Medische concepten, natuurlijke taalverwerking en formele representaties van codeersystemen: een inleiding.

W. Ceusters (1, 2, 3), G. Deville (2, 4), J-L Mommaerts (2)

(1) Cel Informatica, HMKA, Brussel
(2) Office Line Engineering NV, Sint Lievens Houtem
(3) Dienst Medische Informatica, Universiteit Gent
(4) Ecole de Langues Vivantes, FUNDP, Namur

Samenvatting

Binnen de hedendaagse geneeskunde is het aanwenden van codeersystemen voor het documenteren en registreren van medische gegevens een noodzaak. Vooraleer medische concepten, geuit door artsen en paramedici in natuurlijke taal, in hun volle rijkdom weergegeven zullen kunnen worden in een voor de computer begrijpbaar formaat, zullen medici, informatici, taal- en kennistechnologen in multidisciplinair teamverband moeten samenwerken. In dit artikel wordt een eerste inleiding gegeven omtrent deze problematiek. Meer bepaald wordt dieper ingegaan op de mogelijkheden van natuurlijke taalverwerking binnen de geneeskunde, en op de formele representatiedilemma's van complexe medische uitdrukkingen en multi-axiale codeersystemen.

Résumé

La médecine d'aujourd'hui exige l'utilisation des systèmes d'encodage pour la documentation et l'enregistrement des données médicales. Mais avant que des concepts médicaux, exprimés en langage naturel, puissent être représentés dans toute leur richesse sous un format lisible pour l'ordinateur, des cliniciens, des informaticiens, et des spécialistes en ingénierie du langage et de la connaissance devront collaborer étroitement selon un approche multidisciplinaire. Cet article présente une première introduction en cette matière. Plus précisément, il explore les possibilités de traitement de langage naturel en médecine et les problèmes qui y sont associés, tels que la représentation des expressions complexes, et les systèmes d'encodage multi-axiaux.

Summary

In today's medicine, it is mandatory to use coding systems for the documentation and registration of clinical data. But before one can think of representing medical data expressed in natural language, exploiting their full richness, in a format suitable for automatic processing, clinicians, computer scientists, and engineers in language technology and knowledge representation need to collaborate in a multidisciplinary way. In this paper, an introduction is given covering some of the most important issues. More specifically, the possibilities of natural language processing in medicine are introduced, as well as the problems related to the representation of complex medical expressions and of multi-axial coding systems.

Inleiding

De belangrijkste taak van hulpverleners in de gezondheidszorg in het algemeen, en artsen in het bijzonder, is het efficiënt en optimaal lenigen van de gezondheidsbehoeften van de patienten voor wie ze verantwoordelijk zijn. Het beschikken over correcte informatie is een conditio sine qua non voor de adequate uitvoering van die taak (1). Twee klassieke informatiebronnen staan de arts daarvoor ter beschikking: het patientendossier en de medische literatuur. Terwijl het eerste alle belangrijke gegevens groepeert met betrekking tot de (para-)medische voorgeschiedenis van de patient, stelt het tweede de arts in staat om continu op de hoogte te blijven van de nieuwe inzichten op het vlak van de nosologie en de therapie.
De meest gebruikte informatiedrager is voor beide bronnen nog steeds het papier. Ondanks het steeds toenemende belang van informatica en informatietechnologie binnen verscheidene andere sectoren van ons maatschappelijk bestel, blijft de aanwending ervan binnen de gezondheidszorg in het algemeen, en de geneeskunde in het bijzonder, een problematisch punt. Het gevolg hiervan is dat niet alleen de individuele hulpverlener verstoken blijft van de informatie die hij nodig heeft, maar dat ook zowel op het vlak van de regio's en de gemeenschappen, als op nationaal en internationaal vlak, de informatievoorziening ontoereikend is. In een tijd waarin precies deze gezondheidssector met belangrijke budgettaire problemen kampt, zouden, ondermeer met het oog op efficiëntieverhoging (kosten-baten analysen) op verscheidene niveau's (artsen, hospitaalmanagement, ziekteverzekering, ...), adequate gegevens een belangrijke bron van management informatie kunnen uitmaken. Verscheidene oorzaken kunnen voor het relatief beperkt gebruik van moderne informaticatoepassingen in de medische sector verantwoordelijk gesteld worden. Eén van die oorzaken is de beperkte, zelfs vrijwel onbestaande, aanwending van taaltechnologie en computationele linguistiek binnen de medische informatica (2, 3). Tabel 1 geeft een beknopt overzicht van de behoeften terzake, en de daarop van toepassing zijnde computer-linguistische onderzoeksdomeinen.

Behoeften	Computer-linguistische onderzoeksdomeinen
• Automatische vertaling	• corpus-analyse • subtaal modelering • syntactisch-semantisch parsen • formele terminologie • opstellen van lexicons
• Normalisering en standaardisatie	• semantische modelering • formalisatie van terminologieën • opbouw van gestandaardiseerde nomenclaturen
• Automatisch indexeren van documenten	• Optical Character Recognition • tekst extractie • hypertekst
• Man/machine communicatie	• Dialoog modelering • Query - analyse
• Automatische generatie van lexicons	• syntactisch-semantische tagging • lemmatizering
• Telecommunicatie van medische gegevens	• formele taalontwikkeling met "natuurlijke" bouwstenen • gestandaardiseerde nomenclaturen
• Beslissingsondersteuning en opleiding	• hypertekst en hypertekst-achtige paradigma's • co-occurence onderzoek

Tabel 1: behoeften in de medische informatica, en de daarop van toepassing zijnde computer-linguistische onderzoeksdomeinen

Computationele linguistiek in het algemeen, en taaltechnolgie in het bijzonder, zijn wetenschappelijke disciplines die er ondermeer op gericht zijn de communicatie tussen mens en machine op inhoudelijk en interactioneel vlak zo optimaal mogelijk te laten verlopen. Dit betekent dat de machine tijdens die communicatie "perfect" begrijpt wat de mens hem meedeelt of vraagt, zonder beroep te doen op machine-georiënteerde talen (hetzij besturingscommando's of programmeertalen) maar op talen die zo dicht mogelijk aansluiten op gewone, natuurlijke taal.

Knelpunten

Twee elementen, nauw met elkaar verbonden, staan momenteel centraal binnen het medische informatica gebeuren: de ontsluiting van de gegevens aanwezig in de papieren medische dossiers enerzijds, en de communicatie van medische gegevens tussen zorgenverstrekkers en andere partijen in de gezondheidszorg anderzijds.
Electronische medische dossiers zouden voor deze dubbele problematiek een oplossing kunnen zijn, ware het niet dat de behoeften van de diverse partijen, waaronder zowel informatieleveranciers als -gebruikers, zo divers en complex zijn, dat een practische oplossing waarin iedereen zich kan vinden, momenteel nog nooit gerealiseerd werd. Voor het optimaal uitbaten van (medische) informatie moet deze immers in gestructureerde en gecodeerde vorm beschikbaar zijn, terwijl aan de bron (de arts of hulpverlener die de gegevens registreert) bijna uitsluitend van natuurlijke taal gebruik gemaakt wordt. Ondanks de intrinsieke beperkingen blijft het klassieke papieren medisch dossier voor de individuele arts een bruikbaar instrument (4). Dit neemt niet weg dat aan electronische dossiers toch de voorkeur dient gegeven te worden: "*... by their unique potential to improve the care of both*

the individual patients and populations, and concurrently, to reduce waste through continuous quality improvement..." (5). Wetenschappelijk onderzoek heeft anderzijds meermaals uitgewezen dat electronische medische dossierbeheersystemen met directe registratie van gecodeerde informatie, minder goed door artsen geaccepteerd worden dan systemen waarin door middel van vrije tekst informatie opgenomen kan worden. Nochtans bieden deze laatste systemen nauwelijks mogelijkheden voor het genereren van beleidsinformatie, het uitvoeren van klinische studies, het opzetten van audits of kwaliteitscontrole, of het actief ondersteunen van de medische praktijkvoering (dosiscontrole, geneesmiddelenbewaking, allergie beheer, ...). Een zekere bewustwording bij de artsen inzake de voordelen van het gecodeerd registreren is een eerste noodzakelijke stap, maar zeker geen eindpunt. Want zelfs indien hulpverleners er toe gebracht zouden kunnen worden om de rijkdom die de natuurlijke taal hen biedt, aan te wenden om de ziektegeschiedenis van patienten en hun acties in verband daarmee weer te geven, dan nog dient een oplossing gevonden te worden voor de bestaande papieren dossiers. Ook de informatie die zich daarin bevindt, zou immers ontsloten moeten kunnen worden. Tenslotte is met het steeds krachtiger worden van de beschikbare hardware de praktische aanwending van expertsystemen en beslissingsondersteunende toepassingen geen verre toekomstdroom meer. Althans niet wanneer de domeinkennis die daarvoor nodig is, beschikbaar zou zijn onder geformaliseerde vorm. Helaas is momenteel alle medische kennis die besloten ligt onder tekstuele vorm niet direct aanwendbaar voor het ondersteunen van dergelijke geautomatiseerde toepassingen. Het breedschalig gebruik van medische codeersystemen kan aan deze situatie verhelpen. Maar een voorwaarde is dan wel dat hulpmiddelen aangeboden worden waarbij het gebruik van dergelijke systemen voor de arts zo weinig mogelijk belastend zijn. Dit zal in vele gevallen betekenen dat de computer natuurlijke-taalverwerking zal moeten aanwenden als interface tussen de tekstuele input van de arts enerzijds en het codeersysteem anderzijds.
In dit artikel wordt een algemene inleiding gegeven op slechts één aspect van deze materie: de interpretatie van medische termen door computers en de met deze problematiek in verband staande codeersystemen.

Codeersystemen in de geneeskunde

Met het oog op het beter gebruiken van de beschikbare medische informatie werden systemen ontwikkeld om de veelheid aan medische concepten te rubriceren en te classificeren. De geschiedenis van een systeem als bijvoorbeeld ICD-9-CM, ontwikkeld voor de verwerking van morbiditeits- en mortaliteitsgegevens, gaat terug tot 1893 (6). In 1996 zal vermoedelijk de tiende herziene uitgave ervan in gebruik genomen worden voor de verplichte medische registratie in onze Belgische ziekenhuizen.
ICD-9-CM behoort tot de zogenaamde mono-axiale, hiërarchische classificeersystemen. Het systeem is mono-axiaal omdat er slechts op basis van één type concept geclassificeerd wordt, namelijk het concept ziekte of aandoening. Het is een hiërarchisch systeem omdat de verschillende concepten via IS_KIND_OFF relaties met elkaar in verband staan. Bovendien weerspiegelt deze hiërarchische classificatie zich in de codes die aan de concepten gekoppeld werden (tabel 2).

Code	Rubriek
810.	Fractuur van de clavicula.
810.0	Fractuur van de clavicula, gesloten.
810.00	Fractuur van niet gespecificeerd deel van clavicula nno, gesloten.
810.01	Fractuur van sternaal eind van de clavicula, gesloten.
810.02	Fractuur van schacht van de clavicula, gesloten.
810.03	Fractuur van acromiaal eind van de clavicula, gesloten
810.1	Fractuur van de clavicula, open.
810.10	Fractuur van niet gespecificeerd deel van de clavicula nno, open.
810.11	Fractuur van sternaal eind van de clavicula, open.
810.12	Fractuur van schacht van de clavicula, open.
810.13	Fractuur van acromiaal eind van de clavicula, open.

Tabel 2: Structuratie van clavicula-fracturen in ICD-9-CM

De verplichting om ICD-9-CM te gebruiken voor de registratie van aandoeningen naar de overheid toe, heeft ertoe geleid dat heel wat artsen zich rechtstreeks van de ICD-9-CM bedienen voor het registreren van diagnosen in het medisch dossier. Dit is uiteraard een foutief gebruik van het systeem. ICD-9-CM is in de eerste plaats immers een classificatie-systeem, een verzameling van klassen waartoe verschillende, maar uiteraard aanverwante concepten kunnen behoren. In principe zijn de klassen aanwezig in ICD-9-CM geen medische diagnosen die thuis horen in een dossier, maar wel globalisaties van zeer specifieke diagnosen, die in functie van het doel van ICD-9-CM (statistiek), niet preciezer omschreven dienen te worden (tabel 3). In het belang van de patient (therapie, prognose) is een nauwkeuriger beschrijving natuurlijk wel nodig. Trouwens, ICD-9-CM zelf is een variant van ICD-9 die precies op het klinisch gebruik afgestemd werd. Jammer genoeg ruim onvoldoende uitgewerkt.

Precies met het oog op het beter kunnen omschrijven van klinische observaties, diagnosen en behandelingen, werden multi-axiale systemen ontwikkeld. Deze systemen bevatten

Code	Rubriek
320.0	Haemophilus meningitis
320.1	Pneumokokken meningitis.
320.2	Streptokokken meningitis.
320.3	Stafylokokken meningitis.
320.7	Meningitis bij andere elders geclassificeerde bacteriele ziekten.
320.8	Meningitis veroorzaakt door overige gespecificeerde bacterien.
320.9	Meningitis veroorzaakt door niet gespecificeerde bacterien.

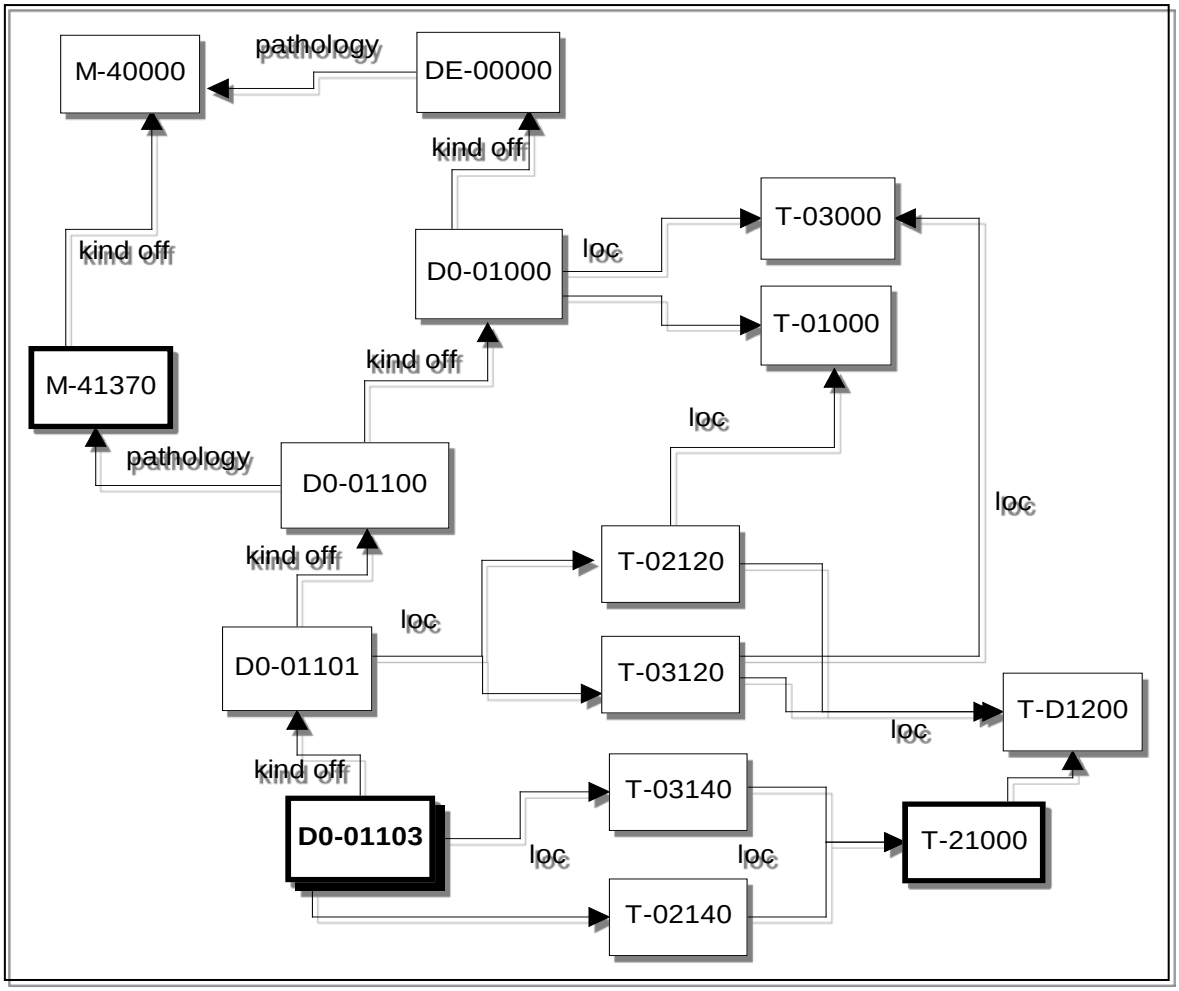
Tabel 3: Volledige en onvolledige diagnostische omschrijvingen in ICD-9-CM

medische concepten die gerubriceerd zijn volgens verscheidene assen. SNOMED® International (7) bijvoorbeeld bevat 11 assen*: morphologie, levende wezens, topologie, aandoeningen, proceduren, chemicaliën, sociale context, functies, physische agentia, beroepen,* en *semantische connectoren*. In totaal werden 132.641 concepten opgenomen. Door het combineren van deze concepten kunnen zeer precieze beschrijvingen gerealiseerd worden. Bovendien is het mogelijk om eenzelfde concept op verschillende manieren te beschrijven, afhankelijk van het doel dat men bij het coderen nastreeft. Zo geeft tabel 4 de

verschillende concepten weer die bij het registreren van een diagnose als *karbonkel van de neus* gebruikt kunnen worden. Figuur 1 geeft onder vorm van een semantisch netwerk de relaties tussen de verschillende concepten weer.

As	Code	Omschrijving
Morfologie	M-40000	Ontsteking
	M-41730	Karbonkel
Topologie	T-01000	Huid
	T-02120	Huid van het aangezicht
	T-02140	Huid van de neus
	T-03000	Onderhuids weefsel
	T-03120	Onderhuid van het aangezicht
	T-03140	Subcutaan weefsel van de neus
	T-21000	Neus
	T-D1200	Aangezicht
Aandoening	D0-01000	Infectie van huid en onderhuids weefsel
	D0-01100	Karbonkel van huid en onderhuids weefsel
	D0-01101	Karbonkel van het aangezicht
	D0-01103	Karbonkel van de neus
Connector	G-C006	Gelocaliseerd in, van, ...

Tabel 4: SNOMED® International codes die gebruikt kunnen worden voor het omschrijven van het begrip *karbonkel van de neus*.



Figuur 1: Semantisch netwerk van begrippen verband houdend met het concept *karbonkel van de neus*. Voor de betekenis van de conceptcodes, zie tabel 4.

Omgaan met medische codeersystemen

Het correct gebruik van medische codeersystemen is geen triviale zaak. Men moet coderen immers beschouwen als het produceren van een zo getrouw mogelijke vertaling van een klinische uitdrukking in natuurlijke medische taal, naar de "taal" van het codeersysteem. In de praktijk gebeurt deze vertaling ofwel door de arts zelf (met het risico op verlies aan informatie zoals hoger beschreven), ofwel door professionele codeerders. Ook dit laatste houdt een risico in omdat deze codeerders a posteriori trachten te interpreteren wat een arts origineel bedoeld heeft. Hoe minder gedetaileerd het codeersysteem is, hoe minder dit risico aanwezig is, maar uiteraard ook hoe minder nauwkeurig de precieze inhoud van een medische uitdrukking gerepresenteerd kan worden. Dit is bijvoorbeeld het geval voor ICD-9-CM.

Systemen als SNOMED® zijn zo complex geworden, dat een efficiënt gebruik ervan zonder de computer als hulpmiddel nagenoeg onmogelijk is. Maar hierdoor worden dan weer andere problemen geïntroduceerd. In de eerste plaats is het nu niet meer een menselijke codeerder die in staat moet zijn te begrijpen wat een arts bedoelt, maar wel een machine. In de tweede plaats moet de machine een perfect begrip hebben van de structuur en de inhoud van het gehanteerde codeersysteem. Het eerste is een probleem van natuurlijke taalverwerking of taaltechnologie, het tweede van semantisch modelleren en kennistechnologie. Bovendien

moet dan ook nog de relatie gelegd worden tussen de semantische representatie van een expressie, en de mogelijke representaties daarvan binnen het codeersysteem.

Natuurlijke taalverwerking in de geneeskunde

Tot voor kort heeft het onderzoek naar natuurlijke taalverwerking in de geneeskunde zich voornamelijk gericht op het automatisch indexeren van artikels en teksten (8, 9). Verscheidene benaderingen hebben in dat verband hun deugdelijkheid bewezen, zij het onder zeer precieze omstandigheden, en betreffende beperkte subdomeinen (10, 11, 12). De eerste beschrijving van zo'n systeem gaat terug tot Luhn die in 1953 het *vector-space model* ontwikkelde waarbij bepaalde eenvoudige taalkundige kenmerken automatisch in documenten herkend konden worden (13). Tegenwoordig ligt het accent meer op *symbolische representaties* (14). Aandachtspunt hierbij is ervoor te zorgen dat de representatie (in de machine) van de objecten en relaties in het behandelde domein overeenstemmen met de menselijke conceptualisatie van het domein (15).
Verscheidene methoden en technieken zijn voorgesteld geweest als oplossing voor specifieke problemen binnen de medische natuurlijke taalverwerking. De meest belangrijke zijn:

- *proximity processing* (16): het extraheren van semantische informatie uit zinnen zonder deze te parsen. Met deze techniek kan behoorlijk wat succes geboekt worden wanneer veel slecht-gevormde zinnen in een tekst voorkomen. Dit is doorgaans het geval in medische rapporten waarbij zinnen geconstrueerd worden zonder gebruik te maken van werkwoorden, en meer van nominalisaties*: blootleggen van de tumor na insnijden van de huid. Vrijprepareren van het gezwel, ...*

- *morfosemantische analyse* (17): het ontleden van woorden in kleinere betekenisvolle eenheden (derma-t-ose, hepa-t-ectomie, hyper-fosfat-urie, ...).

- *conceptual graph theory* (18): een formalisme gebaseerd op Eerste Orde Logica dat tegenwoordig ook veel gebruikt wordt voor kennisrepresentatie in de geneeskunde.

- *linguistic string processing* (19): hierbij ontwikkelt men voor specifieke subtalen een grammatica en bijpassend lexicon door het toepassen van transformationele decompositie en distributieve analysen op representatieve corpora van de subtaal. Hierbij richt men de aandacht in de eerste plaats naar wat artsen in de praktijk blijken te zeggen, en niet naar wat ze theoretisch allemaal *zouden* kunnen zeggen.

Later zullen we dieper ingaan op een taaltechnologische methode voor het semi-automatisch coderen van diagnosen en medische proceduren (20).

Semantisch modelleren van codeersystemen

Wil de computer een hulp zijn voor het registreren van medische gegevens via een codeersysteem, dan moet de machine een perfect begrip hebben van de structuur en de inhoud van het gehanteerde codeersysteem. Dit betekent dat een concept zoals uitgetekend in figuur 1, door middel van complexe datastructuren weergegeven moet worden. De machine moet aldus in staat zijn om af te leiden (*infereren*) dat een *karbonkel van de neus* een *aandoening* is van de *huid en onderhuid van de neus* op basis van een *inflammatoir proces*. Ook moeten de relaties met alle daarmee in verband staande begrippen weergeven kunnen

worden. Gelet op het feit dat reeds voor een relatief eenvoudig concept als *karbonkel van de neus* een complexe structuur opgebouwd moet worden, is het duidelijk dat de representatie van een volledig systeem zoals SNOMED® International een uitermate ingewikkelde zaak is. De complexiteit ligt hem niet in de eerste plaats in de omvang van het systeem, dan wel in de aard van de mogelijke relaties die gelden tussen bepaalde concepten. Zo spreken radiologen (en ook codeersystemen) doorgaans over *opaciteiten in de long*. Maar in werkelijkheid is een *opaciteit* niet iets dat voorkomt in een *long*, maar wel op een *radiografische opname* van een long. Het punt is nu om te bepalen welke maatstaf van precisie gehanteerd zal worden bij het modelleren van zo'n concept. Moet de representatie geschieden op basis van wat doorgaans gezegd wordt, of op basis van de realiteit ?
Een ander probleem is dit van het vereiste detail in de representatie. Moet een *maagulcus* gerepresenteerd worden als een *ulcus van de maag*, of als *een ulcus van de mucosa van de maag* ? In (21) hebben we geargumenteerd dat hier doel-afhankelijke criteria gehanteerd moeten worden.
Nog een aandachtspunt is de mate waarin alternatieve nominalisaties voor "hetzelfde" medisch begrip daadwerkelijk als identiek of als verschillend gerepresenteerd moeten worden. Geneesheren maken in hun taalgebruik geen onderscheid tussen een *longsluiering die is afgenomen*, of *een afname van een longsluiering* (22). Ook niet tussen *een pijnlijke rug*, en *pijn in de rug*. Maar een formeel systeem zal zonder bijzondere instructies dienaangaande deze syntactische varianten als refererend naar twee verschillende objecten beschouwen, met alle semantische gevolgen vandien. Want op de keeper beschouwd, gaat het inderdaad om twee verschillende dingen. Pijn kan bijvoorbeeld in intensiteit afnemen, een rug niet. Een longsluiering kan men zien op één radiografische opname. Een afname van longsluiering is niet zichtbaar op een radiografie. Het is een conclusie gebaseerd op waarnemingen over verschillende radiografieën. Noch kennistechnologie, noch taaltechnologie kunnen deze problemen eenduidig oplossen. Hiervoor moet beroep gedaan worden op de zogenaamde *pragmatiek*, ja zelfs *taalfilosofie*. Maar dat zou ons voor een inleidend artikel te ver leiden.

Besluit

In de hedendaagse geneeskunde is communicatie tussen verschillende partijen een essentieel gegeven. Communicatie kan enkel behoorlijk verlopen wanneer alle partijen elkaar perfect begrijpen. Anderzijds is de hoeveelheid medische informatie die uitgewisseld kan worden dermate toegenomen, dat methoden en technieken aangewend moeten worden om de informatie te beheersen. Medische codeersystemen spelen daarbij een zeer belangrijke rol. De toegenomen complexiteit ervan heeft wel een nood doen ontstaan naar de automatisatie van die systemen. Daardoor heeft er zich een nieuw probleem aangediend: de communicatie tussen mens en machine. Medische uitdrukkingen in natuurlijke taal moeten in een voor de computer begrijpbaar formaat omgezet kunnen worden. Deze taak overlaten aan de arts is een mogelijkheid, maar houdt het gevaar in op minder nauwkeurige weergaven van de realiteit. De machine laten instaan voor de conversie vereist een hechte samenwerking tussen artsen, informatici, linguïsten, taal- en kennistechnologen. De taak is gigantisch, maar niet onmogelijk. Talrijke stukjes van de puzzel werden reeds blootgelegd. Wat rest, is het samenleggen ervan ...