



Koninklijke Bibliotheek

Eindverslag Onderzoekstraject Tekstontsluiting

Auteur	Marian Hellema
Datum	25 januari 2008
Koninklijke Bibliotheek, afdeling RDD/HRD	

Samenvatting

Dit rapport beschrijft de resultaten van het Onderzoekstraject Tekstontsluiting, opgezet door de Koninklijke Bibliotheek (KB). Bij de grootschalige tekstdigitaliseringprojecten van de KB, zoals Staten-Generaal Digitaal (SGD), Databank Digitale Dagbladen (DDD) en Digitalisering Bijzondere Collecties (DBC), is de tekstontsluiting niet optimaal. In de huidige projecten zijn *fulltext* zoekmogelijkheden beschikbaar, gecombineerd met een basisset aan metadata.

Hoewel dit een goede basis vormt, bestaan er meer geavanceerde technieken waarmee de zoekmogelijkheden kunnen worden uitgebreid. Daarnaast speelt een rol dat het veelal om historische teksten gaat. Bij oudere teksten wordt de doorzoekbaarheid belemmerd door spellingsvarianten en doordat de tekenherkenning slechter is dan bij moderne teksten. In dit onderzoekstraject is daarom onderzocht hoe de ontsluiting van historisch tekstmateriaal verbeterd kan worden.

Allereerst is in 2006 begonnen met een inventarisatie van de mogelijkheden van *text retrieval* en *text mining*¹. Vervolgens zijn in 2007 drie pilot-projecten gestart met enkele veelbelovende technieken, met als doel de bruikbaarheid te onderzoeken voor historisch tekstmateriaal uit de KB-digitaliseringsprojecten.

De drie pilot-projecten zijn:

1. Herkennen van spellingsvariatie en OCR-fouten
 - door Martin Reynaert van de onderzoeksgroep ILK, Universiteit Tilburg.
 - en een demonstratie van de mogelijkheden door het INL (Instituut voor Nederlandse Lexicologie, Leiden).
2. Automatische classificatie van teksten aan de hand van thesauri
 - door Irion Technologies, Delft.
3. Automatisch genereren van samenvattingen
 - door Carp Technologies uit Enschede onder eindverantwoordelijkheid van Irion Technologies.

Tijdens dit traject is IMPACT gestart, een Europese project met als belangrijkste doelstelling het verbeteren van de OCR-technologie voor historische teksten. Voor een deel overlappen de werkzaamheden van IMPACT met die van de pilots. Daarom is het beoogde resultaat van IMPACT bij de evaluatie van de resultaten in dit verslag betrokken.

De belangrijkste conclusies van het Onderzoekstraject Tekstontsluiting zijn:

- a) herkenning van spellingsvariatie en OCR-fouten levert op korte termijn de meeste winst op voor het beter doorzoekbaar maken van het gedigitaliseerde tekstmateriaal. De in de pilot ontwikkelde aanpak biedt een uitstekende basis om de vindbaarheid van woorden sterk te verbeteren. Deze techniek is in principe bruikbaar voor alle collecties (taal- en periodeonafhankelijk). Inmiddels is een vervolgtraject gestart, waarin deze techniek binnen de technische infrastructuur van de KB wordt geïmplementeerd.
- b) automatische classificatie van teksten aan de hand van thesauri dient per project op bruikbaarheid te worden beoordeeld. Daarbij moet de meerwaarde van automatische classificatie worden afgewogen tegen de verwachte inspanning.

¹ *Text retrieval* en *text mining* zijn verzamelbegrippen voor technieken waarmee uit ongestructureerde teksten nieuwe informatie kan worden geëxtraheerd. Zie hoofdstuk 2 van dit verslag voor een nadere beschrijving.

- c) software die automatische samenvattingen genereert is mogelijk te implementeren als een *service*. Ieder project kan er dan desgewenst gebruik van maken. De KB zal deze mogelijkheid verder onderzoeken en mogelijk in een vervolgtraject implementeren.
- d) het besluit over het toepassen van tekstontsluitingstechnieken dient zoveel mogelijk na raadpleging van gebruikerspanels te worden genomen. Dit geldt in ieder geval voor de mogelijkheden van samenvattingen en van classificatie.
- e) het IMPACT-project zal naar verwachting na twee jaar sterk verbeterde OCR-technologie opleveren. Daarmee zullen veel van de huidige problemen met de tekenherkenning in historisch tekstmateriaal tot het verleden behoren.
- f) op langere termijn zullen meer geavanceerde technieken worden overwogen om de doorzoekbaarheid van KB-tekstmateriaal verder te verbeteren. De KB zal de ontwikkelingen op dit gebied actief blijven volgen.

Inhoudsopgave

Samenvatting	2
1 Opzet Onderzoekstraject	5
1.1 Aanleiding	5
1.2 Aanpak.....	5
1.3 Verslag resultaten	6
2 Overzicht van technieken voor tekstontsluiting	7
2.1 Named entity recognition	7
2.2 Herkennen van historische spellingvarianten.....	8
2.3 Opsporen van OCR-fouten	8
2.4 Automatische classificatie	8
2.5 Automatische clustering	9
2.6 Automatisch genereren van samenvattingen.....	9
2.7 Information retrieval en text mining	10
2.8 Taalwetenschappelijke analyses.....	10
2.9 Question answering	11
2.10 Tussentijdse conclusie van de inventarisatie.....	11
3 Pilots en hun resultaten.....	12
3.1 Pilot spellingcorrectie en spellingvariatie	12
3.2 Pilot automatische classificatie	13
3.2.1 Resultaten Staten-Generaal Digitaal	13
3.2.2 Resultaten kranten	14
3.2.3 Algemeen.....	15
3.3 Pilot automatische samenvattingen	15
3.4 Resultaten	16
3.5 Demonstratie INL.....	16
3.6 IMPACT-project	17
4 Conclusies	18
4.1 Algemeen.....	18
4.2 Spellingvariatie.....	18
4.3 Classificatie	19
4.4 Samenvattingen	21
4.5 Named entity recognition	21

1 Opzet Onderzoekstraject

1.1 Aanleiding

De KB voert een aantal digitaliseringsprojecten uit waarin grote hoeveelheden tekst worden gedigitaliseerd en *fulltext* beschikbaar worden gemaakt. Het gaat in deze projecten grotendeels om historisch tekstmateriaal en om grote aantallen pagina's:

- Staten-Generaal Digitaal (SGD) digitaliseert parlementaire stukken uit de periode 1814-1995 (2 ½ miljoen pagina's).
- Databank Digitale Dagbladen (DDD) digitaliseert historische kranten uit de 17^e – 20^e eeuw (8 miljoen pagina's).
- Digitalisering Bijzondere Collecties digitaliseert ongeveer 5.000 boeken uit het einde van de 18^e eeuw (1,3 miljoen pagina's).
- Digitalisering nieuwsberichten ANP digitaliseert ANP-nieuwsberichten uit de periode 1937-1985 (1 ½ miljoen pagina's).

Naar verwachting zullen in de toekomst nog meer grootschalige tekstdigitaliseringsprojecten volgen.

Het doorzoekbaar maken van de teksten gaat momenteel op twee manieren: er worden metagegevens toegevoegd en de teksten worden in hun geheel (op woord) doorzoekbaar gemaakt door het toepassen van tekenherkenning (OCR, *Optical Character Recognition*). Bij grootschalige digitaliseringsprojecten is over het algemeen slechts mogelijk een beperkte hoeveelheid metagegevens toe te voegen.

Deze ontsluiting (beperkte metadata in combinatie met *fulltext* zoekmogelijkheden) biedt een goede basis om de teksten te doorzoeken, maar is niet optimaal. Het zoeken in deze omvangrijke collecties zal vaak een groot aantal zoekresultaten opleveren, waardoor een eindgebruiker niet (makkelijk) de gewenste informatie vindt. Daarnaast wordt de doorzoekbaarheid bij oudere teksten bemoeilijkt door de historische spellingvarianten. Bovendien is de kwaliteit van de tekenherkenning bij oudere teksten minder goed dan bij moderne teksten, wat de doorzoekbaarheid eveneens belemmert.

De vraag die aan dit Onderzoekstraject ten grondslag ligt, is in hoeverre de ontsluiting van de teksten verbeterd kan worden met geavanceerde technieken, zoals *text retrieval* of *text mining*. Vanwege de grootschaligheid van de digitaliseringsprojecten moeten de technieken zoveel mogelijk geautomatiseerd kunnen worden; veel handmatig werk is niet haalbaar.

1.2 Aanpak

Het eerste deel van het Onderzoekstraject bestond uit een reeks inventariserende gesprekken met bedrijven, instellingen en onderzoeksgroepen die zich bezighouden met tekstontsluiting. Het doel van deze inventarisatie was zicht te krijgen op de bestaande technieken voor tekstontsluiting.

De inventarisatie heeft geleid tot een overzicht van verschillende technieken. De belangrijkste conclusie was dat het zinvol was een aantal technieken in pilot-projecten verder te onderzoeken. De doelstelling van deze pilots was te onderzoeken in hoeverre de technieken bruikbaar zijn voor teksten uit de KB-digitaliseringsprojecten.

De drie pilot-projecten zijn:

1. Herkennen van spellingsvariatie en OCR-fouten
 - door Martin Reynaert van de onderzoeksgroep ILK, Universiteit Tilburg.
 - en een demonstratie van de mogelijkheden door het INL (Instituut voor Nederlandse Lexicologie, Leiden).
2. Automatische classificatie van teksten aan de hand van thesauri

- door Irion Technologies, Delft (<http://www.irion.nl/>).
- 3. Automatisch genereren van samenvattingen
 - door Carp Technologies uit Enschede onder eindverantwoordelijkheid van Irion Technologies (<http://carp-technologies.nl/>).

De tekstbestanden die gebruikt zijn voor de pilots komen uit de projecten Staten-Generaal Digitaal en Historische Kranten in Beeld (een pilotproject dat vooraf is gegaan aan Databank Digitale Dagbladen).

Tijdens dit traject is tevens IMPACT gestart, een project waarin 15 Europese partners samenwerken, met als één van de hoofddoelstellingen het verbeteren van de OCR-technologie voor historische teksten. De KB leidt het IMPACT-project en het INL is een deelnemende partner. Omdat er raakvlakken zijn met de pilots tekstontsluiting, zijn de beoogde werkzaamheden van IMPACT bij de evaluatie van de resultaten van dit Onderzoekstraject betrokken.

1.3 Verslag resultaten

Dit verslag bevat de resultaten van het onderzoekstraject. Hoofdstuk 2 geeft een overzicht van de bestaande technieken van tekstontsluiting, die in de inventariserende gesprekken naar voren zijn gekomen. Hoofdstuk 3 beschrijft de pilot-projecten en gaat in op de werkzaamheden van het INL en van het IMPACT-project. In hoofdstuk 4 worden conclusies getrokken.

2 Overzicht van technieken voor tekstontsluiting

Uit de inventariserende gesprekken is een aantal technieken naar voren gekomen die interessant kunnen zijn voor tekstontsluiting van KB-materiaal. In dit hoofdstuk worden deze technieken kort beschreven.

Veelgebruikte begrippen en definities:

- *information extraction*: het onttrekken van informatieve gegevens aan tekstmateriaal. Hieronder valt onder meer *named entity recognition*, d.w.z. het geautomatiseerd herkennen van persoonsnamen, plaatsnamen, organisatienamen, datum-waarden e.d.
- *information retrieval*: het zoeken en vinden van informatie in tekstmateriaal door het doorzoeken van de tekst of metagegevens. Indexeertechnieken zijn hierbij van groot belang, evenals de ranking van zoekresultaten.
- *text retrieval*: dit is hetzelfde als information retrieval, toegepast op ongestructureerd tekstmateriaal (i.t.t. gestructureerde databases).
- *text mining*: het vinden van tevoren onbekende informatie of kennis in teksten. Hierbij worden *data mining* technieken (geavanceerde statistische technieken) toegepast op tekstueel materiaal. Het gaat hierbij om het ontdekken van relaties tussen gegevens.
- achterliggende technieken die worden toegepast om bovenstaande doelen (information extraction, information retrieval, text retrieval, text mining) te bereiken:
 - Natural Language Processing (NLP): technieken die gebaseerd zijn op kennis van taal (woordenlijsten, grammatica, semantische kennis).
 - statistische technieken (bv. woordfrequenties).

Bovenstaande begrippen zijn overkoepelende begrippen. De KB heeft zich in dit onderzoekstraject vooral gericht op een aantal concrete tekstontsluitingstechnieken., die hieronder worden toegelicht.

2.1 *Named entity recognition*

Bij Named Entity Recognition (NER) worden eigennamen, zoals persoonsnamen, plaatsnamen en namen van organisaties automatisch herkend. Hiervoor kunnen lijsten met bekende "entiteiten" worden gebruikt, zoals lijsten met persoonsnamen. Andere technieken herkennen entiteiten op basis van de context, door de betekenis van de woorden er omheen. Zo kunnen ook entiteiten die niet van tevoren bekend zijn, worden herkend. Bij NER is tevens van belang dat verschillende entiteiten met identieke namen van elkaar onderscheiden worden, bijvoorbeeld vader en zoon George Bush.

NER maakt het mogelijk "entiteiten" te markeren en extra zoekmogelijkheden aan te bieden, bijvoorbeeld een persoonsnaam, die kan doorlinken naar biografische informatie. Herkenning van entiteiten biedt tevens de mogelijkheid het materiaal op nieuwe manieren doorzoekbaar te maken, bijvoorbeeld door het aanbieden van kaarten waarmee de geografische aanduidingen in de tekst doorzocht kunnen worden.

Een project van de KB dat gebruikt maakt van deze techniek is Staten-Generaal Digitaal. Op basis van een lijst met parlementsleden worden alle parlementariërs in de teksten herkend. Dit zal worden gebruikt om biografische informatie aan te bieden en om gemakkelijk te kunnen doorzoeken naar andere plaatsen in de tekst waar de persoon voorkomt. In dit Onderzoekstraject is niet met NER geëxperimenteerd. In het IMPACT-project worden NER-mogelijkheden verder ontwikkeld, die ook voor de KB interessant zijn.

2.2 Herkennen van historische spellingvarianten

Nederlandse historische teksten kennen een groot aantal spellingvarianten..Vóór 1883 hanteerde Nederland zelfs geen officiële spelling en na die datum zijn meerdere malen spellingswijzigingen doorgevoerd. Deze spellingvariatie bemoeilijkt het zoeken van woorden in de tekst. Alle KB-tekstdigitaliseringsprojecten hebben met dit probleem te maken.

Door **herkenning van historische spellingvarianten** kan een eindgebruiker een woord in moderne spelling zoeken (bijvoorbeeld “mens”) en dan plaatsen vinden waar het woord in een oudere spellingvariant staat (“mensch”). Dit verbetert de doorzoekbaarheid van historische teksten aanzienlijk. Eén pilot gaat daarom in op deze techniek, zie hoofdstuk 3, paragraaf 1.

2.3 Opsporen van OCR-fouten

In de meeste digitaliseringsprojecten van de KB worden teksten doorzoekbaar gemaakt met behulp van tekenherkenning (OCR, Optical Character Recognition). De OCR-software is over het algemeen goed in staat de tekens in moderne teksten te herkennen. In het project Staten-Generaal Digitaal wordt in de hedendaagse documenten minstens 99,8% van alle tekens correct herkend. Voor oudere teksten zijn de prestaties van OCR een stuk minder. Voor het pilotproject ‘Historische Kranten in Beeld’ bleek dat slechts zo’n 60-70% van de woorden geheel correct waren herkend. Deels werd dit veroorzaakt doordat de images in dit project van mindere kwaliteit waren, wat de kwaliteit van de OCR nadelig beïnvloedt. De belangrijkste factor is echter dat de OCR-technologie tekortschiet voor historische teksten. Bij ouder tekstmateriaal spelen specifieke problemen, zoals verschillen in lettertype en pagina-indeling, kwaliteit van het papier, beschadigingen, doordrukken van inkt van de achterkant van de pagina e.d. Naar verwachting krijgen alle digitale historische tekstcollecties te maken met veel OCR-fouten.

Bij het opsporen van OCR-fouten wordt geprobeerd bij de correcte woordvariantenvariant alle varianten met OCR-fouten te vinden. De eindgebruiker krijgt hierdoor zoekresultaten die zowel de correcte zoekterm bevat als de varianten van deze term met OCR-fout(en). Een stap verder is de mogelijkheid om de tekst op te schonen, dat wil zeggen de OCR-fouten daadwerkelijk te verbeteren.

Deze techniek levert een belangrijke verbetering van de doorzoekbaarheid van tekstcollecties op. Het kan bovendien bijdragen aan de oplossing van het probleem dat de slechte kwaliteit van de OCR andere tekstontsluitingstechnieken belemmert, zoals automatische classificatie, Named Entity Recognition of het genereren van samenvattingen. Eén van de pilots is daarom gewijd aan deze techniek, zie hoofdstuk 3, paragraaf 1.

Het IMPACT-project besteedt veel aandacht aan het verbeteren van de OCR-technologie voor historische teksten, waardoor in de toekomst betere OCR-resultaten behaald zullen worden. Het achteraf opsporen van OCR-fouten blijft naar verwachting een goede aanvulling, aangezien de OCR-resultaten nooit 100% foutvrij zullen worden.

2.4 Automatische classificatie

Bij automatische classificatie wordt tekst ingedeeld in bepaalde, van te voren vastgestelde, klassen (categorieën). Meestal zijn de klassen ontleend aan thesauri of gecontroleerde vocabulaires en vormen zij een inhoudelijke, onderwerpsgerichte indeling. Krantenberichten kunnen bijvoorbeeld op onderwerp worden geclassificeerd (politiek, sport, cultuur e.d.), maar ook op soort krantenbericht (nieuwsbericht, familiebericht, advertentie e.d.).

Classificatie kan op verschillende manieren worden toegepast om de ontsluiting van teksten te verbeteren:

- de klassen kunnen in de zoekinterface aan de gebruiker worden aangeboden als zoekcriterium of als mogelijkheid om de zoekactie te verfijnen (bijvoorbeeld: alleen in de krantenberichten over cultuur zoeken).
- de klassen kunnen worden gebruikt om gerelateerde documenten aan te bieden (bijvoorbeeld “zoek meer krantenberichten over dit onderwerp”).
- de klassen kunnen worden gebruikt om de *ranking* van zoekresultaten te verbeteren. Als een eindgebruiker bijvoorbeeld als zoekterm “onderwijs” opgeeft, dan worden in de zoekresultaten eerst de teksten getoond die in de klasse onderwijs geïnclassificeerd zijn en dan pas de teksten waar het woord alleen maar in de tekst voorkomt.

Met automatische classificatie is in één van de pilots geëxperimenteerd, zie hoofdstuk 3, paragraaf 2.

2.5 Automatische clustering

Automatische clustering groepeert gelijksoortige teksten. De clusters zijn gebaseerd op de teksten zelf en zijn dus niet afkomstig uit een van tevoren vastgestelde thesaurus, classificatie of andere indeling. Als er nieuwe teksten komen die niet in een bestaand cluster passen, worden nieuwe clusters gevormd. Een voorbeeld van een toepassing is het toegankelijk maken van nieuwsberichten, waarbij de berichten automatisch geclusterd worden, waardoor de eindgebruiker de berichtgeving over een bepaald nieuwsfeit kan volgen. Als er nieuwe onderwerpen in het nieuws komen, worden die herkend als afzonderlijk cluster. Deze techniek wordt ook wel **topic detection** genoemd.

In dit onderzoekstraject is niet met automatische clustering geëxperimenteerd. Fundamentele technieken krijgen in dit stadium de voorrang (met name de spellingvariatie, de OCR-problematiek en de automatische classificatie). Automatische clustering wordt wel als interessante techniek gezien voor KB-tekstcollecties.

2.6 Automatisch genereren van samenvattingen

Automatisch samenvatten van teksten kan op verschillende manieren de doorzoekbaarheid van collecties vergroten:

- de samenvatting van een document geeft de gebruiker snel inzicht in de inhoud van de tekst. Er is onderscheid tussen leesvervangende samenvattingen (waarbij de samenvatting de oorspronkelijke tekst vervangt) en indicatieve samenvattingen (waarbij de samenvatting een indicatie geeft van de inhoud van de tekst, maar waarbij de oorspronkelijke tekst eveneens zichtbaar blijft). Voor KB-tekstcollecties lijken vooral indicatieve samenvattingen interessant, als aanvulling op het tonen van de volledige tekst.
- een korte samenvatting bij het zoekresultaat helpt de gebruiker te bepalen of dit zoekresultaat interessant is. In dergelijke samenvattingen worden liefst de zoekwoorden betrokken die de gebruiker had opgegeven (“snippets”).
- samenvattingen kunnen de classificaties van teksten verbeteren. Het achterliggende principe is dat classificatie beter werkt als de tekst is teruggebracht tot zijn essentie.
- samenvattingen kunnen de *ranking* van zoekresultaten verbeteren. Zoektermen die in de samenvatting voorkomen krijgen een hogere *ranking* dan overige termen.

De eerste mogelijkheid is in één van de pilots onderzocht, zie hoofdstuk 3, paragraaf 3.

2.7 Information retrieval en text mining

Information Retrieval en text mining zijn begrippen die voor verschillende technieken worden gebruikt. Het gaat in het algemeen om het ontdekken van informatie in teksten. Dat wil zeggen dat er niet alleen losse woorden of begrippen in een tekst worden onderscheiden, maar ook verbanden daartussen.

Er zijn systemen ontwikkeld, vooral voor medische teksten, die kunnen omgaan met vragen als "wat zijn de oorzaken van longkanker". Middels taalkundige technieken (zoals syntactische en/of semantische analyse) wordt de informatie uit de teksten zodanig verwerkt dat relevante informatie op inhoudelijke gronden gevonden wordt (dus niet alleen een *fulltext search* naar de woorden "longkanker" en "oorzaak"). Onderzoekers stellen dat de achterliggende principes voor elk soort tekst bruikbaar zijn. Toch is het de vraag of dergelijke technieken goed werken bij de tekstcollecties van de KB. Dit zijn immers geen teksten uit de exacte wetenschappen, maar uit vakgebieden waarin het taalgebruik meer ambigue is, waardoor information retrieval wellicht moeilijker wordt.

Een verwant concept is "Topic Maps". Hiermee worden allerlei soorten informatie uit de teksten aan elkaar gekoppeld, waardoor de gebruiker de informatie op associatieve wijze kan exploreren.

Een ontwikkeling in het vakgebied van de Information Retrieval is het personaliseren van zoeksystemen. De achterliggende gedachte bij personalisatie is dat iedere gebruiker(sgroep) eigen behoeften heeft en eigen betekenissen aan informatie geeft. Informatiesystemen zouden hier rekening mee moeten houden door zich aan te passen aan de behoeften van verschillende gebruikers. Een goed voorbeeld is de toepassing bij de expositie van Beeld en Geluid (Hilversum), waar iedere bezoeker een ring krijgt, die onder meer de geboortedatum bevat. Op basis daarvan wordt het aanbod op de individuele bezoeker toegespitst, die zo bijvoorbeeld de kinderprogramma's uit de eigen jeugd te zien krijgt.

Een andere ontwikkeling in de Information Retrieval is het dialoogsysteem. Dergelijke systemen gaan een dialoog met de gebruiker aan om zo snel mogelijk tot een zo adequaat mogelijk zoekresultaat te komen. Als een gebruiker bijvoorbeeld een ambigu zoekwoord als "jaguar" opgeeft (dier, auto of programmeertaal?), kan er een "wolk" van geassocieerde begrippen worden gepresenteerd, waarmee de gebruiker snel het betekenisdomein kan kiezen dat hem of haar op dat moment interesseert.

In dit Onderzoekstraject Tekstontsluiting is niet met information retrieval en text mining geëxperimenteerd. Het aanpakken van de basisproblemen gaat voor de meer geavanceerde toepassingen als information retrieval en text mining.

2.8 Taalwetenschappelijke analyses

Taalwetenschappers hebben specialistische onderzoeksvragen en tools om teksten te analyseren. Zij zijn onder meer geïnteresseerd in:

- frequenties van woordgebruik (bijvoorbeeld in een bepaalde periode of een bepaald corpus, in vergelijking met een referentiecorpus);
- *part of speech tagging* (het onderscheiden van de grammaticale elementen in de tekst);
- concordantie en *collocation* (de context van woorden);
- semantische analyse;
- *cognates* (woorden die aan elkaar verwant zijn).

Taalwetenschappers zijn een belangrijke doelgroep voor de KB-digitaliseringsprojecten, bijvoorbeeld bij het project Databank Digitale Dagbladen. Hun onderzoeksvragen zijn echter dermate divers en specialistisch dat het niet goed mogelijk is de analysetools die hiervoor

nodig zijn op een algemene website aan te bieden. Taalwetenschappers hebben vooral baat bij het kunnen beschikken over het hele corpus (ruwe tekst), om hiermee met hun eigen analysetools hun onderzoek te kunnen verrichten.

2.9 Question answering

Question Answering houdt in dat eindgebruikers hun zoekvraag in natuurlijke taal kunnen stellen en beantwoord krijgen (bijvoorbeeld “waar is de dichtstbijzijnde pizzeria?”). Het antwoord hoeft niet altijd letterlijk in de tekst te staan, maar kan door middel van taaltechnologie uit de tekst worden afgeleid. Soms zijn er meerdere antwoorden mogelijk en moet het systeem bepalen welk antwoord het meest relevant of het meest correct is. Deze technologie is in dit Onderzoekstraject niet verder betrokken.

2.10 Tussentijdse conclusie van de inventarisatie

De inventarisatie heeft inzicht opgeleverd in technieken waarmee de ontsluiting van de historische digitale tekstcollecties van de KB verbeterd kan worden. De conclusie is dat met het oplossen van een aantal basisproblemen al flinke winst in de doorzoekbaarheid van teksten te behalen valt. Het is niet haalbaar om alle interessante technieken tegelijk nader te onderzoeken en toe te passen.

Dit leidde tot het opzetten van drie pilotstudies: een pilot omtrent de problematiek van spellingvariatie; een pilot waarin automatische classificatie wordt onderzocht; en een pilot waarin automatische samenvattingen worden bekeken.

De overige technieken kunnen voor de KB eveneens interessant zijn. Na de pilots en het eventueel implementeren van de technieken, zouden in de toekomst meer geavanceerde technieken kunnen worden onderzocht en ingezet.

3 Pilots en hun resultaten

Uit de inventarisatie blijkt dat veel technieken met succes worden toegepast. Het tekstmateriaal van de digitaliseringsprojecten van de KB heeft echter een paar bijzonderheden die de toepassing kunnen bemoeilijken:

1. de kwaliteit van de OCR voor oudere teksten is minder goed dan voor moderne teksten. Daardoor bevatten de teksten meer foutieve woordvormen, wat de toepassing van de technieken kan bemoeilijken.
2. het woordgebruik in historische teksten is anders dan in moderne teksten. Er is sprake van oude spellingsvormen en van betekenisverschillen in het woordgebruik. Ook dit kan de toepassing van technieken bemoeilijken.
3. de KB projecten zijn grootschalig. Veel van de technieken zijn uitgetoetst op kleinere collecties, waarbij handmatige bewerking mogelijk is. Bij massadigitalisering dienen de technieken geautomatiseerd te kunnen worden toegepast.

De pilots hebben daarom het doel te onderzoeken in hoeverre ze voor het specifieke materiaal van de KB goede resultaten opleveren. De teksten die hiervoor gebruikt zijn komen uit het project Staten-Generaal Digitaal en uit het pilotproject Historische Kranten in Beeld.

In dit hoofdstuk worden de bevindingen beschreven van:

1. de pilot over spellingvariatie (historische spellingvarianten en OCR-fouten) door Martin Reynaert, ILK, Tilburg.
2. de pilot over automatische classificatie van teksten door Irion.
3. de pilot over het automatisch genereren van samenvattingen door Carp/Irion.
4. de demonstratie door het INL.

Ook wordt kort ingegaan op de doelstellingen van het IMPACT-project.

3.1 Pilot spellingcorrectie en spellingvariatie

Martin Reynaert (ILK, Tilburg) heeft een algoritme ontwikkeld dat een lijst met woordvarianten genereert bij alle woorden in de tekst. In dit algoritme wordt geen onderscheid gemaakt tussen woordvariatie die is ontstaan door OCR-fouten of door historische spelling. Dit algoritme verbetert de *retrieval*: een eindgebruiker vindt bij iedere zoekterm ook de woorden met OCR-fouten en de historische spellingvarianten.

Het algoritme geeft goede resultaten, vooral bij spellingvarianties die een *edit distance* hebben van één of twee². De resultaten worden minder als de kwaliteit van de OCR afneemt, maar zelfs bij teksten met de slechtste OCR-kwaliteit wordt toch een deel van de variatie gevonden.

De behaalde presentaties van het algoritme:

- bij de teksten uit Staten-Generaal Digitaal wordt 89% van alle ongewenste variatie gevonden.
 - 90% van de woordvariatie heeft een *edit distance* van 1 of 2. Voor de fouten met deze edit distances is de **recall** 99,5%: van de 200 varianten die in de tekst voorkomen wordt slechts één variant gemist. De **precisie** bij deze *edit distances* bedraagt 85%: op 100 varianten worden 15 foutieve varianten gegenereerd.
- bij de teksten uit kranten wordt bijna 55% van alle ongewenste variatie gevonden.

² De edit distance (ook wel Levenshtein distance genoemd) wil zeggen het aantal "edit-acties" dat nodig is om de ene woordvariant in de andere te veranderen.

- 55% van de woordvariatie heeft een *edit distance* van 1 of 2. Voor de fouten met deze edit distances is de **recall** 99%: van de 100 varianten die in de tekst voorkomen wordt slechts één variant gemist. De precisie bij deze *edit distances* bedraagt 93%: op 100 varianten worden 7 foutieve varianten gegenereerd.

De hoeveelheid variatie die ontstaat door OCR-fouten bleek vele malen groter dan die door historische spellingvariatie ontstaat. Ook bleek dat OCR-fouten niet erg "systematisch" zijn: bij identieke woorden of lettercombinaties maakt de OCR-software vaak verschillende fouten.

Een belangrijke eigenschap van het algoritme is dat het taalafhankelijk is. In principe kan het zonder aanpassingen op corpora in verschillende talen en uit verschillende perioden worden toegepast. Verbetering van het algoritme wordt bereikt door het uit te breiden met woordenlijsten, (bijvoorbeeld historische lexica, specifiek voor een bepaalde tijd en periode) en door de context van een woord te betrekken in het proces. Ook zijn er nog verbeteringen mogelijk voor hogere *edit distances* dan 2.

Een ander belangrijk gegeven is dat het algoritme geautomatiseerd werkt. Behalve eventuele voorbewerkingen van de teksten (bijvoorbeeld om afbreekstreepjes uit de tekst te verwijderen) is geen handmatig ingrijpen nodig. Wel zal het verwerken van grote corpora veel rekenkracht vergen. Martin Reynaert verwacht dat het algoritme nog efficiënter kan worden gemaakt om goed schaalbaar te zijn.

3.2 *Pilot automatische classificatie*

Irion Technologies uit Delft ontwikkelt onder meer automatische classificatie-software. Deze software moet eerst getraind worden met een testset waarvan de juiste klassen bekend zijn (of "met de hand" worden opgegeven). De software leert te onderscheiden wat de specifieke kenmerken zijn van teksten uit de verschillende klassen. Op basis daarvan kan de software na de trainingsfase nieuwe teksten automatisch classificeren.

Voor de pilot zijn twee collecties en thesauri gebruikt:

1. Staten-Generaal Digitaal, met de Parlementsthesaurus, ontwikkeld en onderhouden door de Tweede Kamer en gebruikt voor handmatige classificatie van parlementaire teksten sinds 1978.
2. Historische Kranten in Beeld, met teksten uit 1918. Voor dit materiaal is gebruik gemaakt van de IPTC³, een internationaal veel gebruikte thesaurus voor nieuwsberichten.

Omdat de software bij het classificeren last kan hebben van spellingvariatie, heeft Irion aandacht besteed aan de problematiek van de historische spellingvariatie en de OCR-fouten. Hiervoor is een *Language Expansion Box* ontwikkeld, waarmee vooraf de teksten zo veel mogelijk worden genormaliseerd naar hedendaags Nederlands zonder spellingsfouten. Voor beide collecties uit de pilot is een afzonderlijke *Language Expansion Box* ontwikkeld.

3.2.1 Resultaten Staten-Generaal Digitaal

De Parlementsthesaurus was niet voor alle termen te gebruiken. De thesaurus bevat namelijk klassen die te veel op elkaar lijken en klassen die eigenlijk niet in een onderwerpsclassificatie thuishoren, zoals geografische klassen. Tevens bleken de trainingsdocumenten niet evenwichtig over de klassen verdeeld te zijn.

³ zie <http://www.iptc.org/pages/index.php>

Uit praktische overwegingen is besloten alleen de klassen van het hoogste niveau uit de hiërarchie van de thesaurus in de pilot te betrekken. Voor de oudere teksten is het ook om inhoudelijke redenen niet zinvol om de thesaurus tot op het diepste niveau in te zetten. De termen op het diepste niveau zijn vaak zo specifiek, dat ze voor oudere teksten niet meer van toepassing zijn (bijvoorbeeld "informatie- en communicatietechnologie" of "NGO's"). De termen op het hoogste niveau zijn algemeen genoeg dat ze ook voor oudere teksten zinvol lijken, zoals "economie", "recht", "defensie" of "cultuur". In totaal zijn er op dit niveau 29 klassen in de Parlementsthesaurus.

Irion heeft haar automatische classificatie-software getraind met trainingsdocumenten uit de periode 1980-2005, die door de deskundigen van de Tweede Kamer zijn geïnclassificeerd. Vervolgens is de getrainde classificatiesoftware gebruikt voor de automatische classificatie van documenten uit de periode 1929-1934.

Uit de evaluatie van de resultaten blijkt:

- een gemiddelde recall van 73% en een gemiddelde precisie van 61%. De KB heeft een evaluatie gedaan, uitgesplitst per soort document. Dit leverde zeer uiteenlopende prestaties op: een recall van 35-90% en een precisie van 30-85%. Deze grote variatie hangt samen met het type documenten: voor sommige documenttypen werkt de classificatie uitstekend, terwijl het voor andere type documenten niet goed werkt.
- bepaalde documenten zijn niet te classificeren, zoals opsommingen van binnengekomen stukken in een vergaderverslag.
- de Language Expansion Box leidt bij teksten uit Staten-Generaal Digitaal niet tot betere prestaties van de classificatie.
- de automatische classificatie-software moet worden bijgetraind, omdat twee klassen vaak ten onrechte worden toegekend, vanwege de onevenwichtigheid in de trainingsdocumenten.

3.2.2 Resultaten kranten

Irion beschikte al over getrainde classificatiesoftware voor de IPTC. Deze is gebruikt voor de automatische classificatie van krantenartikelen uit 1918 (uit Historische Kranten in Beeld).

Uit de geïnclassificeerde teksten blijkt:

- de kwaliteit van de classificatie wordt sterk bepaald door de kwaliteit van de OCR. De artikelen met de minste OCR-kwaliteit gaven voor de classificatie ook de minst goede resultaten.
- in kranten staan regelmatig teksten die niet te classificeren zijn met de begrippen uit de IPTC, bijvoorbeeld puzzels of kalenders.
- hoge recall- en precisiecijfers van 75-100% zijn haalbaar, maar aan een belangrijk deel van de documenten wordt dan geen klasse toegekend (een lage *coverage* van 20%). Breng je de *coverage* omhoog (bijvoorbeeld tot 60-85%) dan is een lagere *recall* en *precisie* (van ongeveer 35-65%) het gevolg. Toelichting: er kan een drempel worden ingesteld voor de mate van zekerheid voor de toegekende classificatie. Hoe hoger de drempel hoe kleiner het aantal documenten waaraan (met zekerheid) een klasse kan worden toegekend. Hoe lager de drempel hoe groter het aantal geïnclassificeerde documenten, maar in de toegekende klassen komen dan meer fouten voor.

- uit evaluatie door de KB blijken bij een *coverage* van bijna 100% een *recall* van 55%-65% en een *precisie* van 40-46% gehaald te worden. In deze evaluatie zijn alleen "echte" nieuwsberichten opgenomen.
- de Language Expansion Box levert bij het krantenmateriaal een verbetering van de classificatie op van ongeveer 20%.

Ook bij dit deel van de pilot rees de vraag in hoeverre de moderne thesaurus bruikbaar is voor het classificeren van oudere teksten. De begrippen uit de moderne thesaurus bleken niet altijd goed te passen bij de teksten uit het begin van de twintigste eeuw. Een voorbeeld hiervan was een bericht uit 1929 waarin melding werd gemaakt van een vliegtuig dat in Bagdad was opgestegen en in Nederlands Indië was geland. Dit werd geclassificeerd als "luchtvaartongeval", terwijl er van een ongeluk geen sprake was. In 1929 had een geslaagde lange-afstandsvlucht nieuwswaarde, terwijl er tegenwoordig wel iets aan de hand zal zijn als een vlucht in het nieuws komt.

3.2.3 Algemeen

Onderzoek wijst uit dat de recall en precisie over het algemeen minstens 80% moeten zijn wil de classificatie goed bruikbaar zijn. De resultaten in de pilot zijn minder goed dan Irion gewend is bij moderne teksten. Zowel voor Staten-Generaal Digitaal als voor de kranten ziet Irion wel mogelijkheden om het gewenste kwaliteitsniveau te behalen. De classificatiesoftware kan bijgetraind worden en de Language Expansion Box verbeterd. Voor Staten-Generaal Digitaal kan de thesaurus nog worden aangepast of kunnen alleen begrippen op het hoogste niveau worden gebruikt. Voor nog oudere teksten moet ook de Language Expansion Box aangepast worden.

Na training kan de classificatie geautomatiseerd worden toegepast op een corpus. Voor iedere nieuwe thesaurus of andere classificatie is echter een nieuwe trainingsprocedure nodig.

Voor de gedigitaliseerde KB-collecties ligt de meerwaarde van automatische classificatie vooral in de extra ontsluiting op onderwerp. Dit kan worden toegepast in combinatie met *fulltext* zoeken: als de eindgebruiker veel zoekresultaten krijgt, kan hij of zij op een bepaald onderwerp "inzoomen". Dit principe wordt "guided navigation" genoemd en is te vergelijken met de mogelijkheid tot verfijnen van de zoekactie zoals die momenteel op de website van Staten-Generaal Digitaal wordt aangeboden (www.statengeneraaldigitaal.nl). Naarmate er minder metagegevens beschikbaar zijn waarmee een collectie doorzocht kan worden heeft automatische classificatie een grotere meerwaarde.

Een andere mogelijke toepassing van classificatie voor de KB ligt in het verbeteren van de *ranking* van zoekresultaten.

3.3 Pilot automatische samenvattingen

Carp Technologies, "onderaannemer" van Irion Technologies, heeft een demonstratieapplicatie ontwikkeld waarmee willekeurige teksten kunnen worden samengevat. De applicatie maakt voor het genereren van de samenvattingen onder meer gebruik van de opbouw van een tekst. Bij moderne krantenberichten staat bijvoorbeeld de kern van een bericht meestal aan het begin van het bericht. Ook wordt gebruikt gemaakt van "signaalwoorden" in de tekst, bijvoorbeeld woorden als "concluderend".

De applicatie kan een samenvatting genereren die bestaat uit belangrijke zinnen uit de tekst, waarbij zoveel mogelijk een nog lopende tekst wordt gevormd. Het is ook mogelijk om de volledige tekst te tonen, waarin de belangrijke zinnen worden gemarkeerd (d.w.z. met een kleur *ge-highlight*). Dit heeft als voordeel dat de gebruiker snel de belangrijke delen van een

tekst ziet, maar ook de tekst eromheen kan lezen. Met opties kan worden ingesteld hoe lang de samenvatting moet zijn. Dit kan door de gebruiker ook interactief worden gedaan, met behulp van een schuifbalk waarmee de samenvatting "on the fly" langer of korter kan worden gemaakt.

3.4 Resultaten

- de teksten uit Historische Kranten in Beeld blijken van een te slechte kwaliteit om zinvol te kunnen samenvatten (d.w.z. te veel OCR-fouten). In deze pilot is daarom niet geëxperimenteerd met oudere krantenteksten, maar alleen met teksten uit het project Staten-Generaal Digitaal.
- de applicatie is geschikt voor het genereren van indicatieve samenvattingen (zie hoofdstuk 2.6). Voor het genereren van *snippets* (op basis van zoektermen) is een kleine aanpassing van de software nodig.
- de applicatie genereert samenvattingen van redelijke kwaliteit. Een informele beoordeling levert een gemiddeld rapportcijfer van 6 à 7 op.
- de kwaliteit van de samenvattingen kan worden verbeterd door de software te trainen met een aantal voorbeelddocumenten die met de hand zijn samengevat. Hiermee kan de software leren inspelen op specifieke kenmerken van de teksten in een bepaalde collectie.
- de demonstratieapplicatie is gebruiksvriendelijk. Het interactief kunnen aanpassen van de lengte van de samenvatting wordt door veel gebruikers aantrekkelijk gevonden, hoewel het de vraag is hoeveel meerwaarde dit heeft.

3.5 Demonstratie INL

Het Instituut voor Nederlandse Lexicologie (INL) beschikt over veel expertise voor taalkundige verrijking van historische teksten. Zij hebben tijdens dit Onderzoekstraject een demonstratie gegeven waarin getoond werd hoe teksten uit Historische Kranten in Beeld uit de periode 1918-1945 kunnen worden verrijkt. De technologie die zij tijdens dit Onderzoekstraject hebben gedemonstreerd, zullen zij ook in IMPACT (als deelnemende partner) inbrengen.

Taalkundige verrijking kan door:

- het opbouwen van een lexicon van het tekstmateriaal op basis van het Woordenboek der Nederlandse Taal (WNT). In dit lexicon zullen ook woordvormen worden opgenomen (lemmatisering).
- OCR-correctie met behulp van de lexicale data uit het voorgenoemde punt. Hierin wordt gekeken naar de *edit-distance* tussen woorden. Woorden worden gelemmatiseerd naar hun moderne vorm. De resultaten kunnen worden gebruikt voor *query expansie*, waarbij de zoekvraag wordt uitgebreid met woordvarianten.
- *retrieval* en koppeling met WNT. De bovengenoemde technieken maken het mogelijk om woorden in moderne spelling te zoeken en ook in oudere spelling te vinden of woorden met OCR-fouten te vinden (*retrieval*). Een andere toepassing is het *linken* van woorden aan lemma's in het Woordenboek der Nederlandse Taal. Daarmee kan aan eindgebruikers de mogelijkheid worden geboden om woorden in het WNT op te zoeken.
- *named entity recognition* op basis van lijsten met "entiteiten".

3.6 IMPACT-project

IMPACT (IMProving ACcess to Text) is een door de EU gesubsidieerd project (gestart op 01-01-2008), waarin 15 partijen samenwerken: nationale bibliotheken, universiteitsbibliotheken, onderzoeksinstituten en commerciële partijen. Het doel is het substantieel verbeteren van technieken waarmee historische tekst (boeken, kranten en ander gedrukt materiaal) op grote schaal omgezet kan worden in digitale tekst. Er wordt onder andere gewerkt aan de verbetering van de OCR-technologie en aan de verrijking van gedigitaliseerde tekst door het bouwen van computerlexica. Tevens worden *collaborative* omgevingen ontwikkeld waarin mensen ge-OCRde teksten kunnen corrigeren.

De coördinatie van het project ligt bij KB, die net als INL meewerkt aan een aantal werkpakketten. De bijdrage van INL voor IMPACT is onder andere:

- het ontwikkelen van historische lexica, in te zetten om de OCR-technologie te verbeteren.
- door deze lexicale data kunnen historische spellingvarianten en OCR-fouten herkend worden. Doel is deze informatie *tijdens* het OCR-proces in te zetten, waardoor de prestaties van de OCR-software verbeteren. Zo verbetert de *retrieval* van teksten, en kan de oorspronkelijke tekst met de historisch juiste spellingvarianten worden gecorrigeerd.
- er wordt een repository gevormd van lijsten met *named entities* die kunnen worden gebruikt voor NER. Dit kan onder meer worden gekoppeld aan *Named Authority Files*, bijvoorbeeld van de Library of Congress.

Bij alle pilots uit dit Onderzoekstraject bleek de matige kwaliteit van de OCR een struikelblok. Daarom is het resultaat van het IMPACT-project van groot belang voor de tekstontsluiting van het KB-materiaal. Als de OCR-technologie verbetert, wordt toepassing van andere tekstontsluitingstechnieken beter mogelijk.

4 Conclusies

4.1 Algemeen

1. De wensen voor tekstontsluiting verschillen per doelgroep. De KB kan en hoeft in de zoekfunctionaliteit die zij aanbiedt echter niet aan alle wensen te voldoen. Met name voor taalkundig onderzoek geldt dat de onderzoeksvragen heel specifiek zijn en specialistische analysetools vergen. Deze onderzoekers zijn er het meest bij gebaat dat zij de beschikking krijgen over het gehele corpus (ruwe tekst), waar zij hun eigen technieken en algoritmes op kunnen toepassen.
2. De slechte kwaliteit van de OCR bij historische teksten vormt een groot struikelblok voor de ontsluiting. Veel van de tekstontsluitingstechnieken worden belemmerd door de slechte OCR-kwaliteit. De beoogde resultaten van IMPACT zijn daarom van groot belang voor alle vormen van tekstontsluiting.

4.2 Spellingvariatie

1. Bij het omgaan met spellingvariatie moet onderscheid worden gemaakt tussen:
 - a. het vinden van spellingvarianten om de *retrieval* te verbeteren, d.w.z. zoeken in correcte spelling en ook woorden met OCR-fout vinden; of zoeken in moderne spelling en ook woorden in oudere spelling vinden. Voor dit doel maakt het niet uit of een woordvorm ontstaan is door historische spellingvariatie of door een OCR-fout. Het gaat er om dat de eindgebruiker bij *fulltext* zoekacties ook de spellingvarianten van een woord vindt;
 - b. het verbeteren van OCR-fouten in de tekst zelf, d.w.z. het bepalen van de juiste spellingvariant en die in de tekst opnemen. Voor dit doel is het wel nodig onderscheid te maken tussen historische spellingvarianten en OCR-fouten. Alleen de OCR-fouten moeten worden gecorrigeerd en vervangen door de historisch juiste spellingvariant.
2. Op korte termijn valt voor de KB-projecten de grootste winst te behalen met het verbeteren van de *retrieval* door het vinden van spellingvarianten en OCR-fouten. Dit levert voor de eindgebruiker een aanzienlijke verbetering in de doorzoekbaarheid van de teksten op.
3. De techniek die Martin Reynaert (ILK Tilburg) heeft ontwikkeld, biedt een goed uitgangspunt voor een dergelijke verbetering van de *retrieval*. Afhankelijk van de kwaliteit van de OCR wordt ongeveer 50-90% van de spellingvariatie door zijn aanpak gevonden. Naarmate de kwaliteit van de OCR slechter is, zijn de resultaten ook minder. Zelfs voor de slechtste OCR wordt echter nog zo'n 50% van de spellingvariatie gevonden. Naar verwachting zal de OCR voor de huidige en toekomstige digitaliseringsprojecten niet zo slecht zijn als waarmee in de pilot is gewerkt.
4. In een vervolgtraject zal de techniek van Reynaert worden ingezet voor de digitaliseringsprojecten van de KB. Het in de pilot ontwikkelde algoritme zal worden ingepast in de technische infrastructuur van de KB.
5. De techniek van Reynaert is in principe onafhankelijk van taal en historische periode. De historische lexica die verder ontwikkeld worden binnen het project IMPACT zullen de resultaten nog verbeteren
6. De techniek van Reynaert kan volautomatisch worden toegepast. Hooguit is enige *preprocessing* van de teksten nodig, bijvoorbeeld de afbreekstreepjes uit de tekst verwijderen. In het vervolgtraject zal dit worden geïmplementeerd en worden ingepast in het geheel.

7. Het algoritme van Reynaert vergt naar verwachting veel rekenkracht als het wordt toegepast op grote tekstcorpora. De schaalbaarheid zal in het vervolgtraject worden verbeterd door het efficiënter maken van het algoritme.
8. Reynaerts techniek maakt geen onderscheid tussen historische spellingvarianten en OCR-fouten.
9. Het verbeteren van OCR-fouten in de tekst zelf heeft voor de huidige KB-projecten minder hoge prioriteit dan de verbetering van de retrieval. Het corrigeren van de teksten is vooral van belang indien deze tekst aan de eindgebruiker wordt getoond. In de meeste huidige KB-projecten wordt primair de afbeelding van de pagina gepresenteerd; het tonen van de *fulltext* is een extra optie voor de eindgebruiker. Het verbeteren van de OCR-fouten in de tekst zelf kan ook van belang zijn als voorbereiding bij het classificeren of bij het genereren van samenvattingen.
10. Voor het verbeteren van de OCR-fouten in de tekst zelf is in het kader van de pilots nog geen kant en klare oplossing. De beste optie hiervoor is het verbeteren van het OCR-proces zelf, zodat er minder fouten zullen optreden. Dit is één van de hoofddoelstellingen van het IMPACT-project.
11. De resultaten van IMPACT leiden tot een OCR-technologie die betere OCR-resultaten oplevert, met minder OCR-fouten die achteraf verbeterd moeten worden. Toch is het zinvol het algoritme van Martin Reynaert voor de KB in te zetten: ten eerste om de spellingsvariatie in de huidige digitaliseringsprojecten te vinden en ten tweede omdat de OCR-technologie nooit 100% correcte teksten zal opleveren, waardoor enige correctie achteraf zinvol blijft.

4.3 Classificatie

12. Classificatie biedt de mogelijkheid om teksten op *onderwerp* te ontsluiten. In de huidige digitaliseringsprojecten is dat niet mogelijk door de beperkte hoeveelheid metagegevens. Classificatie biedt ook andere mogelijkheden, zoals bijvoorbeeld het ontsluiten op soort artikel in een krant (familiebericht, advertentie, nieuwsbericht e.d.).
13. De classificatie van teksten aan de hand van thesauri of gecontroleerde begrippenlijsten is voor sommige typen documenten geschikt. Zowel bij Staten-Generaal Digitaal als bij de kranten zijn er teksten waarvoor classificatie niet zinvol is.
14. Bij Staten-Generaal Digitaal wordt bij het classificeren met de Parlementsthesaurus een precisie van 35-90% gehaald en een recall van 30-90%. Er is een groot verschil in de prestaties, afhankelijk van het type document.
15. Bij de kranten wordt bij het classificeren met de IPTC een precisie van 40-46% gehaald en een recall van 56-64%. Deze cijfers kunnen aanzienlijk verbeteren (tot recall en precisie-percentages van 75-100%) door de drempelwaarde te verhogen. Dit heeft tot gevolg dat de *coverage* omlaag gaat tot 20%, wat wil zeggen dat er dan in 80% van de gevallen geen klasse wordt toegekend.
16. Classificatie van oudere documenten gaat minder goed dan bij moderne documenten. Het is nog niet duidelijk of de classificatie ook voor teksten van vóór de 20^e eeuw bruikbaar is. De vraag is met name of de bestaande thesauri of gecontroleerde woordenlijsten voor die periode geschikt zijn.
17. Voor de beschikbaarstelling (bv. website) moet worden bepaald hoe classificatie kan worden ingezet. Klassen kunnen als "metagegeven" dienen (bijvoorbeeld als zoekcriterium om zoekvragen te verfijnen) of om de ranking te verbeteren. Wat zinvol is, moet per collectie bepaald worden. Daarbij is raadpleging van gebruikerspanels van groot belang.

18. In het algemeen is een *recall* en een *precisie* van minstens 80% gewenst, wil de classificatie goed bruikbaar. Het vergt de nodige inspanning om de classificatie voor Staten-Generaal Digitaal of Databank Digitale Dagbladen op voldoende kwaliteitsniveau te krijgen. Dat brengt tevens de nodige kosten met zich mee. Dit geldt eveneens voor classificatie van eventuele andere collecties.
19. Voor Staten-Generaal Digitaal en Databank Digitale Dagbladen zal aan gebruikerspanels worden voorgelegd of classificatie zinvol is. Het is de vraag of de meerwaarde van classificatie in deze projecten opweegt tegen de benodigde inspanning, omdat het materiaal al op voldoende andere manieren doorzoekbaar is. Voor andere projecten zal de afweging afzonderlijk moeten worden gemaakt. Voor gedigitaliseerde collecties die niet veel andere zoekmogelijkheden bieden (omdat er weinig metagegevens zijn), kan automatische classificatie een goede mogelijkheid zijn.
20. Classificatie levert de beste inhoudelijke ontsluiting als zij voor iedere collectie afzonderlijk wordt toegepast. Dit betekent dat er voor iedere collectie een geschikte thesaurus moet worden gekozen en dat de classificatiesoftware afzonderlijk moet worden getraind. Mogelijk moet er binnen een collectie nog per historische periode worden bijgetraind.
21. Een alternatief is het classificeren van meerdere collecties met behulp van één thesaurus of classificatie. Dit heeft als voordeel dat de ontsluiting van al deze collecties uniform is en dat de techniek efficiënter en goedkoper kan worden ingezet. Het nadeel is dat de classificatie inhoudelijk minder goed is toegespitst op de specifieke collecties.
22. Na training van de software is het classificeren een volautomatische proces.
23. De invloed van lage OCR-kwaliteit op het classificatieproces is duidelijk merkbaar. De *Language Expansion Box* van Irion biedt in veel gevallen een verbetering. Deze LEB moet per collectie worden aangepast en mogelijk ook nog per periode. Mogelijk bieden andere technieken om de OCR te verbeteren eveneens een verbetering van de classificatie, zoals het algoritme van Martin Reynaert.
24. Waarschijnlijk verbetert de kwaliteit van de classificatie het meest als de OCR-technologie verbetert. De OCR-resultaten worden dan beter en er is minder noodzaak tot het verbeteren van OCR-fouten achteraf. De technologie die in het kader van IMPACT wordt ontwikkeld, lijkt hiervoor goede mogelijkheden te bieden.
25. Het beoordelen van de kwaliteit van classificatie blijft tot op zekere hoogte een subjectieve kwestie. Het blijkt dat er ook bij handmatige classificatie door deskundigen onderlinge verschillen zijn. Volgens een onderzoekje in het kader van het STITCH-project⁴ (*Semantic Interoperability to Access Cultural Heritage*) was er tussen een aantal titelbeschrijvers van de KB een onderlinge overeenstemming van 80%.
26. Het ontwikkelen en bijhouden van thesauri is gespecialiseerd werk, waarvoor veel inhoudelijke kennis van het domein en van thesauri nodig is. In het kader van de digitaliseringsprojecten van de KB is het niet haalbaar om thesauri te ontwikkelen. Wel kunnen bestaande thesauri worden gebruikt en voor specifieke projecten kunnen eenvoudige classificaties worden ontwikkeld, bv. “genres” (soorten berichten) bij kranten.

⁴ STITCH is een project binnen het NWO-researchprogramma CATCH (Continuous Access to Cultural Heritage) waarin de KB, de Vrije Universiteit en het Max Planck Instituut samenwerken. Het onderzoeksproject houdt zich bezig met Semantic Web en specifiek met het combineren van verschillende vocabulaires. Zie <http://www.cs.vu.nl/STITCH/>.

4.4 Samenvattingen

27. Het is belangrijk om onderscheid te maken tussen de mogelijke manieren waarop samenvattingen kunnen worden ingezet. De beslissing om samenvattingen toe te passen, hangt hier vanaf.
28. De eerste mogelijkheid is de samenvatting inzetten als vervanging van de oorspronkelijke tekst. Deze mogelijkheid is voor KB-digitaliseringsprojecten niet zo relevant. Over het algemeen wordt veel waarde gehecht aan het tonen van de oorspronkelijke tekst. De kwaliteit van de samenvatting is onvoldoende om als vervanging te dienen.
29. De tweede mogelijkheid is de "indicatieve samenvatting", waarbij de samenvatting de lezer helpt om te ontdekken waar een tekst over gaat. De in de pilot ontwikkelde applicatie (samenvatter) toont de hele tekst en markeert de belangrijkste zinnen (*ge-highlight*). Dit biedt de lezer sneller inzicht in de tekst, terwijl de hele context van de gemarkeerde zinnen kan worden bekeken. Dit is voor de gedigitaliseerde tekstcollecties van de KB een interessante mogelijkheid.
30. Een derde mogelijkheid van de applicatie is het genereren van "snippets" bij zoekresultaten. Dat zijn hele korte samenvattingen, bestaande uit korte zinnen of zinsnedes, die gebruikers helpen te bepalen welke zoekresultaten interessant zijn. Deze mogelijkheid is voor de KB interessant, maar de behaalde kwaliteit zou vergeleken moeten worden met de mogelijkheden van de huidige zoekmachine van de KB, Verity.
31. De door Carp ontwikkelde applicatie biedt samenvattingen van redelijke kwaliteit. Na enige *fine tuning* lijkt de kwaliteit voldoende voor indicatieve samenvattingen en snippets.
32. De manier waarop de samenvattingen worden getoond (zinnen highlighten en schuifbalk) is aantrekkelijk en gebruiksvriendelijk. De KB zal onderzoeken of de applicatie als onafhankelijke service kan worden geïmplementeerd, zodat de samenvatter in allerlei omgevingen kan worden aangeboden.
33. Het is de vraag hoeveel toegevoegde waarde de samenvatter voor eindgebruikers heeft. Raadplegen van gebruikerspanels moet hierover meer duidelijkheid geven.
34. Het beoordelen van de kwaliteit van samenvattingen is subjectief. Verschillende personen hebben een ander oordeel over wat een goede samenvatting is. Carp citeert onderzoek waaruit blijkt dat in samenvattingen die door mensen gemaakt zijn slechts 50-60% van de zinnen overeenkomen.

4.5 Named entity recognition

35. NER biedt de mogelijkheid om teksten op nieuwe manieren doorzoekbaar te maken en om bij "entiteiten" extra informatie of *services* aan te bieden, bijvoorbeeld biografische informatie bij persoonsnamen of linken naar een elektronische encyclopedie met achtergrondinformatie over de *entiteit*.
36. Voor Staten-Generaal Digitaal wordt de gewenste named entity recognition binnen de KB gerealiseerd.
37. Voor de overige projecten zullen de resultaten van het IMPACT-project worden afgewacht, waarin onder meer lijsten met namen van personen, plaatsen e.d. worden ontwikkeld ten behoeve van NER.
38. Per digitaliseringsproject moeten worden besloten hoe NER zinvol kan worden ingezet.