TERMINOLOGIE: OP HET SNIJVLAK VAN AMBACHT EN TECHNOLOGIE

Klaar Vanopstal, Lieve Macken, Els Lefever, Marjan Van de Kauter, Joost Buysschaert & Véronique Hoste

Terminologie is niet weg te denken uit een opleiding waarin taal centraal staat. In ieder vakgebied - en dan denken we zowel aan het medische, juridische en technische domein, als aan grammatica, literatuur, taaltechnologie, logica, recht - komt terminologie expliciet dan wel impliciet aan bod. Het is dan ook niet verwonderlijk dat er vanuit verschillende hoeken interesse bestaat voor de discipline. In een geglobaliseerde wereld maakt meertalig terminologiebeheer bovendien ook efficiënte communicatie mogelijk over de taalgrenzen heen.

Terminologie is de studie van termen en hun gebruik. Termen op hun beurt kunnen we omschrijven als woorden die gebruikt worden binnen een specifiek vakgebied. Terminologie kan vanuit verschillende perspectieven worden benaderd, bijvoorbeeld prescriptief versus descriptief. Hoewel het doel van beide aanpakken hetzelfde is, namelijk het opstellen van 'een' terminologie, is er toch een verschil in benadering. Waar men binnen de prescriptieve terminologie streeft naar een consensus rond het gebruik van een bepaalde set termen (standaardisatie), is descriptief terminologiewerk veeleer een reflectie over het gebruik van termen in een bepaald vakgebied, zonder daarbij aan enige vorm van regelgeving te doen.

Ook aan onze vakgroep werd terminologie reeds vanuit verschillende perspectieven bestudeerd. Enerzijds is er het puur manuele terminologische en terminografische werk, waarbij na een uitgebreide studie van de termen en hun gebruik termenbanken worden opgesteld die ook extra informatie bevatten, zoals

synoniemen, definities en contexten. Deze termenbanken kunnen op een min of meer ontologische manier worden gestructureerd, bijvoorbeeld met verwijzingen naar boven- en ondertermen. Anderzijds worden ook systemen ontwikkeld voor termextractie, die (voorlopig) enkel gericht zijn op het opstellen van zogenaamd 'platte' termenlijsten, zonder enige semantische informatie dus. Op de wip tussen deze invalshoeken zit het ABOP-project, een *authoring* tool die onder meer wetenschappelijke terminologie detecteert en, wanneer mogelijk, ook vervangt. Daarnaast werd binnen de vakgroep ook een doctoraatsonderzoek uitgevoerd naar het gebruik van terminologie, meer bepaald van een gecontroleerd vocabularium, in medische information retrieval.

In wat volgt geven we een overzicht van de terminologieprojecten en overige initiatieven die aan onze vakgroep al werden gerealiseerd, en andere die nog op til zijn.

Terminografie

Al sinds 1987 loopt het *MeSH Termbase Project*[1] (MeSH = Medical Subject Headings)[1-4], dat aanvankelijk gericht was op het creëren van een tweetalige medische termenbank (Engels-Nederlands), maar intussen ook is uitgebreid met een beperkt aantal Franse vertalingen. Het terminologische onderzoek wordt hoofdzakelijk uitgevoerd door masterstudenten, die in het kader van hun masterproef een selectie van termen uit de originele MeSH-thesaurus bespreken en vertalen. Het GenTerm-fiche, ontworpen door het Centrum voor Terminologie[2], wordt als leidraad gebruikt voor het terminografische werk. Ingevulde GenTerm-fiches worden in een databank bewaard die de studenten tijdens colleges kunnen raadplegen als ondersteuning bij medische vertalingen.

Studenten kunnen ook bijdragen leveren aan het *Farma-project*, een zijproject van MeSH Termbase waarin de focus ligt op terminologie

[1] http://www.cvt.ugent.be/mesh.htm

uit het vakgebied van de farmaceutica. Ook dit project spitst zich toe op Engels en Nederlands, al zijn er in 2012-'13 ook twee *Farma*-masterproeven met Italiaans ingediend.

Een ander terminologieproject dat gedragen wordt door studenten, is het *EDiCT*[3]*-project*[5] (Electronic Dictionary of Communication Technology). Communicatie wordt hier gezien als een breed domein, dat onder meer public relations, marketing, reclame en journalistiek omvat. Ook binnen dit project wordt het GenTerm-fiche gebruikt als basis voor het terminografische werk. Ondertussen werden al verschillende beknopte afgeleide versies gepubliceerd van deze termenbank (ELeCT, Electronic Lexicon of Communication Technology [6]).

Sinds 2011-'12 loopt er ook een project in het kader van de EMT-werking (European Master of Translation), waarbij studenten via hun masterproef bijdragen kunnen leveren aan de *IATE-termenbank* van de Europese Unie. Leden van het Centrum voor Terminologie aan de vakgroep Vertalen, Tolken en Communicatie geven overigens al sinds 2004 geregeld terminologietraining aan EU-vertalers.

Ook het *JuriGenT-project*[4] is gebaseerd op input van de studenten. Voor meer informatie over dit project verwijzen we graag naar de bijdrage van Patricia Vanden Bulcke en Carine De Groote aan deze bundel.

Dat ook tolken afnemers zijn van termenlijsten is genoegzaam bekend. Over de specifieke vereisten van een 'tolkenglossarium' is binnen de vakgroep al vaker van gedachten gewisseld en er zijn initiatieven in ontwikkeling.

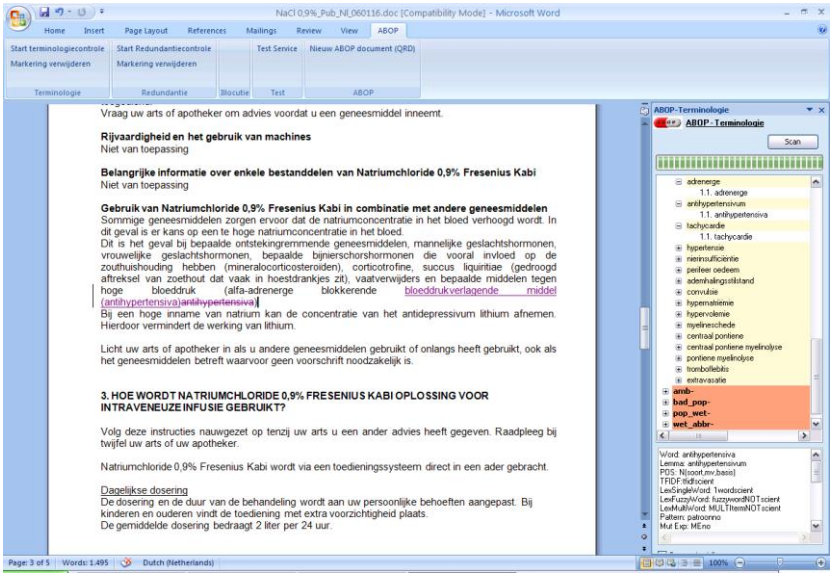
---

[2] http://www.cvt.ugent.be
[3] http://www.cvt.ugent.be/edict.htm
[4] http://www.cvt.ugent.be/jurigent.htm

Een pleidooi voor de erkenning van terminografisch werk als gedegen wetenschappelijk onderzoek vindt u in de bijdrage van Joost Buysschaert.

Terminologie in toepassingen

Het doel van het *ABOP*[5]*-project* (**A**utomatische **B**ijsluiter**Op**timalisatie)[7-8] van het Language and Translation Technology Team (LT3)[6] was om schrijvers en vertalers van bijsluiters te ondersteunen in het opstellen van medische bijsluiters. De leesbaarheid van bijsluiters wordt door verschillende factoren beïnvloed, waaronder het gebruik van al dan niet wetenschappelijke terminologie. De ABOP-software detecteert automatisch wetenschappelijke terminologie en vervangt die door populaire termen om zo de leesbaarheid van de bijsluiters te vergroten. Wanneer geen populaire term voorhanden is, wordt een korte begripsomschrijving toegevoegd.



[5] http://www.lt3.ugent.be/en/projects/abop/
[6] http://www.lt3.ugent.be/

Afbeelding 1: Schermafdruk van ABOP-toepassing in Word

Afbeelding 1 toont een schermafdruk van de tool, waarin een wetenschappelijke term (*antihypertensiva*) wordt vervangen door een korte begripsomschrijving (*bloeddrukverlagend middel*) met de wetenschappelijke term tussen haakjes. In het scherm rechts wordt extra informatie gegeven over de tekstuele kenmerken (bv. woordsoort, lemma, TF-IDF-score) op basis waarvan de term als wetenschappelijk geclassificeerd werd.

Het ABOP-project werd gerealiseerd in partnerschap met Artesis Antwerpen en een gebruikerscommissie van farmaceutische bedrijven alsook dienstverlenende bedrijven voor die farmaceutische sector. Het werd gefinancierd door het Tetra-fonds van het IWT.

In het doctoraatsonderzoek 'Impact of language skills and system experience on medical information retrieval'[9] werd terminologie in een andere toepassing bestudeerd: medische information retrieval. Dit interdisciplinaire onderzoek bestond uit een theoretische studie van terminologie enerzijds, en een empirisch onderzoek naar het gebruik van terminologie (MeSH) in PubMed anderzijds. Uit de empirische studie bleek onder meer dat het gebruik van MeSH ook voor Nederlandstaligen een nuttig hulpmiddel is bij het zoeken van medische informatie in PubMed, ook al is dit vocabularium voorlopig enkel beschikbaar in het Engels. Verder bleek dat niet alleen de kennis van het Engels een invloed had op de efficiëntie van een zoekopdracht, maar ook de ervaring met het zoeksysteem en met de specifieke terminologie.

Automatische terminologie-extractie

Waar in de terminografische projecten manueel gedetailleerde informatie wordt opgezocht voor termen, wordt bij termextractie getracht om delen van dit proces te automatiseren. Het resultaat hiervan is een platte termenlijst, die verder kan worden verrijkt via manueel opzoekingswerk.

Termextractieprojecten als TExSIS [10] en PSA[11] zijn ontstaan uit de nood aan eenduidige termenlijsten. Enerzijds hebben deze lijsten een descriptief karakter doordat ze gebaseerd zijn op schriftelijke communicatie (bv. jaarverslagen, handleidingen, etc.). Daarnaast heeft de aanpak in het PSA-project ook een prescriptief karakter, aangezien een duidelijk onderscheid wordt gemaakt tussen geprefereerde en niet-geprefereerde termen.

Waar TExSIS en PSA uitgingen van paralleltekster zullen twee toekomstige projecten vertrekken van vergelijkbare corpora (SCATE en ExPECT). Bij paralleltekster wordt het zoekveld voor vertalingen aanzienlijk verkleind door de expliciete bronzin-doelzinrelaties. De grootste uitdaging voor termextractie uit vergelijkbare corpora is dat net deze bronzin-doelzinrelaties ontbreken.

PSA

Het 'Étude doublons et synonymes thésaurus *APV-project'* was een samenwerking tussen LT3, Peugeot-Citroën en het vertaalbedrijf Telelingua. Het project had tot doel het terminologiegebruik in de handleidingen van automobielbedrijf PSA consistenter te maken. Dit werd bereikt door in de databanken die gebruikt worden voor het compileren van alle technische documentatie, semantische en morfologische varianten te vervangen door één bepaalde term (bv.: de termen *Leichtmetallscheibenrad*, *Leichtmetallrad*, *Alufelge*, *Aluminiumscheibenrad*, *Aluminium-Scheibenrad*, *Scheibenrad Aluminium* en *Leichtmetall-Scheibenrad* werden vervangen door *Metallscheibenrad*); en dit voor alle twintig talen in de portefeuille van PSA.

Om dit te kunnen verwezenlijken werden eerst twintig bilinguale termenlijsten op een automatische manier geëxtraheerd, telkens met het Frans als spiltaal. Vervolgens werd voor elke set van synonieme termen een voorkeursterm aangeduid door een werknemer van PSA. Uiteindelijk werden alle synoniemen dan door deze voorkeursterm vervangen in de databases van alle twintig ondersteunende talen,

waardoor uiteindelijk de woordenschat in de resulterende documentatie van PSA uniformer en eenduidiger werd.

TExSIS

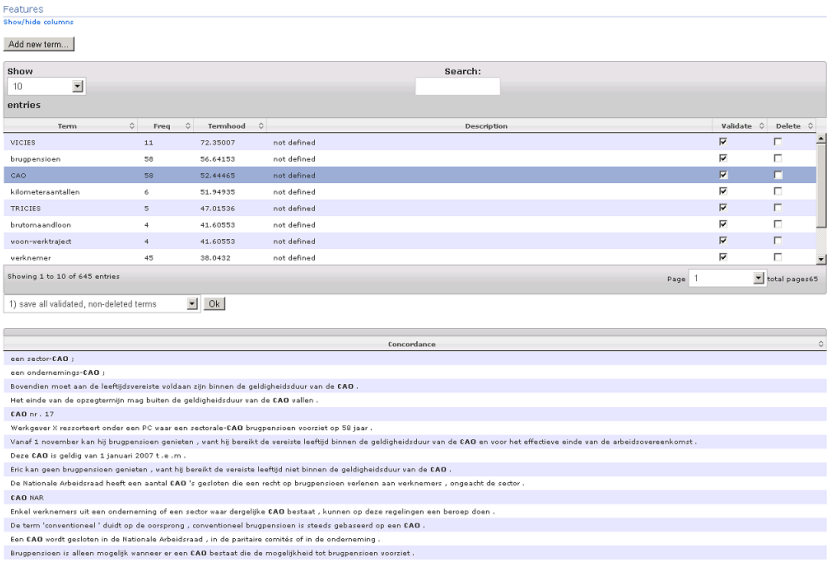
Wanneer we kijken naar een- of meertalige bedrijfsdocumenten zoals folders, jaarverslagen, online- en offline-ondersteuning en producthandleidingen, wordt het al snel duidelijk waarom een consequent gebruik van terminologie ook voor bedrijven belangrijk is. Om het eenduidig en consistent gebruik van terminologie in bedrijven te ondersteunen (denk maar aan machinevertaling, CAT en informatiemanagement) werd binnen het *TExSIS*[7]*-project* (**T**erminology **Ex**traction for **S**emantic **I**nteroperability and **S**tandardization) een extractiemodule ontwikkeld die zowel mono- als multilinguale termenlijsten extraheert. Hoewel het systeem op zich taalonafhankelijk werkt, werd in het kader van dit project gefocust op vier verschillende talen (Nederlands, Engels, Frans en Duits).

Onderstaande schermafdruk (afbeelding 2) toont de interface van de TExSIS-extractietool, waarin naast de term zelf ook informatie wordt gegeven over frequentie en termhood, een maat die aangeeft in welke mate een term met een bepaald domein wordt geassocieerd. De gebruiker kan hieraan zelf een definitie toevoegen. Onderaan op het scherm wordt iedere term in zijn context getoond, zodat vertalers een idee krijgen van het gebruik van de term.

TExSIS werd gefinancierd door het IWT in het kader van het Tetra-programma en werd ondersteund door een commissie van gebruikers uit de bedrijfswereld onder peterschap van de Nederlandse Taalunie. De gebruikerscommissie bestond uit vertegenwoordigers van vertaalbedrijven (CrossLanguage, Telelingua, Yamagata Europe), software- en IT-bedrijven (TextKernel, Mentoring

[7] http://www.lt3.ugent.be/en/projects/texsis/

Systems, Actonomy) en eindgebruikers uit verschillende sectoren (Jan De Nul, WTCB, SDWorx).



Afbeelding 2: de TExSIS-interface

SCATE

Het *SCATE- project* (**S**mart **C**omputer-**A**ided **T**ranslation **E**nvironment) is een onderzoeksproject waarvoor LT3 en andere Vlaamse universitaire partners (onderzoeksgroepen CCL, ESAT/PSI, LLIR en Thomas More van de Katholieke Universtiteit Leuven en de EDM onderzoeksgroep van de Universiteit Hasselt) de handen in elkaar slaan. Het project wordt gefinancierd door het IWT via het SBO-financieringskanaal voor strategisch basisonderzoek. Het hoofddoel van het project is de ontwikkeling van een geavanceerde vertaaltool die verschillende vertaalbronnen integreert en zo de efficiëntie van vertalers drastisch kan verbeteren. Tijdens het vertalen worden meer en betere vertaalsuggesties aangereikt, wat uiteindelijk resulteert in een hogere efficiëntie van het vertaalproces.

LT3 is binnen het SCATE-project onder andere verantwoordelijk voor de ontwikkeling van de terminologiecomponent, die meertalige

termenlijsten op een automatische manier extraheert uit vergelijkbare corpora. Daarbij wordt in eerste instantie een termenlijst uit de eentalige teksten afgeleid, waarna vervolgens de link gelegd wordt tussen de corresponderende termen in de verschillende talen.

ExPECT

In het *ExPECT-project* (**Ex**traction of **P**arallel **E**lements from **C**omparable **T**exts) wordt onderzocht in hoeverre termextractie op basis van vergelijkbare corpora kan worden ingezet ter ondersteuning van vertalers en tolken. We gaan hierbij uit van een situatie waarin er geen parallelle data voor het domein/taalpaar voorhanden zijn. Voor de extractie van bilinguale termenlijsten vertrekken we vanuit eerder ontwikkelde methodologieën en wordt nagegaan in welke mate die ook geëxtrapoleerd kunnen worden naar het Nederlands. In eerste instantie worden monolinguale termen geëxtraheerd met behulp van de TExSIS-tool, die in een volgende fase aan hun equivalent(en) in de andere taal zullen worden gelinkt door de contexten van kandidaat-vertalingen te vergelijken (de zogenaamde 'distributionele hypothese').

In de evaluatiefase zal de output op twee manieren beoordeeld worden: naast de (traditionele) berekening van *precision* en *recall* zal de tijdswinst bij en kwaliteit van vertalingen en tolkopdrachten met terminologische ondersteuning van het systeem worden geëvalueerd.

Zoals vermeld ligt de focus bij termextractie op het creëren van platte lijsten met termen. In de toekomst willen we meer de nadruk leggen op de automatische aanmaak van ontologieën, wat aan bod komt in het *MuST*[8]- (**Mu**ltilingual Corpora for the automatic **S**tructuring of

[8] http://www.lt3.ugent.be/en/projects/must/

Terms) en het *SentiFM*[9]*-project* (**Senti**ment Mining for **F**inancial **M**arkets).

Referenties

[1] Buysschaert, J. (1996). Terminografie met de computer: twee voorbeelden. In J. Buysschaert (Red.), *De computer van de Germanist* (pp. 11-24). Gent: BGG.

[2] Buysschaert, J. & Robberecht, P. (2001). Enkele informatiseringsaspecten van het MeSH-Vertaalproject. In W. Vandeweghe et al. (Red.), *Polyfonie. Opstellen voor Paul van Hauwermeiren* (pp. 55-64). Gent: Mercator Hogeschool.

[3] Buysschaert, J. (2006). The development of a MeSH-based biomedical termbase at Hogeschool Gent. In P. Zweigenbaum et al. (Eds.), LREC 2006 Satellite Workshop W08. Acquiring and representing multilingual, specialized lexicons: the case of biomedicine (pp. 39-43). Genova.

[4] Buysschaert, J. (2006). Exploiting an English-and-Dutch biomedical termbase: the search for an ideal format. *Equivalences,* 33, pp. 33-42.

[5] Buysschaert, J. (2009). Quantifying Anglicisms in French, German and Dutch business communication. In S. Slembrouck, M. Taverniers & M. Van Herreweghe (Eds.), *From will to well: Studies in linguistics offered to Anne-Marie Simon-Vandenbergen* (pp. 69-77). Gent: Academia Press.

[6] Buysschaert,J., Vanopstal K. & Kovács, L. (Eds.) (2012) ELeCT 3.2. Electronic lexicon of communication terminology. Gent, Communication & Cognition, Electronic publication.

[9] http://www.lt3.ugent.be/en/projects/sentifm/

[7] Delaere, I., Hoste, V., Peersman, C., Van Vaerenbergh, L., & Velaerts, P. (2009). ABOP, Automatic Optimization of Patient Information Leaflets. In S. Cardey (Ed.), *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages*. Presses Universitaires De Franche-Comté, Besançon, France.

[8] Hoste, V., Vanopstal, K., Lefever, E., & Delaere, I. (2010). Classification-based scientific term detection in patient information. *Terminology, 16*(1), 1-29.

[9] Vanopstal, K. (2013). Impact of language skills and system experience on medical information retrieval. Zelzate: University Press.

[10] Macken, L., Lefever, E., & Hoste, V. (2013). TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment. *Terminology, 19*(1), 1-30.

[11] Lefever, E., Macken, L., & Hoste V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.*