Begrijpelijkheid en Aantrekkelijkheid van een Medisch QA-systeem

Masterthesis
Faculteit Communicatie en Cultuur
Opleiding Communicatie- en Informatiewetenschappen
Specialisatie Bedrijfscommunicatie en Digitale Media/Communicatie Design

Begeleider: *Prof. dr. A.P.J van den Bosch*

Maurice Onderwater

September 2009

Voorwoord

Voor u ligt een scriptie die evaluatiemethoden aanreikt voor Question-Answering systemen. Tevens vindt u in deze scriptie een onderzoek naar manieren om antwoorden gegenereerd door een Restricted Domain Question Answering systeem, namelijk Rolaquad, te beoordelen.

Hierbij wil ik iedereen bedanken die de laatste jaren en vooral de laatste paar maanden in mij geloofd heeft. Het heeft niet altijd mee gezeten. Hoe verder ik de scriptie uitstelde, hoe lastiger het werd om de draad weer op te pakken. Uiteindelijk is het me dan toch gelukt en heb ik de moed verzameld en een einde gebreid aan een eindeloos lijkend verhaal dat nog veel langer had kunnen worden.
Ik wil Kirsten bedanken, die mij altijd weer opbeurde met de voor mij al legendarische uitspraak "goed bezig, studienerdje", Rice Cools (medisch expert) met het beoordelen van de antwoorden en Bob Bishoen voor het aandachtig doornemen en verbeteren van mijn scriptie. Ook wil ik mijn vader bedanken die mij altijd (ook financieel) is blijven steunen, ondanks de "dreigementen" dat hij het volgende jaar echt niets meer zou betalen. Ten slotte wil ik mijn begeleiders (Antal van den Bosch en Sander Canisius) bedanken voor het geduld dat ze met mij hebben gehad.

Nu op naar het leven van de werkende man. Ik heb er zin in!

Maurice Onderwater

Augustus 2009

Samenvatting

Rolaquad is een Question Answering systeem (QA-systeem) dat nog niet optimaal werkt. De betrouwbaarheid van de gegeven antwoorden is zeer laag en daarom dient gezocht te worden naar methoden die bijdragen aan de verbetering van de techniek en de presentatie van antwoorden.

In deze scriptie worden twee vragen beantwoord:

1 *Welke methode(n) is/zijn geschikt om Rolaquad zowel technisch als inhoudelijk te evalueren?*
2 *Waar zijn in de door Rolaquad gegenereerde antwoorden inhoudelijk verbeteringen te behalen?*

Het scoren van een systeem aan de hand van de *Mean Reciprocal Rank* is een onderdeel van een QA-systeemevaluatie. Het bepaalt de 'accuratesse' en wordt gemeten nadat technische verbeteringen aan de QA-algoritmes zijn aangebracht. Rolaquad scoort een *Mean Reciprocal Rank* van 0,09.

Er zijn twee typen QA-systemen, namelijk het Open Domain en het Restricted Domain QA-systeem. Open Domain QA-systemen beantwoorden zo goed als alle vragen. Restricted Domain QA-systemen alleen over een specifiek onderwerp.

Open Domain systemen worden veelal geëvalueerd binnen laboratoriumexperimenten, vaak internationaal georganiseerd in de vorm van competities. De meest belangrijke zijn: TREC (voor Engelstalige QA-systemen), NTCIR (voor Japanstalige QA's) en CLEF (voor multilinguale of "Europeestalige" QA's). Open Domain evaluatiemethoden zijn niet geschikt om een Restricted Domain systeem te beoordelen. Onderzoek wijst uit dat er onder meer problemen ontstaan met de vragentestset en de documentcollectie (Diekema *et al.*, 2004).

Voor Restricted Domain systemen zijn weinig evaluatiemethoden beschikbaar. Mede daarom hebben Diekema *et al.* (2004) een eigen methode opgesteld die verschillende aspecten van het systeem, zoals 'systeemprestaties', kan beoordelen. Deze methode mist echter een inhoudelijke evaluatie van de antwoorden op andere aspecten, zoals 'interessantheid'.
Maes *et al.* (1996) hebben twee schalen opgesteld die instructieve teksten kunnen beoordelen op 'aantrekkelijkheid' en 'begrijpelijkheid'. Deze schalen zijn mogelijk ook geschikt om de door Rolaquad gegenereerde antwoorden inhoudelijk te beoordelen.

Uit onderzoek (Theune *et al.*, 2006) is gebleken dat een lang antwoord beter is dan een kort antwoord. Dit onderzoek kijkt naar een ander aspect van een antwoord: 'directheid'. Een direct antwoord is een antwoord dat we verwachten van een specialist. Een indirect antwoord is een antwoord dat niet aan het begin, maar ergens in het midden of aan het eind van een stuk tekst te vinden is.

Aan de hand van een experimenteel onderzoek, waaraan 61 proefpersonen meewerkten, is getracht te achterhalen waar inhoudelijk verbeteringen te behalen zijn in de antwoorden van Rolaquad. De resultaten laten zien dat (1) de gebruikers van Rolaquad (leken) het verschil zien tussen een juist en onjuist antwoord, en (2) de waardering voor 'aantrekkelijkheid' en 'begrijpelijkheid' van een antwoord hangt ook van de 'correctheid' (juist/onjuist).
De gemiddelde 'aantrekkelijkheid' en 'begrijpelijkheid' kan verbeterd worden door middel van een hogere 'accuratesse' of door middel van een techniek die zoveel mogelijk antwoorden direct aan de gebruiker presenteert.

Vervolgonderzoek zou gericht kunnen worden op de vraag hoe het komt dat de gebruikers (leken) van Rolaquad zien of een antwoord al dan niet juist is.

Inhoudsopgave

1 Inleiding

Deze scriptie beschrijft het onderzoek naar diverse evaluatiemethoden en de toepassing van (enkele van) deze methoden op een Restricted Domain Question Answering systeem in ontwikkeling, namelijk Rolaquad.

Rolaquad vormt een onderdeel van het grotere NWO/IMIX[1] project waarbinnen door een aantal universiteiten[2] samengewerkt werd aan een interactief, multimodaal QA-systeem (Question-Answering-systeem, of Vraag-Antwoordsysteem). Dit systeem is in staat via een gesproken dialoog vragen op medisch gebied te beantwoorden. Rolaquad analyseert en beantwoordt de door een spraakherkenner omgezette tekst.

Rolaquad dient nog veel technische en semantische verbeteringen te ondergaan. Daarom wordt dit onderzoek uitgevoerd om de ontwikkelaars van Rolaquad een beter inzicht te verschaffen in de knel- en verbeterpunten van hun systeem. Tevens wordt getracht evaluatiemethoden aan te reiken die ook in latere stadia van ontwikkeling toegepast kunnen worden. Het onderzoek is drieledig: allereerst wordt de *Mean Reciprocal Rank* (MRR) berekend om de 'accuratesse' van Rolaquad te bepalen. Dit wordt gedaan met behulp van 185 medische vragen. De resultaten zullen na latere aanpassingen worden gebruikt om te kunnen zien of er vooruitgang is geboekt. Daarbij wordt gebruik gemaakt van de semantische *tagging* van Rolaquad, dat het systeem gebruikt om antwoorden te koppelen aan vragen, en de juistheid van de gevonden antwoorden. Vervolgens wordt onderzocht welke methoden geschikt zijn om Rolaquad zowel technisch als inhoudelijk te evalueren. Ten slotte zal onderzocht worden of en waar inhoudelijke verbeteringen te behalen zijn in de door het systeem gepresenteerde antwoorden. Hierbij wordt nagegaan of de antwoorden inhoudelijk voldoen aan de wensen van de gebruikers. Dit laatste onderzoek zal waar mogelijk ondersteund worden met eerder onderzoek over praktische en theoretische benaderingen over 'het beantwoorden van in natuurlijke taal gestelde vragen' (Theune *et al.*, 2006).

Met behulp van de testresultaten van de eerste twee evaluaties en een respondentenonderzoek, waarmee de wensen van de gebruiker worden onderzocht, wordt een verbeteringsvoorstel ontwikkeld en gepresenteerd aan de ontwikkelaars van Rolaquad.

In de komende hoofdstukken worden de verschillende evaluatiemethoden uiteengezet en vergeleken. Hoofdstuk 1 verschaft inzicht in (1) QA-systemen, (2) de werking en (3) de 'accuratesse' van Rolaquad en reikt (4) diverse evaluatiemethoden aan. Vervolgens wordt één of meerdere evaluatiemethoden uitgekozen die geschikt zijn om Rolaquad te evalueren. Deze methode(n) zal/zullen een antwoord proberen te geven op de in dit hoofdstuk gevormde hypothesen. In hoofdstuk 2 wordt de methode van het onderzoek aangereikt. In dit hoofdstuk wordt onder andere uitgelegd hoe het onderzoek in zijn werk is gegaan. Hoofdstuk 3 presenteert de resultaten van het onderzoek en geeft antwoord op de hypothesen. Er zullen nog geen conclusies verbonden worden aan de resultaten. Dat gebeurt in hoofdstuk 4, waar zowel een kwantitatieve analyse als een kwalitatieve analyse van de resultaten wordt gepresenteerd. Hoofdstuk 5 geeft de zwakten van het onderzoek weer en geeft ideeën voor vervolgonderzoek.

[1] Netherlands Organisation for Scientific Research (NWO)/Interactive Multimodal Information Extraction (IMIX)

[2] Radboud Universiteit Nijmegen, Universiteit van Tilburg, Rijksuniversiteit Groningen, Universiteit Twente en Universiteit van Amsterdam

1.1 Probleemformulering

1.1.1 Aanleiding van het onderzoek

De betrouwbaarheid van de door Rolaquad gegenereerde antwoorden is zeer laag en er dient daarom gezocht te worden naar methoden die kunnen bijdragen aan de verbetering van de achterliggende techniek en de presentatie van antwoorden; de ontwikkelaars hebben behoefte aan een duidelijk inzicht in diverse evaluatiemethoden toegespitst op een Restricted Domain systeem.

1.1.2 Vraagstelling en onderzoeksvragen

De onderzoeken beantwoorden twee vragen. De eerste luidt als volgt:

1 Welke methode(n) is/zijn geschikt om Rolaquad zowel technisch als inhoudelijk te evalueren?

Rolaquad zal in dit onderzoek onderworpen worden aan een inhoudelijke evaluatie. De antwoorden die gegenereerd worden door systeem zullen inhoudelijk beoordeeld worden met behulp van de uit het eerste onderzoek gekozen evaluatiemethode. Met behulp van de resultaten zal een antwoord gegeven worden op de volgende vraag:

2 Waar zijn in de door Rolaquad gegenereerde antwoorden inhoudelijk verbeteringen te behalen?

1.1.3 Wetenschappelijke relevantie van de studie

Op basis van deze studie zal meer duidelijkheid komen over welke evaluatiemethode(n) het meest geschikt is/zijn voor evaluatie van Restricted Domain QA-systemen. Het geeft inzicht op hoe deze methode(n) toegepast kan/kunnen worden op Rolaquad. Uit de resultaten die naar voren komen zal ten eerste moeten blijken in welke mate de evaluatiemethoden toepasbaar zijn. Dit is relevant voor toekomstige QA-evaluaties. Ten tweede komen zwakheden van Rolaquad naar voren. Dat geeft inzicht in kritieke punten van soortgelijke QA-systemen. Dit inzicht biedt steun bij de ontwikkeling van nieuwe Restricted Domain systemen.

1.1.4 Maatschappelijke relevantie van de studie

Met dit onderzoek is sprake van een indirecte maatschappelijke relevantie. Er wordt bijgedragen aan de verbetering van één systeem. Dit systeem geeft inzicht in hoe een medisch QA-systeem functioneert. Deze studie draagt bij aan de ontwikkeling van een goed functionerend systeem dat als doel heeft de gebruiker te helpen een antwoord te krijgen op medische vragen. Mensen hebben baat bij zorg. Iedereen is zelfs afhankelijk van goede medische zorg. Iedere ontwikkeling, zoals het toegankelijker maken van medische informatie door middel van een QA-systeem, draagt bij aan een verbetering van deze zorg.

1.1.5 Doelstelling van onderzoek

Het onderzoek betreft een tweeledig doel. Aan de hand van inhoudelijk evaluaties zal een verbeteringsvoorstel worden gepresenteerd, teneinde de waardering voor de antwoorden te verhogen. Daarnaast moet de evaluatie zo objectief mogelijk zijn. De evaluatie dient daarom van de ontwikkeling gescheiden te worden gehouden. Dit heeft als bijkomend voordeel dat de ontwikkelaars van Rolaquad meer tijd over hebben voor de ontwikkeling en verbetering van het systeem.

1.2 Question Answering Systemen en evaluatiemethoden

1.2.1 QA-systemen

We beginnen met een interessante medische vraag: "Wat dóet het menselijk brein?" Hierop kan het volgende antwoord worden gegeven: "De hersenen hebben een waarnemende, aansturende, controlerende en informatieverwerkende functie." De mens ziet, beweegt, controleert en leert met de hersenen. Al deze functies kunnen tot in perfectie worden uitgevoerd. De één heeft een betere motoriek, de ander heeft betere cognitieve vaardigheden. De mens wordt echter gekenmerkt door beperkingen. We kunnen namelijk niet eindeloos ver springen, oneindig lang hardlopen en oneindig zwaar tillen, noch zijn we niet in staat alles te leren en te weten.
De huidige techniek helpt ons met onze fysieke beperkingen en biedt hulp bij het onthouden en toegankelijk maken van informatie. Met de groeiende hoeveelheid informatie, neemt de vraag naar nieuwe technologiën, die deze informatie toegankelijk maken, toe. Het is niet meer de informatie die we moeten onthouden, maar de kennis hoe we bij deze informatie komen die we ons eigen dienen te maken. "Hoe heet die ene acteur uit die ene film?" Heel even "googlen". "De formule A heeft een raakpunt met formule B op de X-as, maar waar?" De grafische rekenmachine weet hier raad mee. "Ik heb een puist aan de binnenkant van mijn wang, wat kan dit zijn?" Stel uw vraag aan een medisch QA-systeem.

Een QA-systeem is een geautomatiseerd systeem (of een elektronisch programma/systeem) met als taak het beantwoorden van willekeurige in natuurlijk taal gestelde vragen. QA-systemen zijn vooral handig in situaties waarin de gebruiker een specifiek stukje informatie wil achterhalen, maar niet de tijd heeft om alle beschikbare documentatie gerelateerd aan het onderwerp te doorzoeken.

Een QA-systeem kan zowel een open als een gesloten domein hanteren. Wanneer een systeem gebruik maakt van een open domein, behandelt het zo goed als alle vragen. Een dergelijk systeem wordt een Open Domain Question Answering systeem (Open Domain systeem) genoemd. Vragen kunnen variëren van "Wat is de hoogste berg ter wereld?" tot "Wie won de slag bij Waterloo?".[3] Wanneer gebruik wordt gemaakt van een gesloten domein, betekent het dat het systeem alleen vragen behandelt over een specifiek onderwerp, zoals sport of politiek. Een systeem dat gebruikt maakt van een gesloten domein wordt een Restricted Domain Question Answering systeem genoemd.

In de ontwikkeling van QA-systemen zijn twee verschillende benaderingen mogelijk, namelijk *Knowledge-based Question Answering* en *Information Retrieval* (IR). De kennisgebaseerde benadering geeft antwoord op vragen, waarbij de informatie is opgeslagen in een database. De laatste jaren is meer focus onstaan voor een modernere techniek; het IR perspectief. Dit perspectief behandelt vragen door teksten met daarin het antwoord op te halen uit een grote hoeveelheid documenten, op grond van een oppervlakkige gelijkenis met de vraag.
Een QA-systeem bevat een 'vraag classificeermodule' die de 'soort' vraag en het 'soort' antwoord vaststelt. Nadat de vraag geanalyseerd is, worden verschillende modules gebruikt die steeds complexere NLP-technieken (*Natural Language Processing*, natuurlijke taalverwerking) toepassen op een steeds kleiner wordend stuk tekst. De *document retrieval* module tracht documenten te identificeren die het mogelijke antwoord bevatten. Hierna (pre-)selecteert een filter kleine tekstfragmenten uit de documenten. Als laatste wordt door de *answer extraction* module gezocht naar verdere aanwijzingen in de tekst om vast te stellen of het document daadwerkelijk het juiste antwoord bevat.

[3] Een voorbeeld van een Open Domain systeem is "The START Natural Language Question Answering System (http://start.csail.mit.edu/). Het is het eerste online QA-systeem en bestaat sinds December 1993.

Een QA-systeem presenteert de antwoorden op een vergelijkbare wijze als Google. De antwoorden worden gerangschikt aan de hand van relevantie. Deze relevantie wordt bepaald door eerder genoemde technieken. Het antwoord met de hoogste relevantie staat bovenaan, het antwoord met de laagste relevantie onderaan. Sommige systemen geven tien potentiële antwoorden, andere systemen proberen één enkel correct antwoord te geven.

Rolaquad maakt gebruik van de IR benadering waarbij de informatie wordt opgehaald uit een online medische encyclopedie.

1.2.2 Rolaquad[4]

Rolaquad (Robust Language Understanding in Question-Answer Dialogues) is een Restricted Domain systeem op basis van *free text question answering* volgens het IR-principe. Het systeem beantwoordt medische vragen. Antwoorden worden opgezocht in de Spectrum Medical Encyclopedia. Rolaquad analyseert de vragen en de teksten uit deze bron door eerst in de vragen en teksten, aan de hand van een semantische *tagset* die specifiek gericht is op het medische domein, een bepaalde structuur in aan te brengen. De aangebrachte structuur stelt Rolaquad in staat om bepaalde passages uit een tekst - die mogelijkerwijs het antwoord bevat - te linken aan de gestelde vraag.

De semantische tagset bevat vier verschillende annotatieniveaus:

Concepten: Het laagste niveau. Woorden worden gemarkeerd als een bepaald 'concept', zoals een aandoening of een medicijn.

Relaties: Op relatieniveau worden constructies gemarkeerd die concepten op een logische manier met elkaar verbinden.

Secties: Geannoteerde medische teksten zijn onderverdeeld in secties. Een voorbeeld van een veel voorkomende sectie in medische documenten is de sectie "symptomen".

Vraagtypen: Aan elke vraag wordt een vraagtype toegekend. Het vraagtype bepaalt wat voor antwoord er gegeven moet worden. Bijvoorbeeld: de vraag "Hoeveel mensen in Nederland zijn besmet met het HIV-virus" verwacht een hoeveelheid als antwoord.

De eerste twee niveaus worden zowel op de vraag als op de teksten toegepast. 'Vraagtypen' kunnen alleen aan de vragen worden toegekend en secties alleen aan de teksten.

De annotatie wordt gedaan met behulp van een programma, Annotator genaamd. Dit stukje software brengt semi-automatisch (waarbij menselijke annotatoren automatisch gegenereerde suggesties verifiëren) de semantische tags aan op de medische teksten uit Spectrum Medical Encyclopedia.

1.2.3 Vooronderzoek: accuratesse

De ontwikkelaars willen op twee vlakken Rolaquad verbeteren. Allereerst wordt er gestreefd naar een zo hoog mogelijke 'accuratesse' (of exactheid, nauwkeurigheid). Hoe hoog deze 'accuratesse' is, bepalen de ontwikkelaars. Daarnaast willen de ontwikkelaars een zo goed mogelijke inhoudelijke presentatie van de antwoorden (dit bepaalt de waardering voor het antwoord, zie onderzoek 2). Om Rolaquad te testen op 'accuratesse' wordt gebruik gemaakt van de volgende formule:

[4] IMIX Rolaquad Annotatiehandleiding

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{\mathrm{rank}_i}.$$

Deze formule bepaalt de 'accuratesse' van het systeem voor een bepaalde set met vragen. De *reciprocal rank* (RR) van een vraag is de inverse van de plaats waar het eerste correcte antwoord staat. Staat het eerste correcte antwoord bijvoorbeeld op de vierde plaats, dan is de *reciprocal rank* voor die vraag 1/4, ofwel 0,25. Als de vraag meerdere juiste antwoorden oplevert, dan wordt er gekozen voor de rang van het eerste correcte antwoord. Als de vraag geen correcte antwoorden oplevert, dan is de RR nul (0). De *Mean Reciprocal Rank* (MRR) is het gemiddelde van alle RR's van een vragenset (Q). De MRR kan berekend worden voor zowel een Open Domain systeem als een Restricted Domain systeem (Gillard *et al.*, 2006).

Het opstellen van de vragentestset
De vragenlijst is opgesteld aan de hand van de semantische *tagset* van Rolaquad, die bepaalt welke concepten er worden onderscheiden en herkend door het systeem. Met behulp van de tagset is een schema (schema 1.1) opgezet met daarin alle mogelijke relaties binnen en tussen de *tags*. Voor elke relatie zijn minimaal vijf vragen opgesteld. Bijvoorbeeld:

Relatie [definitie]-[aandoening]	Vraag 1:	*Wat is epilepsie?*
Relatie [bijwerking]-[behandeling]	Vraag 71:	*Wat zijn de bijwerkingen van hormoontherapie?*

Er is een lijst van in totaal 185 vragen samengesteld (zie bijlage 2). De vragen zijn bewust niet complex gemaakt; Rolaquad is een systeem ontwikkeld voor medische leken en de verwachting is dat deze groep gebruikers vrij gemakkelijke vragen zullen stellen als "Welk middel helpt tegen hoofdpijn?" of "Welke behandeling helpt tegen wratten?" Wel is geprobeerd de vragen zo divers mogelijk te maken door het gebruik van synoniemen in de vragen. "Wat…" of "Wanneer spreekt men…" suggereert bijvoorbeeld in de meeste gevallen een antwoord met een definitie.
Vervolgens zijn alle vragen ingevoerd in Rolaquad en heeft een medisch expert de antwoorden met een binaire schaal aangemerkt als 'juist' of 'onjuist'.

Resultaat
Rolaquad scoort een MRR van 0,09. Het is aan de ontwikkelaars om te bepalen of dit al dan niet een goede score is.

1.2.4 Open Domain Question Answering evaluatiemethoden

Over onderstaande evaluatiemethoden zijn een aantal naslagwerken beschikbaar. Het moet gezegd worden dat deze scriptie inhoudelijk niet te diep op de methoden zal ingaan. Ook is het van belang te weten dat het hier niet specifiek om methoden gaat, maar om conferenties waar de organisatoren zelf de methoden ontwikkeld hebben.

TREC
TREC (Ellen *et al.*, 2005), the Text REtrieval Conference, is het jaarlijkse evaluatie-evenement van de IR-gemeenschap. De conferentie wordt gesponsord door het Amerikaanse ministerie van Defensie en NIST (National Institute of Standards and Technology). Het evenement is opgedeeld in verschillende onderdelen (*tracks*), zoals *ad hoc retrieval*, *filtering* en 'het beantwoorden van vragen'. Het eindresultaat van de conferentie is een rapport geschreven door de organisatoren en een technisch rapport geschreven door de participanten. TREC focust zich

voornamelijk op *ad hoc retrieval*, waarbij door middel van een query een op relevantie gerangschikte lijst met documenten als resultaat wordt teruggegeven.

NTCIR
NACSIS Test Collection for Information Retrieval (NTCIR[5]) is de Japanse variant van TREC welke wordt gesponsord door the National Institute for Informatics van Tokyo.

Een van de doelen van de conferentie is het aanmoedigen van onderzoek in *Information Access* (IA) door een grote testcollectie en een standaard evaluatie-infrastructuur aan te bieden om vergelijkingen met andere systemen mogelijk te maken.

De nadruk wordt voornamelijk gelegd op (1) IR met Japanse en andere Aziatische talen en cross-linguale IR, (2) de overgang van *document retrieval* naar *information retrieval*, (3) technologieën die informatie uit documenten bruikbaar maken en (4) diverse evaluatiemethoden.

CLEF
De vraag naar een multimodale QA-evaluatiemethode is groot; TREC en NTCIR zijn respectievelijk een op de Engelse en Japanse taal gebaseerde evaluatie. Echter, maar 5% van de wereldbevolking spreekt Engels en een nog kleiner gedeelte Japans. CLEF[6] (Cross-Language Evaluation Forum) levert een infrastructuur om IR/QA-systemen opererend op Europese talen te testen en te evalueren.
Tijdens de conferenties worden twee runs georganiseerd. In de eerste run dienen de aangemelde systemen de vragen te beantwoorden via het Internet. In de tweede run worden de antwoorden opgezocht in een documentencollectie aangeleverd door CLEF. Daarna geeft de organisatie een lijst met de juiste antwoorden. Aan de hand van deze lijst kunnen de deelnemers hun systeem evalueren. Luis *et al.* (2006) geven in hun paper drie evaluatietechnieken die hiervoor gebruikt kunnen worden: *error analysis*, *component removal* en *component substitution*. Bij *error analysis* wordt aan de hand van het systeemlogbestand nagegaan waarom geen juist antwoord is gevonden. De *component removal* evaluatie houdt in dat het systeem nogmaals de vragen beantwoordt, maar in dit geval zonder het gebruik van een of meerdere systeemcomponenten. Op die manier kan worden achterhaald of een component niet eerder een knelpunt vormt bij de zoektocht naar een juist antwoord. De laatste evaluatie gaat na of een component niet beter vervangen kan worden door een ander component met dezelfde of vergelijkbare functionaliteit. Uit een evaluatie[7] van Luis *et al.* (2006) kwam naar voren dat het gebruik van meerdere bronnen de prestaties van hun systeem ten goede kwam.

1.2.5 Restricted Domain Question Answering evaluatiemethoden

[5] http://research.nii.ac.jp/ntcir/outline/prop-en.html
[6] http://www.cultivate-int.org/issue6/clef/
[7] Het Portugeestalige systeem dat hierbij geëvalueerd is, wordt Esfinge genoemd en is te vinden op http://www.linguateca.pt/Esfinge/.

Uit eerder onderzoek (Diekema *et al.*, 2004) is gebleken dat het toepassen van een evaluatiemethode die geschikt is voor een Open Domain systeem op een Restricted Domain systeem voor problemen zorgt in het opstellen van 'testvragen' en 'controleantwoorden' en in de constructie van een documentcollectie. Allereerst worden de vragen beantwoord vanuit een beperkt aantal documenten (Doan *et al.*, 2004), waardoor het vooral op de betrouwbaarheid en volledigheid van deze documenten aankomt. Daarnaast is het zeer lastig voor een systeem om te achterhalen welk deel van een tekst het juiste antwoord bevat. Daarom gebruiken Open Domain systemen zogenaamde *redundancy* technieken, waarmee overtollige informatie wordt opgespoord en weggelaten. Hierbij wordt gebruik gemaakt van het Internet als verificatiemiddel. Het antwoord wordt gekozen aan de hand van de sleutelwoorden uit de gestelde vraag. Hoe vaker het sleutelwoord (of de combinatie van sleutelwoorden) voorkomt, hoe groter de waarschijnlijkheid dat het antwoord in dat document de juiste is. Restricted Domain systemen kunnen deze *redundancy* technieken niet goed toepassen, omdat deze systemen vaak gebruik maken van slechts één, maximaal twee bronnen (Doan *et al.*, 2004). Bovendien werken deze systemen met een zeer domeinspecifieke en uitgebreide terminologie en een grote ambiguïteit. Ten slotte dient een Restricted Domain systeem in opdracht van de klant vaak zeer gedetailleerde vragen te kunnen beantwoorden, wat het samenstellen van een vragenset bemoeilijkt.

Om die reden hebben Diekema *et al.* (2004) een schema opgesteld waarmee een Restricted Domain systeem geëvalueerd kan worden. Zij stellen dat een generieke evaluatie niet voldoende is voor een Restricted Domain systeem. De methode behandelt andere evaluatiecriteria dan Open

Schema 1.2: Evaluatieschema van Diekema *et al.* (2004)

1 System Performance
- 1.1 Speed
- 1.2 Availability / reliability / upness

2 Answers
- 2.1 Completeness
- 2.2 Accuracy
- 2.3 Relevance
- 2.4 Applicability to task / utility / usefulness

3 Database Content
- 3.1 Authority / provenance / Source quality
- 3.2 Scope /extensiveness / coverage
- 3.3 Size
- 3.4 Updatedness

4 Display (UI)
- 4.1 Input
  - 4.1.1 Question understanding / info need understanding
  - 4.1.2 Querying style
    - 4.1.2.1 Question
      - 4.1.2.1.1 NL query
    - 4.1.2.2 Keywords
    - 4.1.2.3 Browsing
  - 4.1.3 Question formulation assistance
    - 4.1.3.1 Spell Checker
    - 4.1.3.2 Abbreviation recognition
- 4.2 Output
  - 4.2.1 Organization
  - 4.2.2 Feedback Solicitation

5 Expectations
- 5.1 Googleness

Domain evaluaties, zoals 'systeemprestaties' en 'bruikbaarheid' (zie schema 1.2). Aan de hand van een open onderzoek onder 25 tot 30 studenten is onderstaand schema ontwikkeld. Het schema bevat 5 hoofddimensies: *System Performance, Answers, Database Content, Display, Expectations*. De

dimensies spreken redelijk voor zich. Alleen 5.1 verdient extra uitleg. *Googleness* staat voor de automatische vergelijking die door gebruikers gemaakt wordt met de zoekmachine Google bij het evalueren en beoordelen van een vergelijkbaar systeem. Een onderzoeksvraag die gesteld kan worden met betrekking tot *Googleness* is bijvoorbeeld: "In vergelijking met Google, vinden de respondenten dit systeem snel/langzaam?"

1.2.6 De gekozen evaluatiemethode

Rolaquad zou deel kunnen nemen aan een (nog niet bestaande) competitie voor Nederlandstalige medische QA-systemen, maar een dergelijke competitie zou met mogelijk maximaal twee andere deelnemers te klein zijn. De drie jaarlijkse bijeenkomsten om een Open Domain systeem te testen en te evalueren, zijn weer niet van toepassing op Rolaquad. Rolaquad is namelijk noch een Engels óf Japans noch een multilinguaal systeem. Daarbij gaat het hier om conferenties, en de methoden zijn om die reden niet 'zelf toepasbaar[8]'. Het is de bedoeling dat systeemontwikkelaars zich (jaarlijks) inschrijven. Het systeem wordt getest aan de hand van vragentestsets. De ontwikkelaars schrijven vervolgens zelf een evaluatie en gaan met behulp van de resultaten na waar verbeteringen te behalen zijn. Waar de vragentestsets van de conferentie vaak hetzelfde blijven, veranderen de evaluatietechnieken ieder jaar. Op die manier moedigen deze conferenties de ontwikkelaars aan om naar nieuwe IR-technieken te blijven zoeken en de oude technieken te verbeteren.
Er zijn bovendien geen Open Domain evaluaties beschikbaar die ontwikkelaars zelf kunnen gebruiken om hun systeem te testen. Dat dit soort evaluaties niet bestaan, is omdat de resultaten van een dergelijke evaluatie niets zou zeggen over hoe goed een systeem nu werkelijk is. Inschrijven voor een conferentie en het "opnemen" tegen andere systemen en de resultaten vergelijken zorgt voor een veel sterkere graadmeter.
Met behulp van een eigen vragentestset en het berekenen van de 'accuratesse' bestaat er een redelijke overlap met de Open Domain evaluatiemethoden. De ontwikkelaars en gebruikers bepalen echter zelf de standaard.
Zoals blijkt uit deze paragraaf zijn zowel TREC, NTCIR en CLEF om meerdere redenen niet geschikt om Rolaquad te evalueren. In de volgende alinea's zullen we kijken naar de Restricted Domain evaluatiemethoden.

Het schema van Diekema *et al.* (2004) is uitermate geschikt om een Rolaquad technisch te evalueren. De methode biedt de mogelijkheid om het systeem te testen op onder andere *performance*, *user-interface design* en *up-to-dateness*. Wat de ontwikkelaars van Rolaquad willen weten is of de antwoorden inhoudelijk ook voldoen. Dimensie 2 in het schema (*answers*) zou een antwoord kunnen geven op deze vraag. Het behandelt aspecten als 'compleetheid' (is het antwoord dekkend), 'accuratesse', 'relevantie' en 'toepasbaarheid' (kan de gebruiker de taak uitvoeren met het gegeven antwoord of mist er nog informatie). De schrijvers bieden echter geen methode aan om de vragen inhoudelijk op aspecten als de mate van interessantheid te testen, en dat specifiek onder de gebruikersgroep van Rolaquad.

Rolaquad is een systeem ontwikkeld als raadpleegmiddel voor leken. Daarom is de verwachting dat gebruikers van Rolaquad meer waarde hechten aan andere aspecten van het antwoord dan alleen de juist- of compleetheid. Dit in tegenstelling tot systemen die als naslagwerk gebruikt worden door experts. De vraag is of de gebruikers van Rolaquad de antwoorden bijvoorbeeld niet te lang, te complex of wel interessant genoeg vinden.

[8] Met 'zelf toepasbaar' wordt bedoeld: "het zelf uitvoeren van een evaluatie, waarbij de resultaten worden vergeleken met eerdere resultaten (voor hetzelfde systeem)".

In Maes *et al.* (1996) wordt een methode voorgelegd om teksten op allerlei aspecten te waarderen. Het waarderen gebeurt aan de hand van twee schalen: 'aantrekkelijkheid' en 'begrijpelijkheid'. De schalen worden gepresenteerd voor instructieve teksten, maar zijn vermoedelijk ook geschikt voor persuasieve en informatieve teksten[9]. De schalen zijn te vinden in hoofdstuk 3.

Sinds kort wordt er steeds meer aandacht gegeven aan Restricted Domain systemen. Daarbij erkennen onderzoekers steeds meer het belang van lange en complete antwoorden (Theune *et al.*, 2006). Ook Lin *et al.* (2006) heeft experimenten uitgevoerd waarin werd aangetoond dat gebruikers het antwoord binnen de context willen hebben, of een antwoord binnen een paragraaf. Een mogelijke verklaring voor dit resultaat is dat het onderzoek alleen "ideale" antwoorden voorlegt die niet door middel van een query uit een document opgehaald zijn. Binnen een systeem als Rolaquad hebben we te maken met antwoorden die ook irrelevante informatie bevatten. Over de lengte van een antwoord gegenereerd door Rolaquad valt dus weinig te zeggen. De verwachting is dat de gebruikers van Rolaquad eerder de voorkeur hebben voor een korter antwoord dat direct antwoord geeft op de vraag, zoals in dit voorbeeld beschreven is:

V: *"Welk middel kan worden voorgeschreven tegen hoofdpijn?"*
A: *"Paracetamol of Ibuprofen kan worden voorgeschreven tegen hoofdpijn."*

Of nog korter:

A: *"Paracetamol of Ibuprofen."*

In dit onderzoek zal daarom een ander onderscheid gemaakt worden, namelijk tussen een direct antwoord en een indirect antwoord. Een direct antwoord moet gezien worden als een antwoord dat we verwachten als we de vraag aan een specialist zouden stellen. Een indirect antwoord is een antwoord waar de gebruiker voor de gewenste informatie op zoek moet in een alinea of paragraaf.

[9] http://www.let.uu.nl/~bregje.holleman/personal/handlmaster07.htm

Hypothesen
In deze scriptie is niet gekozen voor een bestaande voor QA-systemen geschikte evaluatiemethode, maar voor een onderzoeksmethode die gebruikt wordt om (instructieve) teksten inhoudelijk te evalueren, namelijk de schalen van Maes *et al.* (1996). Er zal getest worden of de schalen toepasbaar zijn op een door een QA-systeem gegenereerd antwoord.

Er is een onderzoek opgesteld dat onder andere antwoord moet geven op de vraag of de gebruikers zien of een antwoord juist is en of de juiste en onjuiste antwoorden verschillend scoren op 'aantrekkelijkheid' en 'begrijpelijkheid'. De voorspelling is dat gebruikers niet kunnen zien of een antwoord al dan niet juist is, omdat zij niet over de benodigde medische kennis beschikken (hypothese 1a). Mede om die reden zal een juist antwoord ook niet hoger scoren op 'begrijpelijkheid' en 'aantrekkelijkheid' (hypothese 1b en 1c). Mogelijk staan 'aantrekkelijkheid' en 'begrijpelijkheid' volledig los van de correctheid, ook als de respondenten wél het verschil tussen een juist en onjuist antwoord opmerken (zie inleidende tekst hypothese 2).
Het resultaat hiervan bepaalt hoeveel belang er gehecht moet worden aan bijvoorbeeld de uitkomst van de MRR. De volgende hypothesen zijn opgesteld om een antwoord te geven op deze vraag en op onderzoeksvraag 2:

Hypothese 1a:	*Bij geen van de voorgelegde vragen scoort een juist antwoord hoger dan een onjuist antwoord.*
Hypothese 1b:	*Bij geen van de voorgelegde vragen wordt een juist antwoord beter begrepen dan een onjuist antwoord.*
Hypothese 1c:	*Bij geen van de voorgelegde vragen wordt een juist antwoord aantrekkelijker bevonden dan een onjuist antwoord*

De verwachting is dat de 'score' die gebruikers aan een antwoord geven geen effect heeft op 'aantrekkelijkheid'. Een stukje tekst kan wel interessant zijn, zelfs als het geen antwoord geeft op de vraag (hypothese 2a). Ook is de verwachting dat de 'score' geen effect heeft op 'begrijpelijkheid'. Waarom zou je een stukje tekst niet begrijpen als het volgens jou niet een antwoord is op de vraag (hypothese 2b)? Tussen 'aantrekkelijkheid' en 'begrijpelijkheid' is mogelijk wel een verband; hoe beter een antwoord te begrijpen is, hoe aantrekkelijker het is voor de gebruiker om het te lezen, en visa versa (hypothese 2c).

Hypothese 2a:	*Er bestaat geen verband tussen score en aantrekkelijkheid.*
Hypothese 2b:	*Er bestaat geen verband tussen score en begrijpelijkheid.*
Hypothese 2c:	*Er bestaat een positief verband tussen de aantrekkelijkheid en begrijpelijkheid van een antwoord.*

Er zal tevens gekeken worden in hoeverre de factor 'directheid' van invloed is op de 'score', 'begrijpelijkheid' en 'aantrekkelijkheid'. Het medisch gebied is een vakgebied met een zeer diverse en uitgebreide terminologie en om deze reden én omdat Rolaquad bedoeld is voor leken, is de verwachting dat wanneer een antwoord direct is, het antwoord beter te begrijpen is. Ook op 'aantrekkelijkheid' zal een direct antwoord beter scoren. Deze verwachting komt voort uit de eerdere hypothese dat er mogelijk een positief verband is tussen 'aantrekkelijkheid' en 'begrijpelijkheid'. De derde hypothese luidt daarom als volgt:

Hypothese 3:	*Een direct antwoord scoort hoger op zowel begrijpelijkheid, aantrekkelijkheid als op score.*

2 Methode van onderzoek

2.1 Stimulusmateriaal

Om de data zoveel mogelijk onder controle te houden, werden de vragen en antwoorden in het eerste gedeelte van het onderzoek aan de respondenten gepresenteerd – respondenten konden niet zelf hun vragen bedenken. Hiermee werd voorkomen dat het onderzoeksmateriaal onbruikbaar werd voor analyse. De respondenten werden niet geïnformeerd over het feit dat de vragen en antwoorden uit Rolaquad komen, dit om *priming*[10] te voorkomen.

2.2 Respondenten

Aan dit deel van het onderzoek hebben 61 respondenten deelgenomen. Hiervan waren er 28 vrouw (45,9%) en 32 man (52,5%). In alle gevallen waren de proefpersonen student aan de Universiteit van Tilburg. Van de respondenten zocht 85% minder dan één keer per maand naar medische informatie. 6,7% zocht hier vaker dan één keer per maand naar. De gemiddelde leeftijd van de proefpersonen was 24 (SD = 3.7), van 1 proefpersoon waren de leeftijd en de sekse niet bekend.

2.3 Instrumentarium

In het onderzoek is gebruik gemaakt van vragen aan de hand van een zevenpunts Likertschaal[11]. Maes *et al.* (1996) stellen vragen voor waarmee tekstwaardering en tekstbegrip gemeten kan worden. Er wordt gebruik gemaakt van een *balanced scale*, omdat dit de betrouwbaarheid van het onderzoek verhoogt. In een dergelijke schaal wordt de helft van de vragen negatief en de andere helft positief geformuleerd. In de schalen uit het boek zijn 'begrijpelijkheid' ('tekstbegrip') en 'aantrekkelijkheid' ('tekstwaardering') van teksten opgenomen, maar ook het imago van de producent. Dit laatste werd in dit onderzoek niet opgenomen, omdat de respondenten niet werden geïnformeerd over de herkomst van de antwoorden.

2.4 Procedure

Dit deel van het onderzoek werd op papier afgenomen. Het onderzoek bestond uit in totaal 10 vragen met bij elke vraag een door Rolaquad gegenereerd antwoord. De proefpersonen dienden de vraag en het antwoord van Rolaquad goed te lezen. Vervolgens werd er gevraagd om aan de hand van een schaal een waardering te geven aan het gegenereerde antwoord. De vragen zijn gekozen uit de 185 eerder opgestelde vragen (bijlage 2). Er is gezocht naar 10 vragen (bijlage 1) met zowel een juist als een onjuist antwoord uit een verschillende 'categorie', waarvan alleen van de juiste vragen 5 antwoorden direct en 5 antwoorden indirect antwoord gaven op de vraag.

2.5 Design

Er zijn tien verschillende versies van het onderzoek opgesteld met daarin willekeurig juiste en onjuiste antwoorden. Van elke vraag zijn één juist en één onjuist antwoord opgenomen. De respondent kreeg nooit dezelfde vraag met een verschillend antwoord voorgelegd.
Omdat de respondenten gevraagd werd het huidige systeem te beoordelen, is er alleen gebruik gemaakt van vragen en antwoorden die respectievelijk aan en door het systeem gesteld en gegenereerd zijn; de vragen en antwoorden zijn niet ten behoeve van het onderzoek

[10] Priming refereert naar een hogere gevoeligheid voor bepaalde stimuli door een eerdere ervaring.
[11] Een schaal waarin aan de respondent gevraagd wordt de mate van instemming met een bepaalde uitspraak aan te geven door middel van een meerkeuze antwoordmodel.

gemanipuleerd. De antwoorden die gegeven zijn in het onderzoek waren vooraf geclassificeerd als een direct of als een indirect antwoord op de vraag – dit is alleen gedaan voor het beste antwoord en niet voor de onjuiste antwoorden.

Hieronder volgen de vragen en schalen die voorgelegd zijn in het onderzoek.

Demografische vragen

1. *Wat is uw leeftijd?*
2. *Wat is uw geslacht?*
3. *Hoe vaak hebt u het afgelopen jaar naar medische informatie gezocht?*

Schalen

Juistheid								
Ik geef dit antwoord een score[12] van:								
laag	1	2	3	4	5	6	7	hoog
begrijpelijkheid								
Ik vind het antwoord:								
moeilijk	1	2	3	4	5	6	7	makkelijk
eenvoudig	1	2	3	4	5	6	7	ingewikkeld
onduidelijk	1	2	3	4	5	6	7	duidelijk
onoverzichtelijk	1	2	3	4	5	6	7	overzichtelijk
logisch opgebouwd	1	2	3	4	5	6	7	onlogisch opgebouwd
bondig	1	2	3	4	5	6	7	omslachtig
aantrekkelijkheid								
Ik vind het antwoord:								
interessant	1	2	3	4	5	6	7	oninteressant
afstandelijk	1	2	3	4	5	6	7	aansprekend
afhoudend	1	2	3	4	5	6	7	uitnodigend
boeiend	1	2	3	4	5	6	7	saai
persoonlijk	1	2	3	4	5	6	7	onpersoonlijk
eentonig	1	2	3	4	5	6	7	afwisselend

[12] Om verwarring te voorkomen wordt in deze scriptie een onderscheid gemaakt tussen 'score' (gequoteerd), score (niet gequoteerd) en waardering. Alleen met 'score' wordt de item bedoeld. Het gaat over waardering als het over de schalen 'begrijpelijkheid' en 'aantrekkelijkheid' gaat.

3 Resultaten

In dit hoofdstuk volgt een oppervlakkige presentatie van de resultaten.

Aan de hand van een varimaxrotatie kan de lading van een item voor de schalen worden bepaald. Normaliter wordt deze analyse gebruikt om de verschillende items uit een onderzoek op te delen in schalen. In dit onderzoek was het slechts een controle.

Tabel 3.1: Ladingen van de tekstwaarderingsitems na varimaxrotatie

		'begrijpelijkheid'	'aantrekkelijkheid'
begr3	ingewikkeld - eenvoudig	,166	**,832**
begr1	moeilijk - makkelijk	,223	**,797**
begr4	onoverzichtelijk - overzichtelijk	,367	**,777**
begr3	onduidelijk - duidelijk	,442	**,687**
begr5	omslachtig - bondig	-,008	**,607**
begr6	onlogisch opgebouwd - logisch opgebouwd	,480	**,530**
aant2	afstandelijk - aansprekend	**,812**	,262
aant1	oninteressant - Interessant	**,799**	,174
aant4	saai - boeiend	**,795**	,264
aant3	afhoudend - uitnodigend	**,758**	,292
aant6	eentonig - Afwisselend	**,711**	,092
aant5	onpersoonlijk - persoonlijk	**,621**	,113

Bij een factorlading van .50 of meer is de achtergrond gearceerd. Ieder woord(paar) is voorafgegaan door *ik vind het antwoord.*

In bovenstaande tabel is duidelijk te zien dat de toegepaste schalen ook in dit onderzoek geldig waren.

De interne consistentie van de schalen 'begrijpelijkheid' en 'aantrekkelijkheid' waren goed (respectievelijk Cronbach's $\alpha = .86$ en Cronbach's $\alpha = .87$).

In tabel 3.2 staan apart de waarderingen die gegeven zijn aan juiste en onjuiste antwoorden op dezelfde vraag voor 'aantrekkelijkheid', 'begrijpelijkheid' en 'score'.

Tabel 3.2: Waardering van 'score', 'begrijpelijkheid' en 'aantrekkelijkheid' in relatie met correctheid (de waardering is minimaal 1 en maximaal 7, standaardafwijking staat tussen haakjes)

	juist	onjuist
score	4.62 (1.60)	3.03 (1.70)
begrijpelijkheid	4.31 (1.42)	3.92 (1. 24)
aantrekkelijkheid	3.87 (1.11)	3.41 (1.11)

Op de onderzoeksresultaten is een multivariate analyse losgelaten om te bepalen of er verschillen zijn in de waardering ('score', 'begrijpelijkheid' en 'aantrekkelijkheid') van antwoorden. De 'correctheid' (juist/onjuist) werd gebruikt als conditie.

Tabel 3.3: 'score' in relatie met correctheid

Vraag	A	B	C	D	E	F	G	H	I	J
onjuist	2.79	4.20	2.60	2.84	2.97	**2.35**	3.66	3.63	2.43	2.87
juist	**5.26**	**5.55**	4.94	4.30	4.40	3.40	5.29	4.17	3.94	4.93
significantie	<.001	<.001	<.001	<.005	<.05	<.025	<.001	**=.20**	<.001	<.001
verklaarde variantie	**47.0**	18.7	35.5	15.6	16.0	10.3	24.6	2.8	19.2	32.4

In bovenstaande tabel (3.3) worden voor alle vragen afzonderlijk de gemiddelden van zowel de juiste als de onjuiste antwoorden weergegeven. De 'verklaarde variantie' geeft het effect (in procenten) weer tussen de twee condities. Als de 'verklaarde variantie' significant is dan spreken we van een verschil.

Alleen bij vraag H was er géén significant verschil tussen het juiste en onjuiste antwoord bij dezelfde vraag als het gaat om de 'score'. Respondenten gaven bij negen van de tien voorgelegde vragen een hogere 'score' aan het juiste antwoord.

Tabel 3.4: 'begrijpelijkheid' in relatie met correctheid

Vraag	A	B	C	D	E	F	G	H	I	J
onjuist	**3.17**	**5.32**	4.14	4.38	3.60	3.32	3.26	3.64	4.34	3.95
juist	**5.75**	5.01	3.65	3.76	3.72	4.56	4.78	3.34	3.90	4.61
significantie	<.001	=.23	=.13	<.05	=.68	<.005	<.001	=.21	=.17	<.05
verklaarde variantie	**62.1**	-2.4	-3.8	-9.3	0.3	15.0	26.4	-2.6	-3.2	7.4

In tabel 3.4 is te zien dat voor de vragen A, F, G en J (40%) een positief significant verschil bestond in 'begrijpelijkheid' tussen het juiste en onjuiste antwoord op dezelfde vraag. Voor vraag D werd zelfs een significant negatief verschil gevonden. Dit betekent dat voor deze vraag het onjuiste antwoord door respondenten beter begrepen werd dan het juiste antwoord.

Tabel 3.5: 'aantrekkelijkheid' in relatie met correctheid

Vraag	A	B	C	D	E	F	G	H	I	J
onjuist	**2.67**	4.18	3.82	4.00	3.23	2.74	3.36	3.16	3.00	4.02
juist	3.98	4.24	3.92	4.00	3.74	3.50	**4.62**	3.17	3.60	3.92
significantie	<.001	=.80	=.74	=.94	=.07	<.01	<.001	=.95	<.025	=.72
verklaarde variantie	27.4	0.1	0.2	0.0	5.3	12.6	30.8	0.0	10.3	-0.2

In tabel 3.5 is voor de vragen A, F, G, I was een significant verschil in 'aantrekkelijkheid' te zien tussen het juiste en onjuiste antwoord op dezelfde vraag. Dat is 40% van de voorgelegde vragen.

Hiermee worden hypothese 1a, 1b en 1c verworpen. Voor 'score' is echter een opmerkelijk resultaat gevonden. Voor negen van de tien voorgelegde vragen scoorde een juist antwoord wél beter dan een onjuist antwoord.

Aan de hand van de waardes gegeven voor 'score', 'begrijpelijkheid' en 'aantrekkelijkheid' waren correlaties berekend. Er werd hierbij geen onderscheid gemaakt tussen een juist of een onjuist antwoord. De correlatie tussen 'aantrekkelijkheid' en die van 'begrijpelijkheid' bedroeg .66 ($p < .01$). Er werd tevens een significante correlatie gevonden voor 'aantrekkelijkheid' met 'score' en 'begrijpelijkheid' met 'score' (respectievelijk .66 ($p < .01$) en .49 ($p < .05$)). Met dit resultaat worden zowel hypothese 2a als 2b verworpen. Hypothese 2c wordt hiermee bevestigd.

In dit onderzoek is bewust gezocht naar vijf vragen die direct en vijf vragen die indirect antwoord gaven op een antwoord. De vragen A, B, E, G en J gaven in dit onderzoek een direct antwoord op de vraag.

Tabel 3.6: Waardering van 'score', 'begrijpelijkheid' en 'aantrekkelijkheid' voor de juiste antwoorden in relatie met directheid van het antwoord (score is minimaal 1 en maximaal 7, standaardafwijking staat tussen haakjes)

	direct	indirect
score	5.09 (1.41)	4.15 (1.65)
begrijpelijkheid	4.78 (1.37)	3.84 (1.32)
aantrekkelijkheid	4.10 (1.10)	3.64 (1.08)

Bij 'score' was er een significant verschil op te merken tussen een antwoord direct en een antwoord indirect gegeven op een vraag ($t(303)=5.36$, $p < .001$). Bij 'begrijpelijkheid' was er een significant verschil op te merken tussen een antwoord direct en indirect gegeven op een vraag ($t(304)=6.12$, $p < .001$). Bij 'aantrekkelijkheid' was er een significant verschil op te merken tussen een antwoord direct en indirect gegeven op een vraag ($t(304)=3.71$, $p < .001$). Deze resultaten bevestigen hypothese 3.

4 Analyse

4.1 Kwantitatieve analyse van de resultaten

Het onderzoek had tot doel te achterhalen welke methoden geschikt zijn om een Restricted Domain systeem te evalueren en waar inhoudelijk verbeteringen te behalen zijn in de door Rolaquad gegenereerde antwoorden. Op de eerste vraag is vanuit bestaande theorie een antwoord gegeven: er zijn geen Open Domain evaluatiemethoden geschikt om een Restricted Domain systeem als Rolaquad te evalueren. Diekema *et al.* (2006) hebben een geschikt schema opgesteld dat zich voornamelijk gecentreerd heeft op een technische evaluatie van een Restricted Domain systeem. De tweede onderzoeksvraag is op verzoek van de ontwikkelaars opgenomen in deze scriptie. Er zijn hypothesen opgesteld die indirect een antwoord zullen geven op deze vraag. Op basis van de resultaten kunnen echter geen conclusies gevormd worden die stellen dat het systeem "onvoldoende", "voldoende" of "goed" scoort.

Uit de resultaten (hypothese 1a, 1b en 1c) komt naar voren dat respondenten zien dat een antwoord juist of onjuist is; bij negen van de tien voorgelegde antwoorden scoorden de juiste antwoorden significant hoger dan de onjuiste antwoorden. Dit maakt de waarde van 'accuratesse' extra waardevol. Hoe hoog deze waarde uiteindelijk dient te zijn, wordt bepaald door de ontwikkelaars. Er kan niettemin voorzichtig gesteld worden dat de 'accuratesse' op dit moment (0,09) aan de lage kant is en dat iedere verbetering die de 'accuratesse' ten goede komt er voor zal zorgen dat de antwoorden hoger zullen scoren. Dit resultaat bepaalt uiteindelijk weer de waardering voor het systeem an sich. Op de schalen 'begrijpelijkheid' en 'aantrekkelijkheid' scoorden de juiste en onjuiste antwoorden voor ongeveer de helft van de vragen verschillend. Voor de overige vragen scoorden de onjuiste antwoorden op 'aantrekkelijkheid' zelfs een aantal keer hoger.

Er bestaat een relatie tussen de 'score' van het antwoord gegeven door de gebruiker en 'aantrekkelijkheid' en 'begrijpelijkheid'. De respondenten hechten blijkbaar waarde aan het al dan niet juist zijn van het antwoord met betrekking tot 'aantrekkelijkheid' en 'begrijpelijkheid'. Daarom zullen de gemiddelden van de schalen 'aantrekkelijkheid' en 'begrijpelijkheid' naar alle waarschijnlijkheid ook hoger scoren als de 'accuratesse' van het systeem hoger wordt.

Met een gemiddelde waardering van 4.31 (juiste antwoorden) en 3.92 (onjuiste antwoorden) voor de 'aantrekkelijkheid' en 3.87 (juiste antwoorden) en 3.41 (onjuiste antwoorden) liggen de waarderingen rond de waarde 4. Dit lijkt niet erg hoog. Het is onbekend hoe andere systemen hierop scoren. Om te zien of het systeem vooruitgang boekt, zal het onderzoek daarom later nog een keer uitgevoerd moeten worden. Bij voorkeur met dezelfde vragen.

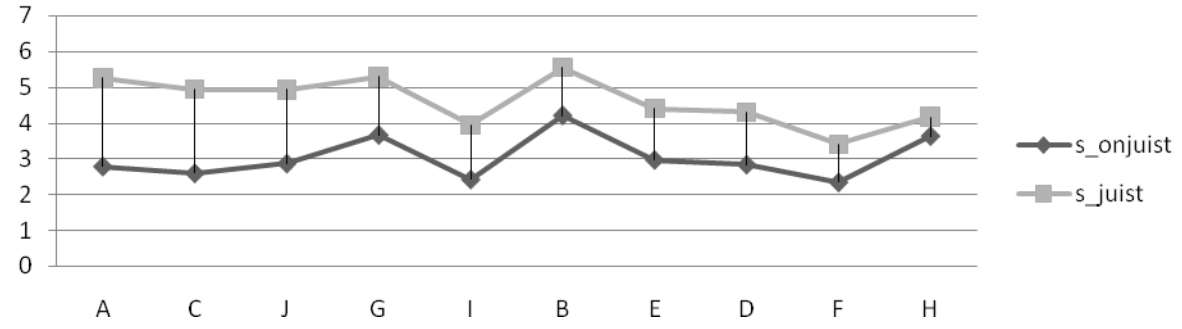
De antwoorden dienen direct gepresenteerd te worden teneinde ze hoger te laten scoren, aantrekkelijker en beter begrijpbaar te maken. Het belangrijkste is of gebruikers al dan niet moeten zoeken naar het antwoord.

4.2 Kwalitatieve analyse van de resultaten

In tabellen 3.3, 3.4 en 3.5 zijn een aantal waarden gearceerd. In dit gedeelte van de scriptie zullen we dieper op deze getallen ingaan. In deze sectie worden geen conclusies getrokken. Het reikt alleen mogelijke verklaringen en interessant materiaal aan voor verder onderzoek.

Bij negen van de tien voorgelegde vragen scoorde het juiste antwoord hoger dan een onjuist antwoord (tabel 3.3). Bij vraag H werd er echter geen significant verschil gevonden.

Grafiek 4.1: Scores van vragen op volgorde van verklaarde variantie



Vraag H: Wat is de oorzaak van een liesbreuk?

Juist antwoord (indirect):

UITWENDIGE BREUK Een uitwendige breuk is een uitstulping van het buikvlies door een al of niet van tevoren bestaande opening, met of zonder inhoud. Hierbij onderscheidt men een breukkanaal (liesbreuk, navelbreuk enz.), een breukzak, die door het buikvlies gevormd wordt, en een breukinhoud, meestal darm en/of net (omentum). Zolang alle onderdelen van een breuk vrij ten opzichte van elkaar beweeglijk zijn, kan men de breuk in de buikholte terugschuiven, maar dit kan belemmerd worden door vergroeiingen. ***Breuken kunnen aangeboren zijn, maar vaker zijn ze verworven door zwakke plekken in de buikwand, onder invloed van verhoogde druk, bijv. bij chronisch hoesten, obstipatie of gezwellen. Een liesbreuk*** *komt het meest voor (85% van alle breuken), en wel voornamelijk bij mannen. De navelbreuk komt vooral voor na de geboorte. Complicaties De voornaamste complicatie bij een breuk is de inklemming (beklemde breuk), die kan optreden wanneer door een verhoging van de druk de breukpoort uitgerekt wordt en bijv. een darmlis in de breukzak wordt geschoven. Wanneer de druk ophoudt, trekt de breukpoort gedeeltelijk dicht en zit de darmlis zodanig ingeklemd, dat de afvoer van het bloed uit de lis afgesnoerd wordt. Deze zwelt daardoor op en ten slotte worden ook de toevoerende slagaders dichtgedrukt en treedt versterf (gangreen) op. De verschijnselen zijn die van darmafsluiting. Behandeling Een beklemde breuk moet altijd geopereerd worden.*

Onjuist antwoord:

beklemd raken van een darmlis. Deze inklemming kan ontstaan als gevolg van een breuk (ingeklemde breuk; zie liesbreuk) of door het beklemd geraken van de darmlis in uitstulpingen van het buikvlies (inwendige beklemming). Soms ook kunnen in de buikholte membranen - die daar ontstaan zijn als gevolg van ontstekingen van het buikvlies, of van operaties - de oorzaak zijn van inklemming van een darmlis. Darminklemming kan aanleiding geven tot darmafsluiting.

De meningen over het onjuiste antwoord op vraag H waren gezien de grote standaardafwijking redelijke verdeeld: 3.63 (1.771). Het is dus goed mogelijk dat een aantal respondenten het onjuiste antwoord ook als juist zagen. 40% van de scores waren namelijk 5 of zelfs 6. Het juiste antwoord voor H scoorde bovendien ook niet erg hoog, vermoedelijk omdat het antwoord ongeveer op de helft van het stuk tekst staat (een 'indirect' antwoord) en de respondent daarom enig zoekwerk moest verrichten.

Als we bijvoorbeeld kijken naar vraag A, waar de verklaarde variantie het grootst was, zien we een vele mate kleinere standaardafwijking: 5.26 (1.182).

Vraag A: Wat is epilepsie?

Juist antwoord (direct):

een plotselinge, tijdelijke toestand van bewustzijnsverlies, gepaard gaande aan een kramptoestand van de spieren, die overgaat in een ritmisch schokken. Zie ook epilepsie.

Onjuist antwoord:

ziekte met onbekende oorzaak. Bij een aantal ziekten (bijv. epilepsie, sommige bloedziekten) kent men een idiopathische vorm naast vormen met een bekende oorzaak. Bij hoge bloeddruk (hypertensie) met onbekende oorzaak spreekt men niet van idiopathische, maar van essentiële hypertensie.

Wat uit de kwantitatieve analyse ook al naar voren is gekomen, is dat directe antwoorden beter scoren dan indirecte antwoorden. Het juiste antwoord op vraag A is een voorbeeld van een direct antwoord. De 'score' van dit antwoord had mogelijk nog hoger gelegen als er nog extra informatie werd weergegeven (Theune *et al.*, 2006). Hetzelfde geldt voor B (hoogste waardering juiste antwoord) en G, waarvan de juiste antwoorden ook hoog 'scoorden'.

Het vermoeden bestond dat de juiste antwoorden op A, B en, in iets minder mate, G hoog scoorden, omdat het onderwerp van de vraag en de medische informatie daaromtrent nog binnen het kennisgebied van de respondenten lagen. Daarom zijn twee mini-onderzoeken uitgevoerd, waarbij zes proefpersonen de tien vragen op volgorde van de kennis die ze over het onderwerp dachten te hebben (moz1) en op volgorde van interessantheid (van het onderwerp, moz2) moesten leggen. Aan de hand van de gemiddelden is een nieuwe rangschikking gemaakt en ter vergelijking naast de rangschikking voor 'score', 'begrijpelijkheid' en 'aantrekkelijkheid' van de juiste antwoorden gelegd. Zie hieronder het resultaat:

Tabel 4.2: rangschikking; moz1 – ik weet het meest over, moz2 – ik vind het meest interessant

Volgorde	1	2	3	4	5	6	7	8	9	10
Score	B	G	A	C	J	E	D	H	I	F
Begrijpelijkheid	A	B	G	J	F	I	D	E	C	H
Aantrekkelijkheid	G	B	D	A	J	C	E	I	F	H
moz1	J	A	I	C	D	E	H	G	F	B
moz2	J	I	C	A	E	B	H	F	D	G

Tabel 4.2 laat zien dat de 'hoogstscorende', de best te begrijpen en meest aantrekkelijke antwoorden op het eerste gezicht geen verband laten zien met hoeveel mensen kennis hebben over het onderwerp en hoe interessant mensen het onderwerp vinden. De 'laagstscorende', de slechts te begrijpen en de minst aantrekkelijke antwoorden laten op het eerste gezicht een licht verband zien.

Vraag F scoorde op zowel het juiste als het onjuiste antwoord het laagst. Een mogelijke verklaring is dat het lastiger is om de informatie uit het juiste antwoord te halen. Daarnaast verwacht je bij een vraag als deze, dat een explicieter onderscheid wordt gemaakt tussen de twee specialismen. Het antwoord geeft aan dát de twee beroepen verschillen, namelijk: een kaakchirurg is een tandarts én arts die de postacademische opleiding tot specialist met goed gevolg heeft doorlopen, maar wát het verschil nu daadwerkelijk is, is niet gegeven.

Vraag F: Is een kaakchirurg hetzelfde als een tandarts?

Juist antwoord (indirect):

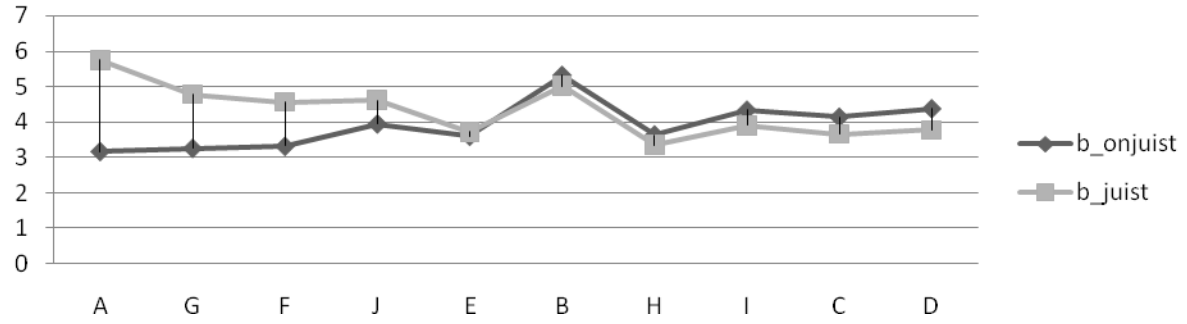
een tandarts én arts die de postacademische opleiding tot specialist mondheelkunde/kaakchirurg met goed gevolg heeft doorlopen. Het is een erkend specialisme.

Onjuist antwoord:

of stomatologie, de specialistische behandeling van afwijkingen en traumata van mond en kaken (zie kaak). In Nederland is het een specialisatie van de tandheelkunde, in verschillende andere landen, o.a. België, van de

geneeskunde. Zie ook kaakchirurg en tandarts.

Grafiek 4.3: 'begrijpelijkheid' van vragen op volgorde van verklaarde variantie



In grafiek 4.3 is te zien dat het juiste en onjuiste antwoord op vraag A respectievelijk het hoogst en het laagst scoren op 'begrijpelijkheid'. De aanwezigheid van moeilijke/niet alledaagse woorden als 'idiopathische' en 'essentiële hypertensie' geeft een mogelijke verklaring voor de lage waardering op 'begrijpelijkheid' voor het onjuiste antwoord op vraag A. In grafiek 4.4 is te zien dat het onjuiste antwoord op vraag A tevens laag scoort op 'aantrekkelijkheid'. Dit valt te verklaren aan de hand van de in het onderzoek bevestigde hypothese 2c, die stelt dat 'begrijpelijkheid' en 'aantrekkelijkheid' positief correleren.
Voor vraag B liggen de waarderingen op 'begrijpelijkheid' van het juiste en onjuiste antwoord hoog en dicht bij elkaar. Mogelijke verklaring: beide antwoorden zijn qua lengte bijna gelijk en bevatten weinig vakjargon.

Vraag B: Wat zijn aambeien?

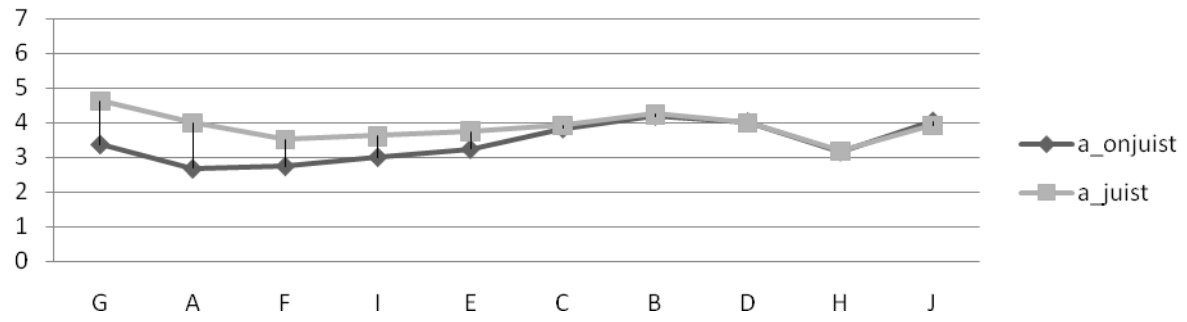
Juist antwoord (direct):

of hemorroïde, plaatselijke, onregelmatige verwijding van een van de sponsachtige kluwens van bloedvaten (adernetten; Lat. enk.: plexus) die in de wand van het onderste deel van de endeldarm en in die van de anus liggen. Zij zijn bedekt met een dun slijmvlies. Men onderscheidt inwendige aambeien (van buiten niet zichtbaar) en uitwendige aambeien (puilen buiten de anus uit als blauwrode knobbels, één tot enkele centimeters groot). Inwendige aambeien kunnen bij sterk persen tijdens bijv. de stoelgang naar buiten komen. Aambeien kunnen week of - als ze stolsels bevatten - hard en vaak zeer pijnlijk zijn.

Onjuist antwoord:

ONTSTAAN Aambeien ontstaan als de druk in de adernetten groter wordt dan de wand kan verdragen. Sommige mensen hebben aanleg voor het ontstaan van aambeien, maar ook wordt het ontstaan bevorderd door een belemmerde bloedafvoer (bijv. door een zittende levenswijze, hoog lichaamsgewicht, chronische verstopping, persen, zwangerschap) of door onvoldoende steun van de dunne aderwanden.

Grafiek 4.4: 'aantrekkelijkheid' van vragen op volgorde van verklaarde variantie



Vraag G: Hoe wordt de diagnose leukemie gesteld?

Juist antwoord (direct):

DIAGNOSE Standaardbloedonderzoek als een bloedbeeld kan de eerste aanwijzing opleveren dat een patiënt aan leukemie lijdt. Het totale aantal witte bloedcellen kan verlaagd, normaal of verhoogd zijn, maar de hoeveelheden rode bloedcellen en bloedplaatjes zijn vrijwel altijd verlaagd. Belangrijker is dat onder de microscoop zeer onrijpe witte bloedcellen (blasten) zichtbaar zijn. Aangezien het bloed gewoonlijk geen blasten bevat, wijst het voorkomen van blasten op de mogelijke aanwezigheid van leukemie. Er wordt echter altijd nog een beenmergbiopsie verricht om de diagnose te bevestigen en vast te stellen om welke vorm van leukemie het gaat.

Het antwoord op G scoort erg hoog op 'aantrekkelijkheid'. Hiervoor zijn onder andere de volgende redenen te bedenken: allereerst is het antwoord wederom een direct antwoord op de vraag. Bovendien matcht het eerste woord uit het antwoord een concept uit de vraag, namelijk "diagnose". Daarnaast zien we weinig moeilijke woorden.

Uit het bovenstaande kunnen we afleiden dat een goed systeem een hoge 'accuratesse' dient te hebben teneinde de gemiddelde waardering van 'begrijpelijkheid' en 'aantrekkelijkheid' te verhogen. Dit komt doordat leken het verschil kunnen zien tussen een juist en onjuist antwoord. Een goed antwoord dient direct te zijn. Dit betekent niet dat het antwoord kort moet zijn (nog steeds geldt hier het onderzoek van Theune *et al.* (2006)). Het belangrijkste is dat een gebruiker niet in een berg informatie moet zoeken. M.a.w. de eerste óf de tweede zin moet een antwoord geven op de gestelde vraag.

De resultaten uit de kwalitatieve bespreking verdient meer onderzoek. Hierover meer in het volgende hoofdstuk.

5 Discussie en conclusies

In dit onderzoek is gekozen voor een opzet waarin de respondenten aan de hand van een schaal een 'score' aan een antwoord gaven. De respondenten had ook gevraagd kunnen worden om aan te geven of een antwoord volgens hen al dan niet juist is. In deze scriptie is bewust gekozen voor de eerste manier. De tweede optie had mogelijk voor (meer) *priming* gezorgd; de indruk wekken dat niet alle antwoorden juist waren, zorgt naar alle waarschijnlijkheid voor dat respondenten met het beantwoorden van de schalen meer rekening gaan houden met de door hen gegeven waardering op 'juistheid'. Het risico van deze opzet is dat niet bekend is in welke mate de respondenten het verschil ook daadwerkelijk weten in plaats van denken. In vervolgonderzoek kan hiermee rekening gehouden worden door de schalen te scheiden van de vraag naar juistheid en door een extra vraag op te nemen die achterhaalt in hoeverre de respondenten zeker zijn van hun 'score'.

De resultaten van het onderzoek lieten een tegenstrijdigheid zien: de respondenten zagen geen verschil in 'tekstwaardering' en 'tekstbegrip' als het ging om juiste en onjuiste antwoorden. Toch correleerden de waarderingen voor 'score' wel met de twee schalen. Waarschijnlijk is het item 'score' niet geschikt om conclusie te trekken in de trant van 'de respondenten zien wat het juiste antwoord is'.

Mogelijk geven de schalen 'aantrekkelijkheid' en 'begrijpelijkheid' een vertekend beeld. De respondenten waren niet op de hoogte gesteld van het feit dat de antwoorden uit een geautomatiseerd systeem afkomstig waren. Naar alle waarschijnlijkheid waren de scores hoger wanneer dit bekend was gemaakt. Sommige respondenten hadden al het vermoeden dat de antwoorden door een computersysteem gegenereerd waren. De voor- en nadelen van het op deze manier onderzoeken zijn lastig uit te sluiten en om die reden was gekozen voor de opzet waarbij de resultaten het meest objectief waren.

Het onderzoek naar 'directheid' is in deze onderzoekssetting niet betrouwbaar. 'Directheid' is niet onderzocht als een aparte conditie. Er zijn meerdere factoren die van invloed waren op het resultaat, waaronder de lengte van het antwoord. De invloed van 'directheid' kan beter onderzocht worden door aan verschillende respondenten dezelfde alinea (met het antwoord) voor te leggen, maar het eigenlijke antwoord van plaats in de tekst te variëren. Mocht uit verder onderzoek blijken dat 'directheid' van invloed is op tekstwaardering, dan zullen hier ongetwijfeld nieuwe technieken voor moeten worden ontwikkeld die antwoorden herformuleren. Een belangrijk gegeven is dat 'directheid' zeer waarschijnlijk een grotere rol speelt in systemen die ontwikkeld zijn voor de leken.

Een combinatie van het waarderen van antwoorden volgens de schalen en het systeem beoordelen aan de hand van het schema volgens Diekema *et al.* (2004) maakt het mogelijk om het systeem zowel technisch als inhoudelijk te testen. Het is aan te raden de technische evaluatie te laten doen door (medisch) experts.

De in deze scriptie gepresenteerde methode zal nog verbeterd moeten worden. Er zal onderzocht moeten worden hoe de schalen het beste voorgelegd kunnen worden aan de gebruikersgroep. Is het laten beoordelen van onjuiste antwoorden bijvoorbeeld niet van invloed op de waardering van de juiste antwoorden? Ook de combinatie van vraag en antwoord is ongetwijfeld van invloed op 'tekstwaardering' en 'tekstbegrip'.

In vervolgonderzoek zal gezocht moeten worden naar kwantitatieve verklaringen voor de kwalitatief geanalyseerde resultaten in hoofdstuk 4. Een interessante onderzoeksvraag is: "waaraan zien leken dat een antwoord juist of onjuist is?"

Literatuurlijst

Anglia Ruskin University. 2009. *Harvard System of Referencing Guide* [Online] (Geüpdatet 26 Mei 2009)
Beschikbaar op: http://libweb.anglia.ac.uk/referencing/harvard.htm [Toegang op 27 Juli 2009]

Canisius, S., Bosch, A. van den & Daelemans, W. *IMIX Rolaquad Annotatiehandleiding: ILK Research Group Technical Report 04-04*

Costa, L.F. & Sarmento, L., 2006. Component Evaluation in a Question Answering System. *Proceedings of LREC 2006.* Genoa, Italy.

Diekema, A.R., Ozgur Y. & Liddy E.D., 2004. Evaluation of restricted domain question-answering systems. *Workshop on Question Answering in Restricted Domains. 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, pp.2-7.

Doan-Nguyen, Hai, Kosseim & Leila, 2004. The Problem Of Precision In Restricted-Domain Question Answering. Some Proposed Methods Of Improvement. *Workshop on Question Answering in Restricted Domains. 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, pp.8-15.

Ellen, Voorhees M. & Harman, D.K., 2005 *TREC: Experiment and Evaluation in Information Retrieval.* Cambridge, Massachusetts: MIT Press.

Gillard L., Bellot P. & El-bèze M., 2006. Question Answering Evaluation Survey. *5th International Conference on Language Resources and Evaluation (LREC2006)*, Genoa, Italy.

Lin, J., & Katz, B., 2006. Building a reusable test collection for question answering. *Journal of the American Society for Information Science and Technology, 57(7)*, pp.851-861.

Maes, A., Ummelen, N. & Hoeken, H., 1996. *Instructieve teksten. Analyse, ontwerp en evaluatie.* Bussum: Coutinho.

NTCIR (NII Test Collection for IR Systems) Project. 1997. *From Original Proposal: Scope of the Project* [Online] (Geüpdatet Maart 1998)
Beschikbaar op: http://research.nii.ac.jp/ntcir/outline/prop-en.html [Toegang op 20 Juli 2009].

Peters, C. 2002. *Evaluating Cross-Language Systems the CLEF Way* [Online] (Geüpdatet 11 Februari 2002)
Beschikbaar op: http://www.cultivate-int.org/issue6/clef/ [Toegang op 20 Juli 2009]

Theune, M., Krahmer, E., Bosma, W., Marsi, E. & Hooijdonk, C., 2006. Mag het ietsje meer zijn? Antwoordpresentatie in een QA-systeem. *DIXIT, tijdschrift over toegepaste taal- en spraaktechnologie, 4 (1),* pp.12-15.

Bijlagen

Bijlage 1: 10 vragen voor het onderzoek

Vraag A

Wat is epilepsie?

Juist antwoord (direct):

een plotselinge, tijdelijke toestand van bewustzijnsverlies, gepaard gaande aan een kramptoestand van de spieren, die overgaat in een ritmisch schokken. Zie ook epilepsie.

Onjuist antwoord:

ziekte met onbekende oorzaak. Bij een aantal ziekten (bijv. epilepsie, sommige bloedziekten) kent men een idiopathische vorm naast vormen met een bekende oorzaak. Bij hoge bloeddruk (hypertensie) met onbekende oorzaak spreekt men niet van idiopathische, maar van essentiële hypertensie.

Vraag B

Wat zijn aambeien?

Juist antwoord (direct):

of hemorroïde, plaatselijke, onregelmatige verwijding van een van de sponsachtige kluwens van bloedvaten (adernetten; Lat. enk.: plexus) die in de wand van het onderste deel van de endeldarm en in die van de anus liggen. Zij zijn bedekt met een dun slijmvlies. Men onderscheidt inwendige aambeien (van buiten niet zichtbaar) en uitwendige aambeien (puilen buiten de anus uit als blauwrode knobbels, één tot enkele centimeters groot). Inwendige aambeien kunnen bij sterk persen tijdens bijv. de stoelgang naar buiten komen. Aambeien kunnen week of - als ze stolsels bevatten - hard en vaak zeer pijnlijk zijn.

Onjuist antwoord:

ONTSTAAN Aambeien ontstaan als de druk in de adernetten groter wordt dan de wand kan verdragen. Sommige mensen hebben aanleg voor het ontstaan van aambeien, maar ook wordt het ontstaan bevorderd door een belemmerde bloedafvoer (bijv. door een zittende levenswijze, hoog lichaamsgewicht, chronische verstopping, persen, zwangerschap) of door onvoldoende steun van de dunne aderwanden.

Vraag C

Wat is een de Latijnse benaming voor suikerziekte?

Juist antwoord (indirect[13]):

Diabetes mellitus ('suikerziekte') is een aandoening waarbij de bloedspiegels van glucose (een eenvoudige suiker) abnormaal hoog zijn doordat het lichaam niet voldoende insuline afgeeft of dit niet adequaat gebruikt. Vaak wordt de volledige naam diabetes mellitus gebruikt in plaats van alleen diabetes om deze aandoening te onderscheiden van de betrekkelijk zeldzame aandoening diabetes insipidus. De bloedglucosespiegel varieert in de loop van de dag. De spiegel stijgt na een maaltijd en daalt binnen twee uur weer tot een normaal niveau. De normaalwaarde van de bloedglucosespiegel ligt 's ochtends in nuchtere toestand normaal tussen 4,0 en 6,4 mmol/l bloed. Twee uur na het gebruiken van voedsel of drank met suiker of koolhydraten is deze bloedglucosespiegel gewoonlijk lager dan 6,7 tot

[13] Wellicht rijst de vraag waarom het juiste antwoord op vraag C geen direct antwoord is; het antwoord wordt immers toch op de eerste regel gegeven? De reden dat het juiste antwoord in dit onderzoek geen direct antwoord is, is omdat de lezer niet uit de tekst kan opmaken dat 'Diabetes mellitus' ook daadwerkelijk de Latijnse benaming is.

7,8 mmol/l. De normaalwaarden vertonen gewoonlijk een lichte maar toenemende stijging bij personen boven de 50 jaar, vooral bij mensen die een zittend leven leiden. Insuline (een hormoon afgegeven door de alvleesklier) speelt de hoofdrol bij het handhaven van de juiste bloedglucosespiegel. Insuline maakt het mogelijk glucose naar de cellen te transporteren zodat deze energie kunnen produceren of glucose kunnen opslaan tot deze nodig is. De stijging van de bloedglucosespiegel na eten of drinken prikkelt de alvleesklier tot insulineproductie, wat een verdere stijging van de bloedglucosespiegel voorkomt en ervoor zorgt dat deze geleidelijk daalt. Omdat de spieren glucose gebruiken om in hun energiebehoefte te voorzien, kan de bloedglucosespiegel ook dalen tijdens lichamelijke activiteit.

Onjuist antwoord:

(afk.: GTT), onderzoek dat wordt uitgevoerd om suikerziekte op te sporen. De test wordt voorafgegaan door een koolhydraatrijk dieet gedurende drie dagen. Vervolgens moet de patiënt nuchter naar het laboratorium komen waar bloed wordt afgenomen; daarna drinkt hij wat suikerwater. Om het halve uur, twee uur lang, wordt bloed afgenomen. In deze bloedmonsters wordt dan het bloedsuikergehalte bepaald en uit deze uitslagen is te zien hoe het lichaam op de toegediende suiker heeft gereageerd; m.a.w. of de patiënt suikerziekte heeft of niet.

Vraag D

Wat zijn de overeenkomsten van hepatitis a, b, c en e?

Juist antwoord (indirect):

Hepatitis is een ontsteking van de lever door een willekeurige oorzaak. Hepatitis wordt meestal veroorzaakt door een virus, voornamelijk een van de vijf hepatitisvirussen: A, B, C, D of E. Hepatitis komt minder vaak voor ten gevolge van een andere virusinfectie, zoals de ziekte van Pfeiffer, gele koorts en cytomegalievirusinfectie. De belangrijkste niet-virale oorzaken van hepatitis zijn overmatig alcoholgebruik en geneesmiddelen. Hepatitis kan acuut verlopen (korter dan een halfjaar) of chronisch. De ziekte doet zich over de gehele wereld voor. Het hepatitis-A-virus wordt overgebracht via de ontlasting van de één naar de mond van de ander. Een dergelijk overdracht is meestal het gevolg van slechte hygiëne. Vooral in ontwikkelingslanden komen door water of voedsel overgebrachte hepatitis-A-besmettingen veel voor. Soms is het eten van rauwe schelpdieren de oorzaak. Ook komen vaak opzichzelfstaande gevallen voor, meestal door overdracht van persoon op persoon. De meeste hepatitis-A-infecties veroorzaken geen symptomen en gaan onopgemerkt voorbij. Het hepatitis-B-virus wordt minder gemakkelijk overgebracht dan het hepatitis-A-virus. Het virus kan via besmet bloed of besmette bloedproducten worden overgedragen. In de Verenigde Staten is bloedtransfusie echter zelden de oorzaak van besmetting met het hepatitis-B-virus, omdat voldoende voorzorgsmaatregelen worden genomen om besmetting via bloed te voorkomen. Vaak vindt overdracht plaats tussen drugsgebruikers die dezelfde naalden gebruiken of via seksueel contact tussen zowel heteroseksuele als homoseksuele partners. Zwangere vrouwen met hepatitis B kunnen het virus tijdens de geboorte overdragen op hun baby. De kans op besmetting met het hepatitis-B-virus is groter bij patiënten die gedialyseerd worden, bij kankerpatiënten en bij ziekenhuispersoneel dat veel met bloed in aanraking komt. Ook is de kans groter bij mensen die zich in een afgesloten omgeving bevinden (zoals gevangenissen en psychiatrische inrichtingen), waar rechtstreeks lichamelijk contact plaatsvindt. Hepatitis B kan worden overgebracht door gezonde mensen die chronisch drager zijn van het virus. Het is onduidelijk of het virus kan worden overgebracht via insectenbeten. In veel gevallen is de oorzaak van hepatitis onbekend. In bepaalde delen van de wereld, zoals het Verre Oosten en delen van Afrika, leidt hepatitis B tot veel gevallen van chronische hepatitis, cirrose en leverkanker. Het hepatitis-C-virus veroorzaakt minstens 80% van de gevallen van hepatitis ten gevolge van bloedtransfusie, plus veel verspreide gevallen van acute hepatitis. Het virus wordt het meest overgebracht door drugsgebruikers die elkaars naalden gebruiken. Overdracht via seksueel contact is zeldzaam. Hepatitis C is verantwoordelijk voor veel gevallen van chronische hepatitis en enkele gevallen van cirrose en leverkanker. Om onbekende redenen hebben mensen met een alcoholische leverziekte vaak ook hepatitis C. Deze combinatie van ziekten leidt tot een groter functieverlies van de lever dan één van deze ziekten alleen zou veroorzaken. Een klein aantal gezonde mensen blijkt chronisch drager van hepatitis C te zijn. Het hepatitis-D-virus komt alleen voor als coïnfectie bij hepatitis B en maakt de infectie met hepatitis B ernstiger. Drugsverslaafden lopen een relatief groot risico. Het hepatitis-E-virus veroorzaakt

af en toe epidemëeen die vergelijkbaar zijn met epidemieën die door hepatitis A zijn veroorzaakt. Tot nu toe zijn deze epidemieën alleen in ontwikkelingslanden voorgekomen.

Onjuist antwoord:

KANS OP EEN SOA In Nederland lopen elk jaar ca. 100 000 mensen een SOA op. De kans op een SOA is zeer klein wanneer men veilig vrijt (zie veilig vrijen). Er is echter een aantal SOA's dat ook op andere wijze kan worden overgebracht. Hiv-infectie en hepatitis B kunnen ook worden overgedragen via bloed-bloedcontact. Dit kan gebeuren als drugsgebruikers gebruikte naalden en spuiten aan elkaar doorgeven en bij tatoeage of piercing met niet-steriele apparatuur of op niet-hygiënische wijze. Een onbehandelde SOA maakt de kans op de overdracht van en besmetting met het hiv-virus tienmaal zo groot. Bij de geboorte kan een baby besmet raken als de moeder besmet is met: hiv-virus, hepatitis B-virus, syfilisbacteriën, chlamydia's, herpes-simplexvirus of gonokokken.

Vraag E

Wat doet hemoglobine?

Juist antwoord (direct):

(afk.: Hb) of bloedkleurstof, rood gekleurd eiwit, dat als bloedkleurstof voorkomt in de rode bloedcellen en dat het zuurstoftransport naar de weefsels als functie heeft. In de longen is een hoog percentage van het hemoglobine aan zuurstof gebonden, in zuurstofarme weefsels wordt zuurstof afgegeven. Behalve aan zuurstof bindt hemoglobine zich ook sterk aan koolmonoxide, waardoor het zuurstoftransport vermindert. In het menselijk lichaam komen verschillende hemoglobinen voor, die zich van elkaar onderscheiden in het eiwitgedeelte. Normaal is het hemoglobine A, bestaande uit vier ketens van elk 146 aminozuren. In ongeboren kinderen en in pasgeborenen komt het foetale hemoglobine (Hb-F) voor, gekenmerkt door een groter bindingsvermogen voor zuurstof. Na de geboorte wordt het foetale hemoglobine in ca. 4 maanden volledig afgebroken, wat een oorzaak is van de bij pasgeborenen voorkomende geelzucht.

Onjuist antwoord:

SYMPTOMEN Bij vergiftiging met cyaanverbindingen kleurt de huid blauw tot blauwachtig zwart. Het zijn ademhalingsvergiften die de hemoglobine van het bloed onwerkzaam maken.

Vraag F

Is een kaakchirurg hetzelfde als een tandarts?

Juist antwoord (indirect):

een tandarts én arts die de postacademische opleiding tot specialist mondheelkunde/kaakchirurg met goed gevolg heeft doorlopen. Het is een erkend specialisme.

Onjuist antwoord:

of stomatologie, de specialistische behandeling van afwijkingen en traumata van mond en kaken (zie kaak). In Nederland is het een specialisatie van de tandheelkunde, in verschillende andere landen, o.a. België, van de geneeskunde. Zie ook kaakchirurg en tandarts.

Vraag G

Hoe wordt de diagnose leukemie gesteld?

Juist antwoord (direct):

DIAGNOSE Standaardbloedonderzoek als een bloedbeeld kan de eerste aanwijzing opleveren dat een patiënt aan leukemie lijdt. Het totale aantal witte bloedcellen kan verlaagd, normaal of verhoogd zijn, maar de

hoeveelheden rode bloedcellen en bloedplaatjes zijn vrijwel altijd verlaagd. Belangrijker is dat onder de microscoop zeer onrijpe witte bloedcellen (blasten) zichtbaar zijn. Aangezien het bloed gewoonlijk geen blasten bevat, wijst het voorkomen van blasten op de mogelijke aanwezigheid van leukemie. Er wordt echter altijd nog een beenmergbiopsie verricht om de diagnose te bevestigen en vast te stellen om welke vorm van leukemie het gaat.

Onjuist antwoord:

VORMEN Allereerst is er een onderscheid tussen acute leukemie en chronische leukemie. Deze indeling is gebaseerd op de mate van uitrijping van de abnormale cellen. Bij acute leukemie rijpen de cellen vrijwel niet uit en vindt er in korte tijd een ophoping van onrijpe cellen plaats. Deze onrijpe cellen worden ook wel blasten genoemd. Acute leukemie geeft vaak in korte tijd klachten. Bij chronische leukemie is het proces trager. De cellen rijpen nog redelijk goed uit. Chronische leukemie geeft minder snel klachten en wordt dan ook vaak bij toeval ontdekt als er om een andere reden bloedonderzoek wordt gedaan. Naast het onderscheid tussen acute en chronische leukemie is er een onderscheid te maken tussen myeloïde leukemie en lymfatische leukemie. Dit heeft te maken met het type witte bloedcel (leukocyt) dat de leukemie veroorzaakt. Bij de myeloïde leukemie zijn de granulocyten betrokken en bij de lymfatische leukemie de lymfocyten. Samenvattend zijn er dus vier belangrijke soorten leukemie: acute myeloïde leukemie (AML), acute lymfatische leukemie (ALL), chronische myeloïde leukemie (CML) en chronische lymfatische leukemie (CLL). Acute lymfatische leukemie komt vooral bij kinderen voor, acute myeloïde leukemie komt bij jongvolwassenen en volwassenen voor. Chronische myeloïde leukemie wordt met name op de middelbare leeftijd vastgesteld, terwijl chronische lymfatische leukemie voornamelijk bij oudere mensen voorkomt.

Vraag H

Wat is de oorzaak van een liesbreuk?

Juist antwoord (indirect):

UITWENDIGE BREUK Een uitwendige breuk is een uitstulping van het buikvlies door een al of niet van tevoren bestaande opening, met of zonder inhoud. Hierbij onderscheidt men een breukkanaal (liesbreuk, navelbreuk enz.), een breukzak, die door het buikvlies gevormd wordt, en een breukinhoud, meestal darm en/of net (omentum). Zolang alle onderdelen van een breuk vrij ten opzichte van elkaar beweeglijk zijn, kan men de breuk in de buikholte terugschuiven, maar dit kan belemmerd worden door vergroeiingen. Breuken kunnen aangeboren zijn, maar vaker zijn ze verworven door zwakke plekken in de buikwand, onder invloed van verhoogde druk, bijv. bij chronisch hoesten, obstipatie of gezwellen. Een liesbreuk komt het meest voor (85% van alle breuken), en wel voornamelijk bij mannen. De navelbreuk komt vooral voor na de geboorte. Complicaties De voornaamste complicatie bij een breuk is de inklemming (beklemde breuk), die kan optreden wanneer door een verhoging van de druk de breukpoort uitgerekt wordt en bijv. een darmlis in de breukzak wordt geschoven. Wanneer de druk ophoudt, trekt de breukpoort gedeeltelijk dicht en zit de darmlis zodanig ingeklemd, dat de afvoer van het bloed uit de lis afgesnoerd wordt. Deze zwelt daardoor op en ten slotte worden ook de toevoerende slagaders dichtgedrukt en treedt versterf (gangreen) op. De verschijnselen zijn die van darmafsluiting. Behandeling Een beklemde breuk moet altijd geopereerd worden.

Onjuist antwoord:

beklemd raken van een darmlis. Deze inklemming kan ontstaan als gevolg van een breuk (ingeklemde breuk; zie liesbreuk) of door het beklemd geraken van de darmlis in uitstulpingen van het buikvlies (inwendige beklemming). Soms ook kunnen in de buikholte membranen - die daar ontstaan zijn als gevolg van ontstekingen van het buikvlies, of van operaties - de oorzaak zijn van inklemming van een darmlis. Darminklemming kan aanleiding geven tot darmafsluiting.

Vraag I

Wat veroorzaakt een hersenbloeding?

Juist antwoord (indirect):

BLOEDIGE BEROERTES De `bloedige beroertes' maken ca. 20% uit van het totaal. De oorzaak van de bloedige beroerte kan in de hersenen gelegen zijn (intracerebraal) of daarbuiten (extracerebraal). Bij de in de hersenen gelegen oorzaken betreft het doorgaans een bloeding uit een van tevoren zwakke plek in een slagader (men spreekt dan van een hersenbloeding). Bij de buiten de hersenen gelegen bloedingen gaat het meestal om een bloeding tussen de hersenvliezen. Het betreft hier voornamelijk de arachnoïdale bloeding en het subduraal hematoom. Zeldzaam is het epiduraal hematoom.

Onjuist antwoord:

BEHANDELING Er is geen speciale behandeling voor de hersenbloeding. In sommige gevallen kan een operatie om de bloeding te verwijderen de kans op herstel vergroten. Zie ook subduraal hematoom.

Vraag J

Wat gebeurt er bij flauwvallen?

Juist antwoord (direct):

een plotseling tekortschieten van de bloedsomloop in de hersenen, hetgeen zich uit als flauwvallen, bijv. na lang staan, door een te laag bloedsuikergehalte of door nervositeit. Een langdurige collaps duidt vrijwel altijd op shock.

Onjuist antwoord:

HARTKLOPPINGEN Normaal gesproken zijn mensen zich niet bewust van het kloppen van hun hart. Maar onder bepaalde omstandigheden, bij gezonde mensen bijvoorbeeld bij zware inspanning of bij een hevige psychische schok, kan het gebeuren dat ze hun hartslag plotseling voelen. Ze kunnen de indruk krijgen dat hun hart opvallend sterk, snel of onregelmatig klopt. Een arts kan dergelijke symptomen bevestigen door de polsslag te voelen en de hartslag te beluisteren met behulp van een stethoscoop op de borst. Of dergelijke hartkloppingen een teken zijn dat er iets mis is, hangt af van een aantal vragen, zoals of deze verschijnselen altijd onder bepaalde omstandigheden optreden, of ze plotseling of geleidelijk optreden, hoe snel het hart klopt en of de hartslag onregelmatig is, en zo ja, in welke mate. Als de hartkloppingen optreden in combinatie met symptomen als kortademigheid, pijn, zwakte en vermoeidheid of flauwvallen, is er een grotere kans dat ze het gevolg zijn van een abnormaal hartritme of een ernstige onderliggende aandoening.

Bijlage 2: Vragentestset

In deze bijlage de 185 vragen die gebruikt zijn om de MRR van Rolaquad te berekenen. De kolom JA (juiste antwoord) geeft weer op welke plek het juiste antwoord te vinden was. De kolom RR geeft de Reciprocal Rank van de vraag.

Nr	*Vraag*	*JA*	*RR*
	1. Definitie[14]: Aandoening		
1	**Wat is epilepsie?**	**2**	**0,50**
2	**Wat houdt spina bifida in?**	**2**	**0,50**
3	**Wat zijn aambeien?**	**5**	**0,20**
4	**Wat zijn beroepsziektes?**	**0**	**0,00**
5	**Bij hoeveel glazen alcohol per dag spreekt men van alcoholisme?**	**0**	**0,00**
6	**Wat is een emotionele verslaving?**	**0**	**0,00**
7	**Waar komt multiple sclerose het meest voor?**	**0**	**0,00**
	1. Synoniem: Aandoening		
8	**Is myotone dystrofie hetzelfde als de ziekte van Steinert?**	**0**	**0,00**
9	**Hoe wordt een furunkel ook wel genoemd?**	**0**	**0,00**
10	**Hoe wordt een aangeboren alvleesklierinsufficiëntie nog meer genoemd?**	**0**	**0,00**
11	**Wat is een de Latijnse benaming voor suikerziekte?**	**5**	**0,20**
12	**Onder welke naam staat de menopauze ook wel bekend?**	**0**	**0,00**
	1. Soort van: Aandoening		
13	**Is Trichomonas een soa?**	**0**	**0,00**
14	**Welke hartkwalen zijn er?**	**0**	**0,00**
15	**Welke vormen van dementie zijn er?**	**0**	**0,00**
16	**Is de ziekte van waldenström een soort kanker?**	**0**	**0,00**
17	**Valt de ziekte van duchenne onder de aandoening spierdystrofieën?**	**0**	**0,00**
	1. Lijkt Op: Aandoening		
18	**Op welke andere ziekte lijkt de ziekte van Crohn?**	**0**	**0,00**
19	**Is hoofdpijn hetzelfde als migraine?**	**0**	**0,00**
20	**Wat is het verschil tussen jicht en pseudo-jicht?**	**0**	**0,00**
21	**Wat zijn de overeenkomsten van hepatitis a, b, c en e?**	**2**	**0,50**
22	**Wanneer spreekt men van aids en wanneer van een hiv-infectie?**	**0**	**0,00**
23	**Wat is het verschil tussen anorexia nervosa en boulimia nervosa?**	**0**	**0,00**
	2. Definitie: Micro-organisme/stoffen		
24	**Welke allergenen zijn te onderscheiden?**	**0**	**0,00**
25	**Wat is het verschil tussen een bacterie en virus?**	**0**	**0,00**
26	**Wat is Candida?**	**0**	**0,00**
27	**Wat is Listeria?**	**0**	**0,00**
28	**Wat is Staphylococcus aureus?**	**1**	**1,00**
29	**Onder welke omstandigheden vermenigvuldigen bacterieën zich het best?**	**0**	**0,00**
	2. Synoniem: Micro-organisme/stoffen		
30	**Wat is een andere naam voor ringworm?**	**2**	**0,50**
31	**Wat is een synoniem voor vibrio?**	**0**	**0,00**
32	**Onder welke naam staat clostridium botulinum beter bekend?**	**0**	**0,00**
33	**Hoe wordt rabiës ook wel genoemd?**	**0**	**0,00**
34	**Hoe staat H5n1 ook bekend?**	**0**	**0,00**
	3. Definitie: Lichaamsdeel of –stof		
35	**Wat zijn venen?**	**0**	**0,00**
36	**Welke kleur heeft bloed?**	**0**	**0,00**
37	**Hoeveel eierstokken heeft een vrouw?**	**0**	**0,00**
38	**Hoe snel groeit haar?**	**0**	**0,00**
39	**Hoe kan het dat er verschillende kleuren ogen zijn?**	**0**	**0,00**

[14] We spreken hier over 'definitie' in de breedste zin van het woord. Ook de lokatie van voorkomen hoort hier bijvoorbeeld bij.

40	**Is creatine een lichaamseigen stof?**	**0**	**0,00**
	3. Synoniem: Lichaamsdeel of –stof		
41	**Wat is de Latijnse benaming voor monnikskapspier?**	**0**	**0,00**
42	**Heeft bloed nog een andere benaming?**	**0**	**0,00**
43	**Wat is een andere naam voor urine?**	**0**	**0,00**
44	**Is pus hetzelfde als etter?**	**2**	**0,50**
45	**Wat is cerumen?**	**0**	**0,00**
	4. Definitie: Lichaamsfunctie		
46	**Welke functie heeft een [body_part celwand]?**	**0**	**0,00**
47	**Waarvoor dient oorsmeer?**	**0**	**0,00**
48	**Wat doet hemoglobine?**	**4**	**0,25**
49	**Hoe werkt de tastzin?**	**0**	**0,00**
50	**Hoe is de longinhoud te meten?**	**0**	**0,00**
51	**Hoe werkt het menselijke afweermechanisme?**	**0**	**0,00**
	4. Synoniem: Lichaamsfunctie		
52	**Hoe wordt het afweersysteem ook genoemd?**	**0**	**0,00**
53	**Wat zijn andere namen voor kotsen?**	**0**	**0,00**
54	**Wat is een synoniem voor urineren?**	**0**	**0,00**
55	**Wat is een ander woord voor geeuwen?**	**0**	**0,00**
56	**Hoe staat digestie ook wel bekend?**	**0**	**0,00**
	5. Definitie: Specialist		
57	**Wat is een uroloog?**	**0**	**0,00**
58	**Welke functie heeft een klinisch patholoog?**	**0**	**0,00**
59	**Wat doet een KNO arts?**	**0**	**0,00**
60	**Op welk gebied is een cardioloog gespecialiseerd?**	**0**	**0,00**
61	**Wat is een osteopaat?**	**0**	**0,00**
62	**Is een huisarts een specialist?**	**0**	**0,00**
	5. Synoniem: Specialist		
63	**Is een kaakchirurg hetzelfde als een tandarts?**	**1**	**1,00**
64	**Is een verloskundige hetzelfde als een gynaecoloog?**	**1**	**1,00**
65	**Wat is een andere naam voor een voetdeskundige?**	**0**	**0,00**
66	**Is een psychiater iets anders dan een psycholoog?**	**0**	**0,00**
67	**Hoe wordt pathologische anatomie tegenwoordig genoemd?**	**0**	**0,00**
	A. Aandoening – Symptomen		
68	**Welke symptomen heeft aids?**	**0**	**0,00**
69	**Waaraan herken ik dat ik griep heb?**	**0**	**0,00**
70	**Wat zijn de klachten bij leukemie?**	**0**	**0,00**
71	**Waarom heb ik diarree?**	**0**	**0,00**
72	**Hoe kan ik een herkennen? diagnoses, prevents**	**0**	**0,00**
73	**Mijn is waar kan dit aan liggen?**	**0**	**0,00**
	B. Aandoening – Komt voor bij		
74	**Komt een chlamydia-infectie meer voor bij mannen?**	**0**	**0,00**
75	**Welke mensen hebben een verhoogde kans op hart- en vaatziekten?**	**0**	**0,00**
76	**Vanaf welke leeftijd begint botontkalking?**	**0**	**0,00**
77	**Waarom komt cellulites meer voor bij vrouwen?**	**0**	**0,00**
78	**Kunnen negroïde mensen verbranden?**	**0**	**0,00**
79	**Bij welke groepen mensen komt agyrie voor?**	**0**	**0,00**
80	**Bij welke type huid komen basaliomen het meest voor?**	**0**	**0,00**
	D. Aandoening – Eigenschap		
81	**Is hemofilie erfelijk?**	**1**	**1,00**
82	**Is alkaptonurie te genezen?**	**0**	**0,00**
83	**Is de verslaving die kan ontstaan bij het binnenkrijgen van teveel cafeïne emotioneel of lichamelijk?**	**0**	**0,00**
84	**Kan de ziekte van hodgkin terugkeren?**	**0**	**0,00**
85	**Zijn verlammingsverschijnselen als gevolg van het Guillain-barré syndroom blijvend?**	**0**	**0,00**
86	**Wanneer is een cyste kwaadaardig?**	**0**	**0,00**
	E. Aandoening – Diagnose		
87	**Hoe stelt mijn dokter borstkanker vast?**	**0**	**0,00**

88	**Op welke manier is autisme te herkennen?**	**0**	**0,00**
89	**Hoe kunnen blaassamentrekkingen worden aangetoond?**	**1**	**1,00**
90	**Welke methode gebruikt men om onvruchtbaarheid bij mannen vast te stellen?**	**0**	**0,00**
91	**Wat kan wijzen op slaaptekort?**	**0**	**0,00**
92	**Hoe wordt de diagnose leukemie gesteld?**	**1**	**1,00**
	F. Aandoening – Micro-organisme		
93	**Welke bacterieën veroorzaken een infectie?**	**0**	**0,00**
94	**Hoe ontstaat een longontsteking?**	**0**	**0,00**
95	**Welke bacterie veroorzaakt er reizigersdiarree?**	**1**	**1,00**
96	**Hoe kan ik besmet worden met het rota virus?**	**0**	**0,00**
97	**Welke bacterie kan spontane abortus veroorzaken?**	**0**	**0,00**
	G. Aandoening – Lichaamsdeel		
98	**Naar welke organen kan een bijbaltumor zich uitzaaien?**	**0**	**0,00**
99	**Waar komen likdoorns voornamelijk voor?**	**1**	**1,00**
100	**Wat heb ik als ik een aneurysma heb?**	**0**	**0,00**
101	**Welke hartaandoeningen komen vaak voor bij mensen met het syndroom van down?**	**0**	**0,00**
102	**Wat is dysmelie?**	**0**	**0,00**
103	**Wat zijn klompvoeten?**	**0**	**0,00**
	H. Aandoening – Overdrachtsvorm/Oorzaak		
104	**Hoe komt het dat ik cholera heb?**	**0**	**0,00**
105	**Wat is de oorzaak van een liesbreuk?**	**3**	**0,33**
106	**Wat veroorzaakt een hersenbloeding?**	**7**	**0,14**
107	**Kan ik herpes krijgen van zoenen met iemand die hiermee besmet is?**	**0**	**0,00**
108	**Kan ik door zonnen uitdrogen?**	**0**	**0,00**
109	**Welke bacteriën veroorzaken krentenbaard?**	**0**	**0,00**
110	**Aan welke stof heeft het lichaam een tekort bij het adrenogenitaal syndroom?**	**0**	**0,00**
111	**Hoe ontstaan allergische reacties?**	**0**	**0,00**
112	**Wat gebeurt er bij flauwvallen?**	**3**	**0,33**
113	**Door welke bacterie wordt framboesia veroorzaakt?**	**0**	**0,00**
	I. Aandoening – Lichaamsfunctie		
114	**Waarom menstrueer ik niet? prevents**	**0**	**0,00**
115	**Waarom krijg ik pijn in mijn buik als ik na het eten ga sporten?**	**0**	**0,00**
116	**Waarom zweet ik als ik koorts heb?**	**0**	**0,00**
117	**Waarom word ik geel als ik een leverdisfunctie heb?**	**0**	**0,00**
118	**Waarom heb ik een branderig gevoel bij het plassen?**	**0**	**0,00**
	J. Aandoening – Advies		
119	**Hoe kan ik prostaatkanker voorkomen?**	**0**	**0,00**
120	**Is rugpijn na schaatsen te voorkomen?**	**0**	**0,00**
121	**Hoe kun ik me behoeden voor een te hoog cholesterol?**	**1**	**1,00**
122	**Waar beschermen antioxidanten mij tegen?**	**0**	**0,00**
123	**Hoe voorkom ik een soa?**	**0**	**0,00**
124	**Kunnen twee mensen met het downsyndroom gezonde kinderen krijgen?**	**0**	**0,00**
	K. Aandoening – Duur/Tijdstip/Periode		
125	**Hoe lang duurt het voor een botbreuk om te genezen?**	**0**	**0,00**
126	**Waarom ben ik langer dan 9 maanden zwanger?**	**0**	**0,00**
127	**Na hoeveel tijd verdwijnen wratten?**	**0**	**0,00**
128	**Wanneer heb ik het meest last van hooikoorts?**	**0**	**0,00**
129	**Na hoeveel tijd kan een ziekte chronisch worden?**	**0**	**0,00**
130	**Binnen hoeveel tijd verdwijnen aften?**	**0**	**0,00**
131	**Wat is de normale hoeveelheid slaap dat iemand benodigd?**	**0**	**0,00**
132	**Na hoeveel tijd kan syfilis worden getest?**	**0**	**0,00**
133	**Op welk moment van de dag zijn de symptomen van griep het hevigst?**	**0**	**0,00**
	L. Aandoening – Behandeling		
134	**Wat helpt tegen rosacea?**	**0**	**0,00**
135	**Welke oplossingen bestaan er voor bedplassen?**	**0**	**0,00**
136	**Hoe is gingvitis te behandelen?**	**0**	**0,00**
137	**Hoe kan acromegalie in de actieve fase worden behandeld?**	**0**	**0,00**
138	**Bij welke aandoeningen kan chiropractie uitkomst bieden?**	**1**	**1,00**
139	**Mag paracetamol gebruikt worden om koorts tegen te gaan bij iemand met een hoge bloeddruk?**	**0**	**0,00**
	M. Behandeling – Duur/Tijdstip/Periode		
140	**Wat is de overlevingskans na vijf jaar bij leverkanker?**	**0**	**0,00**

141	**Om de hoeveel tijd wordt een wrat behandeld bij cryotherapie?**	**0**	**0,00**
142	**Waartoe dient de 20-wekenecho?**	**0**	**0,00**
143	**Hoe lang is de werking van een volledige tetanus vaccinatie?**	**0**	**0,00**
144	**Hoeveeldaags dien ik feneticilline in te nemen?**	**0**	**0,00**
	N. Bijwerking – Duur/Tijdstip/Periode		
145	**Kan nurofen gebruikt worden tijdens de laatste drie maanden van de zwangerschap?**	**0**	**0,00**
146	**Gaat bij het gebruik van laxeermiddelen verstopping na verloop van tijd over?**	**0**	**0,00**
147	**Binnen hoeveel tijd na chemotherapie kan haaruitval plaatsvinden?**	**0**	**0,00**
148	**Hoe lang duren bij een heroineverslaving ziekteverschijnselen bij afkicken?**	**0**	**0,00**
149	**In welke periode na het inbrengen van implanon is hoofdpijn mogelijk?**	**0**	**0,00**
	O. Behandeling – Bijwerking		
150	**Wat zijn de bijwerkingen van hormoontherapie?**	**0**	**0,00**
151	**Wat is lipodystrofie?**	**0**	**0,00**
152	**Heeft psychostimulantia bijwerkingen?**	**1**	**1,00**
153	**Wat zijn de bijwerkingen als ik de medicijnen budesonide en formotero combineer?**	**0**	**0,00**
154	**Welke verschijnselen kunnen optreden bij behandeling met SSRI antidepressiva?**	**0**	**0,00**
	P. Symptomen – Duur/Tijdstip/Periode		
155	**Hoe snel na een tekenbeet kan rode huiduitslag zichtbaar zijn?**	**0**	**0,00**
156	**Na hoeveel tijd moet hoesten bij een verkoudheid normaliter verdwijnen?**	**0**	**0,00**
157	**Hoe lang na besmetting met de ziekte van Pfeiffer uit zich doorgaans vermoeidheid?**	**0**	**0,00**
158	**Welke symptomen kunnen zich enkele weken na de krab van een kat uiten?**	**0**	**0,00**
159	**In welke periode is er de grootste kans op astma door hooikoorts?**	**0**	**0,00**
160	**Wat is incubatietijd?**	**0**	**0,00**
	Q. Symptomen – Advies		
161	**Hoe is een slechte adem te voorkomen?**	**0**	**0,00**
162	**Hoe zorg ik dat ik niet vervel?**	**0**	**0,00**
163	**Hoe kan ik de klachten aan de bekkenbodem na een bevalling voorkomen?**	**0**	**0,00**
164	**Hoe bestrijd ik de jeuk bij schaamluis?**	**0**	**0,00**
165	**Hoe kan ik voorkomen dat ik droge ogen krijg bij het dragen van lenzen?**	**0**	**0,00**
	R. Komt voor bij – Symptomen		
166	**In welke leeftijdscategorie komen puisten het meeste voor?**	**0**	**0,00**
167	**Waaraan merkt een man dat hij besmet is met gonoeroe?**	**0**	**0,00**
168	**Wat zijn de symptomen van puberen?**	**0**	**0,00**
169	**Welke mensen raken sneller in paniek?**	**1**	**1,00**
170	**Rond welke leeftijd komt haaruitval door discoïde lupus erythematodes het meest voor?**	**0**	**0,00**
	S. Symptomen – Behandeling		
171	**Hoe zijn menopauze symptomen te behandelen?**	**0**	**0,00**
172	**Is obstipatie bij darmkanker te behandelen?**	**0**	**0,00**
173	**Hoe valt de oververmoeidheid bij Pfeiffer te behandelen?**	**0**	**0,00**
174	**Wanneer is het verstandig om geen morfine te gebruiken bij de bestrijding van pijn?**	**0**	**0,00**
175	**Hoe dien je een kneuzing te behandelen?**	**0**	**0,00**
	T. Specialist – Diagnose		
176	**Waar kan ik voor een echo terecht?**	**0**	**0,00**
177	**Wie kan vaststellen dat ik leukemie heb?**	**0**	**0,00**
178	**Wie stelt reuma vast?**	**0**	**0,00**
179	**Mag een dokter een zelfdiagnose doen?**	**0**	**0,00**
180	**Hoe betrouwbaar is de diagnose van een specialist?**	**0**	**0,00**
	U. Specialist – Behandeling		
181	**Welke specialist behandelt acne?**	**0**	**0,00**
182	**Kan de kaakchirurg ook kiezen trekken?**	**0**	**0,00**
183	**Wie behandelt ziektes aan de ogen?**	**0**	**0,00**
184	**Wie maakt tijdens de behandeling gebruik van radioactieve stoffen?**	**0**	**0,00**
185	**Door wie wordt een open hartoperatie uitgevoerd?**	**0**	**0,00**