media, terwijl een analoge telefoon bij de 'oude' media hoort. Het **Internet**, **videogames**, **computers**, **digitale film**, **virtual reality**, **digitale fotografie**, **mobiele telefonie** en meer digitale media vallen hier dan onder. Onder de 'oude' media verstaan we dan traditionele film, televisie, pers en fotografie [etc.] http://nl.wikipedia.org/wiki/Nieuwe_media

Uit bovenstaande kunnen we concluderen dat het domein *nieuwe media* (net als andere domeinen) geen duidelijke grenzen kent. Dat geeft niet. Ons doel in deze eerste fase is het kennismaken met het vakgebied, met de gerelateerde andere domeinen en subgebieden en het maken van een basis voor de conceptuele structuur voor onze toekomstige termenbank.



Uit de informatie in Wikipedia blijkt ook dat *nieuwe media* een subdomein vormt van *media*, dat in totaal 26 subcategorieën en 132 ingangen bevat in Wikipedia. Deze 132 ingangen zijn in de meeste gevallen vaktermen met definities. (Zie: <http://nl.wikipedia.org/wiki/Categorie:Media>)

Op basis van bovenstaande informatie kunnen we een eerste schets maken van het vakgebied van de *nieuwe media* en van de plaats die het inneemt binnen het vakgebied *media*. Bovendien hebben we een groot aantal basistermen (zie vetgedrukt) die relevant zijn voor het vakgebied en kunnen worden gebruikt voor het opsporen van verdere vaktermen voor onze termenbank.



3.1.2 Informatiebronnen

Welke bronnen kunnen worden gebruikt voor het verduidelijken van concepten, voor het bouwen van tekstcorpora en voor het extraheren van termen?

Als eerste regel mogen wij stellen dat algemene woordenboeken in de meeste gevallen geen goede bronnen zijn als het om vaktaal gaat. Woordenboeken zijn gemaakt op basis van frequentieelstn uit algemene taalcorpora en bevatten als gevolg weinig vaktermen. Als er wel vaktermen zijn opgenomen zijn deze vaak niet correct gedefinieerd. Dat lijkt vreemd, maar zoekt u maar eens in een goed algemeen woordenboek naar vaktermen waarmee u inhoudelijk bekend bent. Soms zult u tevreden zijn, maar vaak ook niet! De alfabetische structuur van algemene woordenboeken is ook minder geschikt om thematisch te zoeken naar termen en concepten.

Welke bronnen dan wel?

Echte vaktalige bronnen zoals studieboeken, vakbladen en begrippenlijsten met definitie zijn veel betrouwbaarder en bevatten meer termen dan een algemeen woordenboek. Als u vaktalige bronnen voor handen heeft, kunt u bijvoorbeeld zoeken naar passages waarin de termen worden uitgelegd. Zulke passages worden **contextuele definities** genoemd. Goede contextuele definities zijn geschreven door vakexperts, zijn recent en gezaghebbend.

Als u noteert uit welke bronnen u uw contextuele definities haalt, kunt u anderen laten zien hoe betrouwbaar ze zijn. Zo kan een ander van uw werk profiteren. En u van dat van een ander. En u profiteert langer van uw eigen werk, want u weet zelf na een jaar nog steeds wat uw bronnen waren.



"OKAY, I'VE PUT IT IN THE RECYCLE BIN, NOW WHAT?"

3.1.3 CEOV aanbevelingen

Nog enkele tips uit *CEOV Aanbevelingen voor Terminologie* wat betreft de betrouwbaarheid van bronnen:

- een wetenschappelijke en technische uitgave is meestal betrouwbaarder dan een algemene uitgave;
- een wetenschappelijke en technische uitgave is betrouwbaarder in de brontaal dan in vertaling;
- een bericht in een vakblad is meestal betrouwbaarder dan een artikel over hetzelfde onderwerp in een dag- of maandblad;
- een normatieve officiële tekst is betrouwbaarder en meer bindend dan een niet normatieve officiële tekst;
- een wetenschappelijke en technische uitgave die alleen maar op het vakgebied van de te behandelen termen en begrippen is gericht, is betrouwbaarder dan een vergelijkbare uitgave die dat vakgebied enkel oppervlakkig aanraakt;
- de auteurs van vakteksten zijn meer geloofwaardig wanneer ze zich in hun hoofdtaal uitdrukken;
- informatie die door verschillende en onafhankelijke bronnen wordt bevestigd, biedt meer zekerheid.

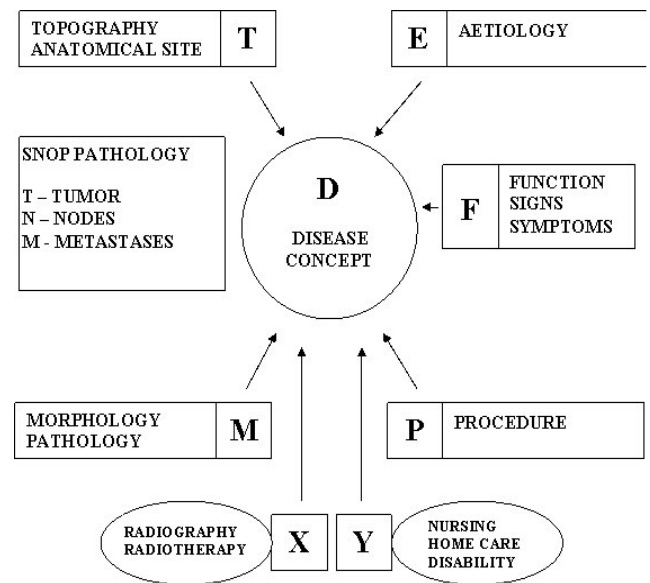
(Bron: [CEOV Aanbevelingen voor terminologie](#))



3.2 DATAMODEL

Voor u begint met termextractie is het belangrijk dat u een aantal beslissingen neemt m.b.t. het toekomstige ontwerp van de termenbank. In deze ontwerpfase moet er gekeken worden welke informatie en op welke manier wordt opgeslagen in de terminologische databank.

Er zijn in de afgelopen decennia talloze werken verschenen over de ideale termenbank en over wat voor informatie termenbanken moeten bevatten. Vertalers, vertaalbureaus en onderwijsinstellingen maken tot op heden gebruik van de meest simpele versie van termenbanken (Excel-lijsten of Access-bestanden) tot complexere oplossingen (WebWordSystem, I-term, Multiterm etc). Andere software voor terminologiebeheer vindt u hier: <http://taalunieversum.org/taal/terminologie/terminologiemangement/>.



Wij zijn van mening dat een goed bruikbare termenbank in het ideale geval minimaal de onderstaande velden bevat:

- conceptuele velden (= *conceptueel niveau*)
- bron- en doeltaaltermen en varianten (= *term niveau*)
- bronvermeldingen en definities en/of contexten en/of voorbeelden eventueel collocaties en grammaticale informatie zoals woordsoort, meervoud etc. (= *attribuut niveau*)

Bovendien is het belangrijk dat de termenbank

- op alle niveaus/ velden doorzoekbaar is en dat
- de data in vertaalsoftware kan worden gebruikt (d.m.v. import/export functies in de bekende dataformaten).

Dus om aan de eisen van systematisch terminologiebeheer te kunnen voldoen, dient een termenbank bovenstaande velden en functionaliteiten bevatten, of het nou een heel complex systeem is of een simpel database software.

Een voorbeeld van een terminologische ingang:

Term:	appreciatie
Domein:	macro-economie
Subdomein:	beurs
concepttype:	financiële proces
woordklasse:	zn
gram:	telbaar
definitie-1	Koersstijging van een valuta door vraag en aanbod. -> bron-1: <i>Economie</i> 1500 termen van A tot Z
definitie-2	ander woord voor waardeinstijging. Wordt gebruikt in de valutahandel om de waardeinstijging van de ene valuta ten opzichte van een andere aan te geven. Tegenovergesteld van depreciatie. -> bron-2: <i>Financieel Woordenboek</i>
definitie-3	Een appreciatie is de waardeinstijging van de ene munt ten opzichte van een andere. Een beweging in een wisselkoers resulteert erin dat men met een bepaalde munt meer van een andere geldeenheid kan kopen, de eerste munt is dus geapprecieerd. De term appreciatie is het tegenovergestelde van depreciatie, en wordt gebruikt bij veranderingen van flexibele wisselkoersen. -> bron-2: http://www.morningstar.nl/nl/news/article.aspx?articleid=76126&categoryid=488&lang=nl-NL&vabdfrom
voorbeeld:	Als bijvoorbeeld vorige week een euro 90 dollarcent kostte, en deze week een euro 95 dollarcent waard is, is de waarde van de euro ten opzichte van de dollar geapprecieerd. -> bron-2: http://www.morningstar.nl/nl/news/article.aspx?articleid=76126&categoryid=488&lang=nl-NL&vabdfrom
collocaaties:	forse appreciatie; opmerkelijke appreciatie; lichte appreciatie, appreciatie van X munt(en); appreciatie van X valuta; appreciatie van X koers, een appreciatie van X %, X % appreciatie; appreciatie van X valuta tegenover Y valuta; appreciatie van X valuta ten opzichte van Y valuta.

3.3 CORPUS & TERMEXTRACTIE

Na de fase van domeinafbakening en het ontwerpen van het datamodel kunt u beginnen met het verzamelen van vakteksten om een tekstcorpus te bouwen.

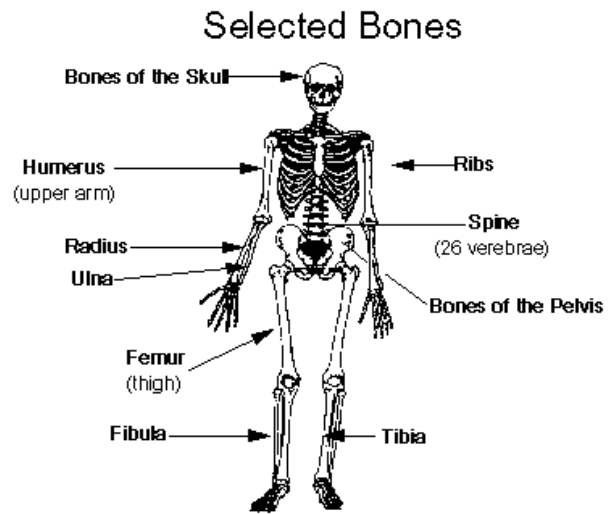
Het verzamelen van teksten begint in feite al in de eerste fase, tijdens het kennismaken met het geselecteerde vakgebied. Het doel van het aanleggen van een tekstcorpus is tweeledig: teksten bevatten niet alleen informatie over het domein zelf maar ook vaktermen, voorbeelden, definities en contexten.

Zowel de feitelijke informatie over het domein als de talige informatie kunnen worden gebruikt bij het verzamelen van termen en het maken van definities voor de toekomstige termenbank.

Afhankelijk van uw doelen kunt u één- of meertalig termen extraheren. Als u termen in meerdere talen wilt extraheren, heeft u voor elke taal afzonderlijk een verzameling teksten nodig. Een corpus dat teksten bevat uit hetzelfde vakgebied in verschillende talen wordt als **vergelijkbaar corpus** aangeduid. Tekstverzamelingen waarbij bronteksten en vertalingen worden gealigneerd vormen een **parallel corpus** (ook wel een **vertaalgeheugen** genoemd).

Meertalig terminologiebeheer wordt gedreven door dezelfde principes als ééntalig terminologiebeheer. Tijdens meertalig terminologiebeheer worden vaak ééntalige bronnen

(terminologieverzamelingen, documenten, glossaria etc.) in verschillende talen aan elkaar gekoppeld en als multilinguale databanken aangeboden.

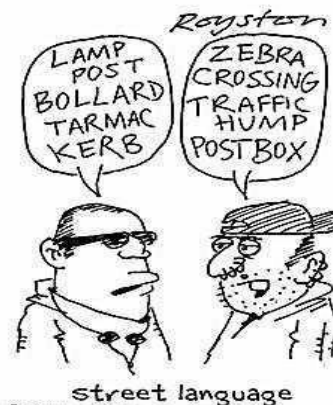


3.3.1 Ééntalig

Ééntalige (of monolinguale) termenlijsten en termenbanken kunnen worden gebruikt voor verschillende doeleinden. Een ééntalige termenverzameling is uitermate geschikt voor **SEO** (search engine optimization = het optimaliseren van de vindbaarheid van een site door zoekmachines), voor het bouwen van **thesauri** en voor het thematisch aanbieden van informatie op websites (door bijvoorbeeld overheidsinstanties).

Daarnaast kunnen zulke verzamelingen worden gebruikt door vakexperts binnen een bepaald vakgebied, voor kennismanagement doeleinden, voor het bestuderen van een vakgebied en als ondersteuning voor studieboeken.

Niet meteen voor de hand liggend maar u kunt ééntalige terminologie ook inzetten tijdens het vertalen. U kunt vaktermen uit de brontaal beter begrijpen m.b.v. een ééntalige termenlijst met definities (ook wel 'begrippenlijst' genoemd) en d.m.v. een termenlijst in de doeltaal kunt u controleren of de door u gekozen vertaal-equivalent inderdaad hetzelfde concept aanduidt.



3.3.1.1 Verzameling teksten (ééntalig)

Een goede Engelstalige site met een cursus over het maken van corpora vindt u [hier](#). In de Appendix van deze online cursus geeft de schrijver, John Sinclair, **tips voor het aanleggen van corpora**. De belangrijkste hiervan zijn:

1. Een corpus blijft altijd indicatief (en nooit definitief) dus u moet niet proberen naar volledigheid te streven! Bovendien maakt teveel tekstmateriaal de meeste taalsoftware traag. Het is beter om de corpusgrootte beperkt te houden maar wel teksten van zo divers mogelijke bronnen op te nemen.
2. Als u voldoende teksten heeft verzamelt, kunt u beginnen met het onderzoeken van uw corpus m.b.v. computerprogramma's. Maar voordat u daaraan begint kunt u het beste **een back-up maken van de teksten in het oorspronkelijke formaat**.
3. Vervolgens converteert u de teksten in 'plain text formaat' (.TXT, het liefst in UTF8) m.b.v. bijvoorbeeld een texteditor.
4. Om later de bron van de tekstfragmenten te kunnen achterhalen dient u gebruik te maken van **bronvermeldingen**.

Er bestaan verschillende online tools en desktop applicaties voor het verzamelen van teksten en het bouwen van tekstcorpora (Zie hieronder).

and to make it available to those developing corpora today. The modest aim is to build on some previous work in developing provisional de facto standards for the issue of standardization in developing tools for corpus annotation, and especially for dialogue annotation, developing a workbench and an evaluation of some of the issues involved in developing a spoken language corpus, with a reminder corpus builders to stop developing the corpus. While it is important to use at some stage of building a linguistic corpus. Little or no knowledge of a ready-made or can guess about its linguistic detail. Ideally a corpus should be a practice of adding interpretative linguistic information to a corpus. For example, the fact is that linguistic annotation cannot be done a priori other than what their meaning is, in linguistic terms. As an example, I have already had practice for different levels of linguistic annotation. The main message here is significant as any of its intrinsic linguistic properties, if indeed the two can be that address a great variety of linguistic issues ranging from morphological. Digital resources, particularly linguistic corpora, are designed to serve and be available in those developing corpora today. The modest aim of this Guide is to be called internal criteria. Corpora should be designed and constructed they are chosen. Since electronic corpora became possible, linguists have been tagging the 'Urwon Family' of corpora (consisting of the Urwon Corpus, 10 resources, particularly linguistic corpora, are designed to serve many different purposes. In creating and tagging corpora, particularly large ones assembled in this way and need to imply that all corpora should conform to LEXES standards. In English language corpora still dominate the field of corpus linguistics. It is usually hoped that corpora will be made available for other languages and widely used for language corpora, it is often trivial to migrate from old corpora and not appropriate for language corpora (see <https://www.allm.org/standards>). In one of the earliest specialised corpora Xordash: Hoe's corpus of textbooks

Tools voor het verzamelen van teksten

WebReaper

WebReaper is een gratis [webcrawler](#) (of webspider). Met behulp van een crawler kunt u veel corpusteksten gratis, snel en met minimale moeite verzamelen en overzetten naar uw harde schijf.

U hoeft alleen een URL in te voeren of een filter in te stellen en WebReaper verzamelt automatisch de relevante pagina's van het internet.

Webcorp

Een online tool waarmee u het internet kunt doorzoeken naar teksten die bepaalde zoekwoorden bevatten. Bijvoorbeeld als u zoekt op 'virtual reality' krijgt u enkele honderden tekstfragmenten en websites die ook andere (gerelateerde) termen bevatten zoals 'mass media', 'hypermedia', 'cyberspace' etc. Deze teksten kunt u vervolgens gebruiken voor het maken van een vakcorpus. (<http://www.webcorp.org.uk/>)

DMOZ

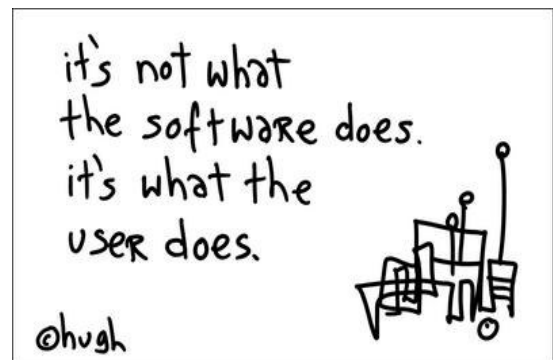
Een andere mogelijkheid is het gebruiken van DMOZ (Open Directory Project) waarin webpagina's thematisch zijn geordend. Hier vinden we 49 Nederlandse sites en 268 Engelstalige sites die zich op het vakgebied van *Nieuwe Media* richten. De teksten op deze sites kunnen dienen voor het opbouwen van een corpus van relevante teksten. (<http://www.dmoz.org/World/Nederlands/>)

TerminoWeb

Het TerminoWeb Platform biedt een experimenteel semi-automatisch corpusbouw- en analyse-instrument aan, bestemd voor terminologen en onderzoekers in de terminologie. TerminoWeb heeft in zijn huidige vorm drie belangrijke functies:

- het zoeken op het Internet naar kennisrijke documenten binnen een bepaald domein of vakgebied
- termextractie
- het analyseren van kennisrijke contexten

Helaas is TeminoWeb voorlopig alleen beschikbaar voor onderzoeksdoeleinden en alleen voor de talen Engels en Frans. (<http://terminoweb.iit.nrc.ca/TE.html>)



3.3.1.2 Tekstanalyse & Termextractie (ééntalig)

Termextractie is het computerondersteund proces waarbij op basis van een reeks elektronische teksten lijsten van potentieel interessante termen worden samengesteld. Ééntalige termenlijsten (ook wel "begrippenlijsten" genoemd) bevatten bovendien vaak definities, voorbeelden en andere relevante contextuele en conceptuele informatie.

Ééntalige termenlijsten met definities komen vaak voor in studieboeken of in vakliteratuur. Ook steeds meer bedrijven



"You haven't got dyslexia the instructions are in Polish."

3.3.2.1 Verzameling teksten (meertalig)

Als u een twee- of meertalige termenbank wilt maken, heeft u voor elke taal een verzameling teksten nodig van het desbetreffende vakgebied. In het ideale geval gebruikt u hiervoor uw eigen vertalingen/ vertaalgeheugens (**parallele corpora**) of die van een andere partij. Veel informatie op websites is beschikbaar in meerdere talen.

Deze informatie kunt u gebruiken om uw eigen parallel corpus te bouwen. Er worden ook steeds meer vertaalgeheugens van verschillende projecten en organisaties beschikbaar gesteld op het internet.

Een andere manier om een meertalig corpus te bouwen is het verzamelen van teksten uit hetzelfde vakgebied in verschillende talen (**vergelijkbare corpora**). U kunt dit op dezelfde manier doen als het verzamelen van ééntalige materiaal (Zie: 3.3.1.1).



Tools voor het verzamelen van teksten (meertalig)

Dit zijn enkele tools waarmee u naar meertalig informatie kunt zoeken op het Internet.

Linguee - <http://www.linguee.com/>

2lingual - <http://www.2lingual.com/>

TSM - <http://www.ttn.ch/TSM.ASP>

WeBiText - <http://www.webitext.com/bin/webitext.cgi>



Als u op zoek bent naar een overzicht van online beschikbare vertaalgeheugens of parallele corpora klik hier:

http://en.wikipedia.org/wiki/Parallel_text

3.3.2.2 Tekstanalyse & Termextractie (meertalig)

Meertalige termextractie is het computerondersteunde proces waarbij op basis van een reeks elektronische teksten meertalige lijsten van potentieel interessante termen worden samengesteld.

In elk vertaalproject speelt de identificatie van equivalenten voor gespecialiseerde termen een grote rol. Vakgebieden zoals informatietechnologieën, recht en geneeskunde hebben een grote hoeveelheid specifieke terminologie. En daar komt bij dat veel klanten voorkeur voor specifieke terminologie zullen hebben.

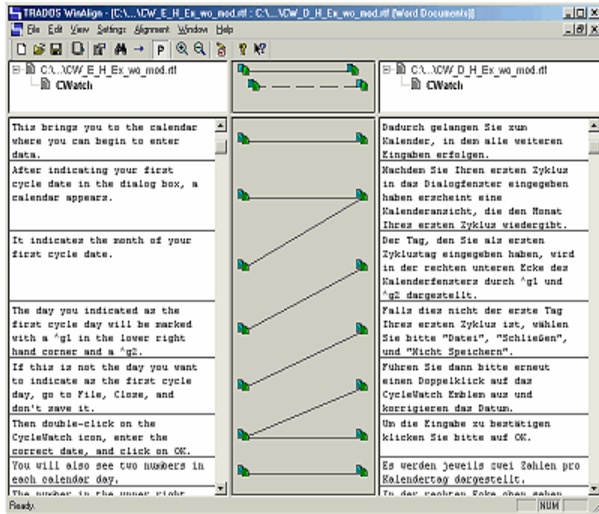
Verschiede taalanalysetools kunnen worden gebruikt bij het samenstellen van termenlijsten en het invullen van terminologische databanken. Dit gebeurt d.m.v. extractie van potentiële termen uit een selectie van elektronische teksten en hun vertalingen (parallele corpora).

Alignment

Vertalers en vertaalbureaus beschikken vaak over een grote hoeveelheid vertaald materiaal dat ze mogelijk willen omzetten in een vertaalgeheugen. **Alignmentprogramma's vergelijken een brontekst met de vertaling om de**

overeenkomende segmenten (zinnen, titels etc.) aan elkaar te koppelen en tot vertaaleenheden in een vertaalgeheugen samen te voegen.

Afhankelijk van de aard van de bron- en doeltaksten kan het zijn dat het alignmentprogramma een aantal foute verbindingen maakt (Zie onderstaande screenshot). **Het met elkaar verbinden van vertaalsegmenten is dan ook een semi-automatisch proces waarvan het resultaat gecontroleerd en eventueel verbeterd moet worden.**



Het verbeteren en corrigeren van de rauwe output is een intensieve en tijdrovende klus waarbij men over een flinke dosis taal- en softwarekennis moet beschikken. Naast commerciële programma's bestaan er tegenwoordig open source softwarepakketten en zelfs (gratis) on-line diensten.

Parallele en vergelijkbare corpora

Voor het maken van twee- of meertalige termenbanken heeft u een verzameling teksten nodig in twee- of meer talen. In het ideale geval heeft u bronteksten en hun vertalingen die u kunt aligneren met als resultaat uw eigen **parallele corpora** (of **vertaalgeheugens**).

Een parallel corpus is voor veel doeleinden uiterst nuttig. We vermelden het hier omdat het niet alleen (als elk tekstcorpus) informatie en termen in context bevat, maar die termen ook nog eens in context vertaalt. Een goed artikel met veel links naar (gratis) beschikbare parallele corpora vindt u hier: http://en.wikipedia.org/wiki/Parallel_text

Parallele corpora zijn lang niet altijd beschikbaar. Vertalers zullen toch zowel corpora in de bron- als in de doeltaal nodig hebben. Bij gebrek aan parallele corpora, of voor kwesties waarvoor het parallele corpus geen oplossing biedt, kunnen altijd monolinguale corpora van hetzelfde vakgebied in bron- en doeltaal geraadpleegd worden. De context moet dan aangeven in welke betekenis de term gebruikt wordt. Zulke tweetalige, niet-gealigneerde corpora worden ook wel eens **vergelijkbare corpora** genoemd.



Tools voor meertalige tekstanalyse en termextractie

Tekstanalyse



Xbench

ApSIC Xbench is gratis te downloaden op: www.apsic.com. Xbench biedt een gebruiksvriendelijke, uniforme weergave van de tweetalige informatie.

- inzicht in de terminologie van vertaalprojecten
- kwaliteitscontrole van terminologie in vertaalopdrachten
- ondersteunt veel inputformaten
- Ingebouwde browsermogelijkheden



Olifant

Met Olifant, een gratis editor voor *.tmx* (vertaalgeheugen) bestanden, kunt u uw vertaalgeheugens "trimmen". Dat wil zeggen filteren op basis van bepaalde criteria (aantal characters/ segmenten, vaak gebruikte segmenten etc.). Het "trimmen" van vertaalgeheugens is één van de snelste en makkelijkste manieren om termen te extraheren uit meertalige content. Hier geldt wel de regel: hoe meer tekst, hoe beter de resultaat. (<http://okapi.sourceforge.net/Release/Olifant/Help/>)

Alignment

HunAlign is een krachtig alignment software gratis te downloaden [hier](#). Dit programma werd o.a. ook gebruikt voor het maken van de [Europarl parallel corpus](#) van de Europese Parlement. (<http://mokk.bme.hu/resources/hunalign>)

AlignAssist is ook een gratis tool ontwikkeld door Ryan Ginstrom van Felix. In tegenstelling tot HunAlign heeft dit programma een simpele en gebruiksvriendelijke interface. Het is sinds kort mogelijk om alignments te exporteren naar het standaard *.tmx* formaat. Te downloaden hier: (<http://felix-cat.com/tools/align-assist/>).

YouAlign is een online dienst van softwareproducent Terminotix. De Beta-versie van YouAlign is gratis maar wel beperkt tot 5 alignments per dag met een maximum van

1MB per bestand. Het programma werkt met verschillende input- en outputformaten en is makkelijk in gebruik. Klik [hier](#) om YouAlign te gebruiken.

Voor meer online, open source en commerciële termextractors check:

http://en.wikipedia.org/wiki/Terminology_extraction
<http://taalunieversum.org/taal/terminologie/termextractie/>

Voor meer corpus analyse tools check:

<http://taalunieversum.org/taal/terminologie/corpus-analysetools/>



"My computer doesn't understand me!"

3.4 APPLICATIE & BEHEER

Na de domeinafbakening, het ontwerpen van het datamodel, het bouwen van een corpus en het extraheren van termen, kunt u beginnen met het inrichten en vullen van uw terminologische databank. Dit gebeurt d.m.v. het importeren van termen uit verschillende bestanden, tijdens het vertalen m.b.v. plug-ins of handmatig. Het datamodel d.w.z. de structuur en de velden van uw termenbank heeft niet alleen invloed op het proces van termextractie maar ook op de manier hoe u uw termenbank gaat inrichten.

Bij applicatie en beheer horen ook zaken als normalisatie en het regelmatig aanvullen en up-daten van termen, voorbeelden, definities etc. Uw termenbank is in zekere zin uw intellectuele eigendom waarin u al uw kennis over een domein (of verschillende domeinen) opslaat en beheert in de gewenste talen.

Dictionary


gouge [gouʒ]

noun

- a chisel with a concave blade, used in carpentry, sculpture, and surgery.
- an indentation or groove made by gouging.

verb [trans.]

- make (a groove, hole, or indentation) with or as if with a gouge : *the channel had been gouged out by the ebbing water.*
 - make a rough hole or indentation in (a surface), esp. so as to mar or disfigure it : *he had wielded the blade inexpertly, gouging the grass in several places.*
 - (**gouge something out**) cut or force something out roughly or brutally : *one of his eyes had been gouged out.*
- informal overcharge; swindle : *the airline ends up gouging the very passengers it is supposed to assist.*



gouge 1

3.4.1 Definities

Hoe schrijft u goede, beknopte definities?

- U zoekt eerst voldoende informatieve contextuele definities over de term (een stuk of 5 echt informatieve).
- U schrijft ze op met bronvermelding.
- Als u precies begrijpt wat het concept is, maakt u zelf een beschrijving waarin alle informatie van de door u verzamelde definities aanwezig is, een soort "super-definitie". Die moet voldoende uitgebreid zijn, want dat is de beschrijving van wat het concept precies is. Daar baseert u zich straks op bij het vertalen.
- Zo krijgt u een conceptbeschrijving, met als bewijs een aantal contextuele definities en bronnen.
- Dit herhaalt u voor elke taal, en steeds controleert u of uw concepten in die talen volstrekt identiek zijn. Dat kan, aan de hand van uw "super-definitie".

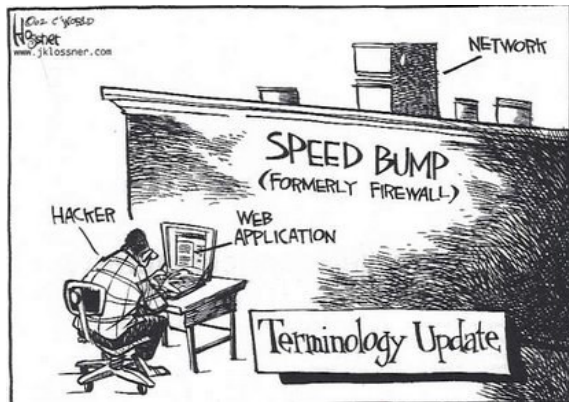
3.4.2 Termenbanken inrichten en bijhouden

Termen importeren

Als u klaar bent met het verzamelen van termen kunt u de verschillende data via een .csv, .txt of .xls bestand importeren naar uw termenbank. De meest gebruikelijke manier is het maken van een excel sheet waarin de verschillende kolommen verschillende informatie bevatten. De meeste terminologische databanken ondersteunen .xls of .csv als import-formaten (Denk aan MultiTerm of aan simpele database-software zoals MS Acces, Filemaker Pro of de gratis en online Grubba.net).

Als u meerdere excel sheets heeft met verschillende informatie en datavelden kunt u deze samenvoegen in uw termenbank. Tijdens het importeren hebben de meeste termenbanken namelijk de mogelijkheid om termen en de bijbehorende velden te synchroniseren. Op die manier kunt u bijvoorbeeld aan een tweetalige termenbank termen in andere

talen toevoegen of uw simpele tweetalige termenlijsten met definities, voorbeelden etc. verrijken m.b.v. begrippenlijsten of ééntalige termenbanken.



Termenbanken bijhouden

Het is belangrijk dat u uw termenbank regelmatig opschooft en up-to-date houdt. U kunt bijvoorbeeld de inhoud van uw termenbank filteren op termen die nog geen vertaling of juist meerdere vertalingen hebben. Voor het bijhouden van uw termenbank heeft u ook de bronvermeldingen hard nodig. U kunt een uitdraai maken van alle termen met bronnen die bijvoorbeeld minstens vijf jaar oud zijn.

Daarnaast is het van belang dat iedereen die het bestand gebruikt kritisch is op wat er gevonden wordt. In de meest geavanceerde setting kan een termenbank online en als multi-user applicatie worden gebruikt. Gebruikers (bijvoorbeeld vertalers) hebben rechten om in te loggen, termen on-line op te zoeken, data te importeren of exporteren en terminologische ingangen (of termfiches) aan te passen. Het is belangrijk dat de terminoloog of de project manager die binnen een organisatie de terminologische databank(en) beheert, de veranderingen kan traceren op basis van datum en gebruikersinformatie, en de wijzigingen kan accepteren of negeren.

Beheer & Normalisatie

Bij terminologiebeheer hoort ook vaak normalisatie. Voor ieder vakgebied waarvoor normen bestaan wordt ook de daarin gebruikte terminologie vastgelegd. Zowel in een hoofdstuk *Termen en definities* in de norm zelf, als in aparte

terminologienormen. Het is van groot belang dat iedereen binnen een vakgebied dezelfde term gebruikt voor hetzelfde begrip, en weet welke definitie daarbij hoort.

Meer informatie over normalisatie als onderdeel van terminologiebeheer vindt u hier:

<http://taaluniversum.org/taal/terminologie/normalisatie/>

4. TOT SLOT

Systematisch terminologiebeheer, zoals in deze cursus beschreven wordt, kost veel tijd. Wij realiseren ons dat. Toch denken we dat deze cursus om een aantal redenen in een behoefte voorziet.

In de eerste plaats is met het opzetten van deze cursus aan vertaalbureaus gedacht. Vertaalbureaus werken op grotere schaal dan individuele vertalers en hebben nog meer baat bij systematisch terminologiebeheer. Deze cursus laat zien wat daar bij komt kijken. We hopen dat de cursus vertaalbureaus stimuleert om in kwaliteit te investeren. Die investering is noodzakelijk; het alternatief is kwalitatief slecht werk leveren voor een iets lagere prijs.

We dachten ook aan studenten van vertaalopleidingen en beginnende vertalers. Studenten kunnen met deze cursus kennismaken met een systematische werkwijze, ook al is het om praktische redenen misschien niet in alle gevallen mogelijk zo'n systematische aanpak in de beroepspraktijk volledig waar te maken. We weten dat sommige opleidingen gebruik maken van de ad-hoc cursus en we hopen dat ook deze cursus in het onderwijs gewaardeerd zal worden.

Natuurlijk is deze cursus ook voor de individuele vertaler bedoeld. We zijn ons ervan bewust dat systematisch terminologiebeheer zoals in deze cursus is beschreven een flinke investering in tijd is, die zich pas op wat langere termijn terugbetaalt. Maar u bent natuurlijk niet verplicht de cursus stap voor stap te volgen. U kunt zich ook laten inspireren, de manier waarop u met terminologie omgaat heroverwegen en bijstellen, u kunt overwegen om één of meer van de besproken tools in te zetten. In ieder geval biedt de cursus u een mogelijkheid om uw werkmethode tegen het licht te houden en eventueel aan te passen. Wij zouden zeer tevreden zijn als de cursus daarvoor geschikt is.

We hopen dat de beide webcursussen in een behoefte voorzien. Ze zijn met zorg gemaakt, maar uiteraard bevatten ze foutjes en andere onregelmatigheden. Als u die ontdekt, horen we dat graag van u. Ook andere kritiek is zeer welkom, zeker als die ertoe kan leiden dat de cursussen beter worden. Zelfs als u vernietigende kritiek hebt die op geen enkele manier kan bijdragen tot het verbeteren van de cursussen ('doek ze maar op!') dan helpt u ons door ons daarvan op de hoogte te stellen. We hopen er echter op dat we u met deze cursussen van dienst zijn en we horen graag wat uw ervaringen zijn.

