Folgert Karsdorp

Het is groen en leeft nog lang en gelukkig

Classificatie van volksverhaalgenres op basis van formules*

Abstract – Different folktale genres make use of various conventional ways of starting a story. How well can we predict the genre of a folktale based on these opening formulas? In this paper I show by means of computational analysis that Dutch and Frisian folktale formulas in many cases already provide enough information to correctly classify a folktale according to its genre. A more in-depth analysis reveals that idiomatic, fully lexically-specified opening phrases serve not only as predictors of genre but also as more abstract schemas with open slots at different levels of abstraction. I characterize the formulas by classifying them according to their primary communicative functions.

1 Introductie

'Er was eens …' is de karakteristieke openingsformule van sprookjes. De formule is geattesteerd in Chaucers *Canterbury Tales* uit 1385, en zelfs nog eerder in Aristophanes' *De Wespen* in het jaar 422. Sprookjes komen in veel talen en culturen voor, met vaak verrassend veel overeenkomsten. Het is dan ook niet verwonderlijk dat andere talen gelijksoortige openingszinnen hebben, zoals het welbekende 'Once upon a time …' in het Engels, 'Il était une fois …' in het Frans en 'Reiz sen senos laikos …' ('ooit, lang geleden in voorbije tijden') in het Lets.

De slotzin van sprookjes laat meer variatie zien. Toch zijn ook hier conventionele patronen aan te wijzen. In het Nederlands kennen we uiteraard het '…en ze leefden nog lang en gelukkig', maar vaak genoeg komt er ook een dier met een al dan niet lange snuit die een einde maakt aan het verhaal. In het Catalaans eindigen verhalen vaak met 'I conte contat, conte acabat' ('verhaal verteld, verhaal afgelopen'). In het Afrikaans is de rijmende slotzin 'Fluit, fluit, die storie is uit' populair.

Niet alleen sprookjes hebben conventionele openingen en afsluitingen. Sagen, legenden, moppen, en raadsels hebben ieder hun eigen kenmerkende frasen. 'Lopen twee mannen …' bijvoorbeeld, met het werkwoord in de eerste zinspositie, is een klassieke manier om een mop te beginnen (zie Bennis 2001). Dergelijke frasen ogen minder gefixeerd en lijken minder formulaïsch van aard. Het hoeven geen twee mannen te zijn: één of drie kan ook. Vijf of meer is echter uitzonderlijk. Het zijn ook niet noodzakelijk mannen, al hebben die (in Nederlandse moppen) een hoge waarschijnlijkheid.[1] En uiteraard hoeft de uitgedrukte handeling er niet een te zijn van lopen. Toch herkennen we een dergelijke frase direct als het begin van een mop. Er lijken dus niet alleen volledig gefixeerde frasen te zijn met symbolische status, maar ook frasen die in een mindere mate lexicaal gefixeerd zijn. Deze

* Mijn dank gaat uit naar Theo Meder, Antal van den Bosch en Marije Koens voor uitgebreid commentaar op eerdere versies van dit stuk.

1 Blijkens een telling in de Nederlandse volksverhalenbank.

frasen hebben één of meerdere open plekken die variabel ingevuld kunnen worden.

Intuïtief lijken volksverhalen dus gebruik te maken van heel stabiele manieren om een verhaal te openen en af te sluiten. Bovendien lijken de verschillende genres gebruik te maken van genreonderscheidende openingen en afsluitingen. De vraag die in dit artikel centraal staat, is of we enige empirische basis kunnen vinden voor deze intuïtie. Als openingen en afsluitingen van volksverhalen zulke stabiele en conventionele elementen zijn, kunnen ze dan worden gebruikt om het genre van een volksverhaal te voorspellen? Deze vraag zal ik onderzoeken met behulp van een computationeel classificatie-experiment. Dit experiment moet inzichtelijk maken welke openingen en afsluitingen goede voorspellers zijn voor een volksverhaalgenre. Zijn dat vooral lexicaal gefixeerde frasen of kunnen we ook schema's ontdekken die gebruikt worden voor een specifiek genre?

Als verschillende openingen als voorspellers dienen voor hetzelfde volksverhaalgenre, is het interessant te onderzoeken wat die formules met elkaar gemeen hebben. Een mogelijke dimensie van overeenkomst is hun functie. Openingsformules dienen verschillende functies, maar uiteindelijk is de primaire functie van formules communicatief van aard:

> Die erzählende Mensch hat die [Eingangsformeln] seines erzählgutes mit verschiedene Funktionen ausgestattet, deren wichtigste wohl sind, die Kommunikation mit dem Zuhörerskreis herzustellen und ihn in das Erzählgeschehen einzuführen (*EM 1975*: 1227).

In een kwalitatieve analyse van de resultaten wil ik onderzoeken welke communicatieve functies de openingen vervullen en welke communicatieve functies dominant en onderscheidend zijn binnen een volksverhaalgenre.

De opbouw van het artikel is als volgt. Ik zal beginnen met een korte beschrijving van het corpus dat ik heb gebruikt (§2). Daarna beschrijf ik de experimentele opzet (§3) en geef ik kwantitatieve resultaten (§4.1). Voordat ik kort de belangrijkste conclusies van het onderzoek bespreek (§5), zal ik in paragraaf 4.2 een kwalitatieve analyse van de resultaten geven.

2 De Nederlandse verhalenbank

De Nederlandse volksverhalenbank[2] bevat ongeveer 42.000 volksverhalen (Meder 2010). De collectie bestaat uit volksverhalen uit verschillende genres, zoals sprookjes, legendes, broodjeaapverhalen en moppen. Er zijn verhalen uit verschillende dialecten van het Nederlands. Elk verhaal in de databank bevat metadata over de taal van het verhaal, de verzamelaar, de plaats en datum van de vertelling, trefwoorden, namen van personages of plaatsen in de verhalen en tot slot het genre. De twee grootste componenten zijn geschreven in het Standaardnederlands en in het Fries. Ik beperk mij in dit artikel tot de verhalen geschreven in deze talen.

2 www.verhalenbank.nl

2.1 Genres

In dit artikel staan zeven volksverhaalgenres centraal. De selectie is gebaseerd op het classificatiesysteem in de volksverhalenbank. Mythes, limericks en liederen laat ik buiten beschouwing omdat die slechts zeer beperkt gerepresenteerd zijn in de databank. Ik geef voor elk genre een korte beschrijving.

– *Sprookje:* Sprookjes zijn spatio-temporeel onbepaalde verhalen. De plot concentreert zich op een enkele protagonist die een reeks opdrachten moet vervullen om zijn of haar doel te bereiken. Sprookjes bevatten vaak elementen van magie en worden als onwaar beschouwd.

– *Sage:* Sagen zijn in tegenstelling tot sprookjes spatio-temporeel bepaald en hebben plaats in het recente verleden van de vertelling. Ze worden (of werden) als waar beschouwd.

– *Broodjeaapverhaal:* Broodjeaapverhalen, ook wel *urban* of *contemporary legends* genaamd, verhalen over het heden. De verhalen worden als waar gepresenteerd en bevatten vaak elementen van mysterie, humor en horror.

– *Mop:* De volksverhalenbank bevat veel korte verhalen die geclassificeerd zijn als mop. Moderne moppen eindigen vaak met een *punchline* en zijn relatief kort. De oudere moppen in de databank kunnen substantieel langer zijn, hebben overeenkomsten met kluchten en boerden en eindigen niet noodzakelijk met een grappige laatste zin.

– *Raadsel:* Hedendaagse raadsels en moppen hebben veel gemeen. Het verschil is dat raadsels voornamelijk beginnen met een vraag gevolgd door een al dan niet grappig antwoord. Oudere raadsels functioneren als puzzels.

– *Kwispel:* De kwispel is een vrij nieuw volksverhaal genre (zie Meder & Burger (2006) voor een uitgebreide beschrijving). Kwispels beginnen met de vaak mysterieuze uitkomst van het verhaal waarna de toehoorders moeten achterhalen welke opeenvolging van plotelementen aan deze uitkomst is voorafgegaan.

– *Persoonlijke narratief:* De meeste genres in de verhalenbank hebben hun oorsprong in een orale traditie en zijn het resultaat van herhaaldelijk hervertellen door verschillende personen over grote(re) geografische afstand. Persoonlijke narratieven verschillen in dit opzicht en verhalen over persoonlijke herinneringen van de verteller.

2.2 Statistiek

Tabel 1 geeft enkele statistieken over de gebruikte collectie. Zoals duidelijk blijkt uit de tabel is de distributie van genres erg scheef. Voor het Nederlands zijn veel voorbeelden van sagen en moppen bekend en slechts een aantal persoonlijke narratieven en kwispels. De Friese verhalen kennen een nog schevere distributie, waarbij vrijwel alle verhalen geclassificeerd zijn als sage.

Tabel 1 De datasets uitgesplitst naar genre.

Genre	*Nederlands*	*Fries*	*totaal*
Sprookje	831	452	1283
Sage	5922	13898	19820
broodjeaapverhaal	2657	14	2671
Mop	2864	2416	5280
Legende	343	14	357
persoonlijke narratief	682	16	698
Raadsel	1560	21	1581
Kwispel	73	0	73
Totaal	**14932**	**16831**	**31763**

3 Experimentele opzet

In deze sectie bespreek ik de experimentele opzet en computationele methoden van het onderzoek. Ik begin met een bespreking van het classificatiesysteem waarin ik kort het basale idee toelicht dat ten grondslag ligt aan de methode van inductieve *machine learning.* Vervolgens zal ik in paragraaf 3.2 de kenmerken bespreken op basis waarvan ik het classifcatiesysteem train. Paragraaf 3.3 geeft een uitleg van de gekozen evaluatiemethodes.

3.1 Classificatiesysteem

Kunnen we op basis van eerdere observaties voorspellingen doen over nieuwe, ongeziene observaties? Deze vraag vormt de kern van veel technieken binnen de discipline van *machine learning*. Laat D het trainingsmateriaal zijn bestaande uit N paren van kenmerken x en uitkomsten y. Het doel is om via inductief redeneren een functie $y = f(x)$ te leren waarmee we voorspellingen kunnen doen over nieuwe voorbeelden. In het geval van genreclassificatie staat y voor het genre waarmee een tekst geclassificeerd is. x staat voor de kenmerken die we extraheren uit een tekst (zie paragraaf 3.2). Het doel is nu om een functie f te leren waarmee we het genre y van een nog ongeziene tekst kunnen voorspellen op basis van de kernmerken x. Er zijn veel verschillende systemen waarmee we een dergelijke functie kunnen leren, waarvan de precieze werking voor de huidige studie niet van groot belang is. In deze studie gebruik ik een *Stochastic Gradient Descent Classifier* zoals die geïmplementeerd is in de Machine Learning Python-bibliotheek *Scikit Learn* (Pedregosa e.a. 2011).[3]

3 Zie www.scikit-learn.org. Ik hanteer L2-normalisatie en draai het algoritme 50 keer, met *smoothing*-parameter $\alpha = 0.0005$. Net als de meeste *smoothing*-parameters moet ook deze parameter handmatig ingesteld worden. Ik heb de parameter geoptimaliseerd met behulp van een *grid-search procedure* op basis van een enkele *fold* en gebruik de waarde α die daarbij het best presteert op F-score.

3.2 Kenmerken

Voor elk volksverhaal extraheer ik het genre van het verhaal, de eerste vijf woorden van het verhaal en de laatste vijf woorden die respectievelijk de openings- en sluitingsformules moeten representeren. Er zijn verschillende manieren om de openingen en afsluitingen te representeren. Ik vergelijk twee methoden.

In de eerste methode gebruik ik *n*-grammen van woorden waarbij *n* voor het aantal aaneengesloten woorden staat. Voor zowel de eerste als de laatste vijf woorden van een verhaal extraheer ik *n*-grammen van lengte 1 tot 5. Voor een openingsfrase als 'Twee mannen lopen op straat' levert dit de volgende 15 kenmerken op:

- unigrammen: *twee*, *mannen*, *lopen*, *op, straat*
- bigrammen: *twee mannen*, *mannen lopen*, *lopen op, op straat*
- trigrammen: *twee mannen lopen*, *mannen lopen op*, *lopen op straat*
- quadrigrammen: *twee mannen lopen op*, *mannen lopen op straat*
- pentagrammen: *twee mannen lopen op straat*

De tweede representatiewijze maakt gebruik van zogenoemde *skip*-grammen, waarbij bepaalde woorden worden overgeslagen. Een zin als 'de heks at het dikke jongetje' bevat de trigrammen *de heks at*, *heks at het*, *at het dikke*, en *het dikke jongetje*. *Skip*-grammen geven ons de mogelijkheid om naast deze trigrammen ook ogenschijnlijk even belangrijke trigrammen te extraheren zoals *at het jongetje* en *heks at het jongetje*, die niet meegenomen worden in reguliere *n*-grammen. Ik extraheer alle *skip*-grammen voor de eerste en de laatste vijf woorden van een verhaal met een maximale *skip* van $k = 4$ (maximaal vier woorden mogen worden overgeslagen). Ik definieer de set van *skip*-grammen voor een zin $w_1, \ldots, w_n$ als:

(1) $$\{w_{i_1}, w_{i_2}, \ldots, w_{i_n} | \sum_{j=1}^{n} i_j - i_{j-1} < k\}$$

waarbij *k* het maximaal aantal skips aanduidt. Een voorbeeld ter verduidelijking. De opening 'Twee mannen lopen op straat' levert de volgende 31 kenmerken op:

- unigrammen: *twee*, *mannen*, *lopen*, *op*, *straat*
- bigrammen: *twee mannen*, *twee lopen*, *twee op*, *twee straat*, *mannen lopen, mannen op, mannen straat, lopen op, lopen straat, op straat*
- trigrammen: *twee mannen lopen*, *twee mannen op*, *twee mannen straat*, *twee lopen op*, *twee lopen straat*, *twee op straat*, *mannen lopen op*, *mannen lopen straat*, *mannen op straat*, *lopen op straat*
- quadrigrammen: *twee mannen lopen op*, *twee mannen lopen straat*, *twee mannen op straat*, *twee lopen op straat, mannen lopen op straat*
- pentagrammen: *twee mannen lopen op straat*

Het aantal kenmerken is met het gebruik van *skip*-grammen meer dan verdubbeld. *Skip*-grammen zijn daarmee een veel krachtigere set van kenmerken dan reguliere *n*-grammen.

Skip-grammen zijn met name interessant omdat ze niet alleen vaste patronen van aaneengrenzende woorden zichtbaar maken. Met *skip*-grammen is het mogelijk openings- of afsluitings*schema's* zichtbaar te maken op verschillende niveaus

van abstractie. Zo zou het kunnen dat een schema als *twee _ lopen op _*, waarbij de *underscores* open plekken aanduiden, sterk geassocieerd is met moppen.

Nguyen e.a. (2012) hebben laten zien dat *n*-grammen van letters (van lengte 2 tot 5) de effectiefste kenmerken zijn voor genreclassificatie van volksverhalen. Zij hebben zich niet beperkt tot de openings- en slotformules, maar hebben alleen naar de teksten als geheel gekeken. Ik vergelijk de hierboven beschreven methoden met een basismodel dat letter-*n*-grammen gebruikt (lengte 2 tot 5) voor de hele tekst en niet alleen openings- en slotformules. Ik kies er bewust voor om niet ook een letter-*n*-grammodel te maken op basis van de openings- en slotformules, omdat de resultaten van letter-*n*-grammen minder helder interpreteerbaar zijn dan die van woord-*n*-grammen of *skip*-grammen.

Binnen de discipline van *Natural Language Processing* en *Computer Science* in het algemeen, is de vergelijking met een basismodel een klassieke experimentele opzet. Ik ben echter niet per se geïnteresseerd in het behalen van betere classificatieprestaties, of het 'beste' model om verhalen naar genres te kunnen classificeren. De vergelijking is daarom vooral ter referentie hoe de beschreven methode zich verhoudt tot de *state of the art*. Een belangrijker doel van deze paper is het ontdekken van nieuwe interpreteerbare en kennisgevende kenmerken die typisch zijn voor de genres.

3.3 Evaluatie

Ik pas 10-*fold cross-validation* toe op zowel de Nederlandse als de Friese dataset, waarbij we de dataset willekeurig verdelen in 10 stukken van ongeveer gelijke grootte. Telkens wordt één tekstdeel (*fold*) als testset gebruikt en de andere negen als trainingsdata. Deze methode biedt de mogelijkheid tot nauwkeurige evaluatie van het potentieel van het model om om te gaan met ongeziene teksten.

De resultaten evalueer ik op *precision*, *recall* en *F-score* (Van Rijsbergen 1979). De maten kunnen zowel voor alle resultaten samen als voor de individuele uitkomstcategorieën (de genres) worden berekend. De *precision* is de proportie van de als genre *y* geclassificeerde documenten die het model correct voorspelt, en wordt berekend door het aantal goed voorspelde documenten van *y* (ware positieven) te delen door alle gevallen die de *classifier* - correct of abusievelijk - als *y* heeft geclassificeerd. De *recall* daarentegen geeft de verhouding van de eigenlijke gevallen van *y* die door het model correct voorspeld worden. De *recall* wordt berekend door de ware positieven te delen door de som van de ware positieven en de documenten die ten onrechte niet als *y* zijn geclassificeerd (foute negatieven). De *F-score* geeft het harmonische gemiddelde van de *precision* en de *recall*. Hieronder zijn de drie definities gegeven:

$$\text{(2)} \qquad \text{Precision} = \frac{\text{Ware positieven}}{\text{Ware positieven + foute positieven}}$$

$$\text{(3)} \qquad \text{Recall} = \frac{\text{Ware positieven}}{\text{Ware positieven + foute negatieven}}$$

$$\text{(4)} \qquad \text{F} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}}$$

4 Resultaten

In deze sectie zal ik de resultaten van de experimenten bespreken. Ik zal beginnen met een kwantitatief overzicht waarin ik de verschillende configuraties evalueer op basis van de genoemde evaluatiematen. Vervolgens zal ik een gedetailleerdere interpretatie geven van de resultaten vanuit een kwalitatief perspectief.

4.1 Kwantitatief overzicht

4.1.1 Nederlands

Tabel 2 geeft de resultaten voor het Nederlands voor beide kenmerkrepresentaties voor openingsfrasen en sluitingsfrasen tegenover de resultaten op basis van letter-*n*-grammen.

Tabel 2 Resultaten voor de Nederlandse verhalen. Voor elke methode is de *precision*, *recall* en macro F-score gegeven.

		precision	*recall*	*F-score (macro)*
n-grammen (woorden)	openingsfrases	0.68	0.56	0.59
	sluitingsfrases	0.41	0.31	0.33
skip-grammen	openingsfrases	0.71	0.56	0.60
	sluitingsfrases	0.39	0.29	0.30
n-grammen (letters)		0.72	0.67	0.68

De resultaten laten duidelijk zien dat openingsfrases betere voorspellers zijn dan sluitingsfrasen. De representatie met *skip*-grammen levert iets betere resultaten dan die met reguliere *n*-grammen, maar het verschil is minimaal. De combinatie van openings- en sluitingsfrases levert voor beide representaties een kleine verbetering op: voor *n*-grammen F = 0.61; voor *skip*-grammen F = 0.62. De hogere precisie in vergelijking met *recall* duidt erop dat het systeem een aantal sterk discriminerende kenmerken heeft gevonden die eenduidig aansturen op een genre. De lagere *recall* laat zien dat met deze kenmerken niet alle gevallen worden afgedekt. De resultaten van de *n*-gram-methode op basis van letters presteert beter dan de andere methodes. Het verschil is echter beperkt en zowel de woord-*n*-grammethode als de *skip*-grammethode hebben een hogere precisie op basis van alleen de openingsfrases.

In onderstaande tabel geef ik nogmaals de scores op basis van de *skip*-gramrepresentatie, maar dit keer uitgesplitst per genre voor een enkele *fold* (voor 10 procent van de data, getraind op de overige 90 procent). Per genre geef ik de F-scores voor openings- en sluitingsfrasen en de combinatie ervan.

Kwispels, moppen, raadsels en sagen zijn goed herkenbaar op basis van de eerste vijf woorden. De extra informatie die sluitingsfrases geven, leidt slechts tot beperkte verbeteringen. Broodjeaapverhalen worden redelijk accuraat voorspeld op basis van de opening. We zien hier dat de combinatie van de eerste en de laatste vijf woorden tot de beste scores leidt. In mindere mate zien we dat ook bij sprookjes en raadsels.

TABEL 3 Resultaten voor de Nederlandse verhalen uitgesplitst naar genre. Voor elk genre is de F-score (micro) gegeven voor openingsfrasen, sluitingsfrasen en de combinatie van de twee.

	openingsfrases	*sluitingsfrases*	*combinatie*	*# verhalen*
Broodjeaapverhaal	0.53	0.39	0.60	238
Kwispel	1.00	0.00	1.00	8
Legende	0.30	0.05	0.26	36
Mop	0.64	0.47	0.67	273
persoonlijke narratief	0.27	0.16	0.36	77
Raadsel	0.89	0.54	0.91	169
Sage	0.77	0.67	0.79	622
Sprookje	0.48	0.27	0.59	70
Gemiddelde	**0.67**	**0.51**	**0.71**	**1493**

4.1.2 Fries

Tabel 4 geeft de globale scores voor de Friese verhalen. De resultaten laten hetzelfde beeld zien als voor het Nederlands. Ook hier fungeren openingsfrases als sterkere indicatoren van het genre. De *skip*-grammethode presteert ook voor het Fries iets beter dan de reguliere *n*-grammen. Wederom presteert de *n*-grammethode met letters het beste, maar ook hier is het verschil niet groot.

TABEL 4 Resultaten voor de Friese verhalen. Voor elke methode is de *precision*, *recall* en macro F-score gegeven.

		precision	*recall*	*F-score (macro)*
n-grammen (woorden)	openingsfrases	0.31	0.24	0.26
	sluitingsfrases	0.35	0.21	0.23
skip-grammen	openingsfrases	0.32	0.29	0.30
	sluitingsfrases	0.29	0.24	0.26
n-grammen (letters)		0.37	0.36	0.36

Ook voor de Friese verhalen geef ik de F-scores op basis van de *skip*-grammethode uitgesplitst naar genre voor openings- en sluitingsfrasen en de combinatie ervan in tabel 5.

De resultaten maken duidelijk zichtbaar dat het systeem in staat is om de grote categorieën (sagen en moppen) met redelijke zekerheid te voorspellen. Dit verklaart ook de veel lagere macro F-scores in Tabel 4. De bijzonder scheve verdeling van de data maakt het moeilijk om voor de minderheidscategorieën discriminerende kenmerken te ontdekken die eenduidig aansturen op een juiste classificatie.

Tabel 5 Resultaten voor de Friese verhalen uitgesplitst naar genre. Voor elk genre is de F-score (micro) gegeven voor openingsfrasen, sluitingsfrasen en de combinatie van de twee.

	openingsfrases	*sluitingsfrases*	*combinatie*	*# verhalen*
Broodjeaapverhaal	0.00	0.00	0.00	3
Legende	0.00	0.00	0.00	1
Mop	0.56	0.45	0.63	237
persoonlijke narratief	0.00	0.00	0.00	1
Raadsel	0.00	0.00	0.00	1
Sage	0.92	0.91	0.93	1395
Sprookje	0.43	0.16	0.40	45
Gemiddelde	**0.85**	**0.82**	**0.87**	**1683**

4.2 Kwalitatief overzicht

De kwantitatieve resultaten besproken in de vorige sectie, hebben er ons een beeld van gegeven welke volksverhaalgenres te voorspellen zijn op basis van openings- en sluitingsformules. Het is te verwachten dat genres die beter te voorspellen zijn, discriminatievere kenmerken hebben in hun openingen en ook minder variatie. Welke kenmerken zijn dat? Wat zijn de meest discriminatieve kenmerken per genre? Dat wil zeggen, welke kenmerken hebben de grootste voorspellende kracht? Welke overeenkomsten vertonen openingen binnen een bepaald genre?[4] Kunnen we de openingen categoriseren naar hun communicatieve functie en daarmee laten zien welke communicatieve functies dominant aanwezig zijn in elk volksverhaalgenre?

De *Enzyklopädie des Märchens* onderscheidt drie primaire categorieën van openingsformules:

1 *Contact*: deze eerste categorie bevat formules die het publiek voorbereiden op het te vertellen verhaal, zoals 'Hoor toe' of 'Ik ga u een verhaal vertellen van ...'. Dergelijke formules hebben de functie om contact met het publiek te leggen en kunnen we categoriseren als 'contactformules'.

2 *Credibilis* of *incredibilis*: vaak wordt een opening gebruikt om het publiek te informeren over de waarheidswaarde van het verhaal: 'wat ik u nu ga vertellen berust op niets dan de waarheid'. Of men doet een beroep op een autoriteit, bijvoorbeeld de grootmoeder, door te beginnen met 'Mijn grootmoeder vertelde altijd ...'. Deze formules vallen onder de categorie '*credibilis* of *incredibilis*'.

3 *Ruimte en tijd*: tot slot zijn er de formules die het verhaal plaatsen in de tijd of ruimte. Zowel ruimte als tijd kunnen bepaald dan wel onbepaald zijn. Sprook-

4 Ik zal in het onderstaande alleen de genres voorspellen waarvoor de classifier een F-score > 0 heeft behaald. Ook zal ik alleen een interpretatie geven van de openingsfrasen omdat hiermee evident betere resultaten werden behaald.

jes zijn het prototypische voorbeeld van verhalen die de toehoorder naar een onbepaalde tijd en onbepaalde plek leiden. Sagen daarentegen, zijn doorgaans spatio-temporeel bepaald. Deze formules zal ik scharen onder de categorie 'ruimte en tijd'.

4.2.1 Nederlands

De belangrijkste voorspellers voor broodjeaapverhalen zijn de volgende kenmerken:

(5) *in Nederland, Amsterdam, poema, Amerikaanse, Druten, Londen, jarige, een familie, cola, geruchten, aids, Amerika, een _ stel, echtpaar, politie, categorie, dit vertelde, sinaasappels, Robert, Rotterdam, kunstgebit, studenten, dit _ verhaal, gerucht, Ede, Den Haag, een _ vrouw, Apeldoorn, inclusief, internet, afgelopen, Robbert, dagblad, vakantie, de politie, as, Franse, vriendin, Brussel, lezen, het lijkt, rubriek, Volkert, motorrijder, een _ meisje*

Veel van de kenmerken hebben betrekking op locaties (*Nederland, Londen, Amsterdam, Amerika, Druten, Rotterdam*) en plaatsen het verhaal in een temporeel gebonden setting (*afgelopen week / maand, augustus, (op) vakantie*). De feitelijkheid (*credibilis*) van het verhaal wordt ondersteund door een externe autoriteit bij het verhaal te betrekken zoals 'een vriendin [...] vertelde een verhaal' of 'dit vertelde mijn biologieleraar'. Een andere expliciete opening binnen de categorie 'credibilis of incredibilis' is 'dit verhaal is echt gebeurd'. Verder zijn er verschillende entiteiten die typisch een rol spelen in broodjeaapverhalen en daarmee een onderscheidende functie vervullen, zoals *cola*, *aids* en *tuinkabouter*. Verschillende broodjeaapverhalen beginnen met de contactformule *dit _ verhaal* als in 'dit verhaal is echt gebeurd' of 'dit verhaal hoorde ik eind ...' waarmee expliciet aangegeven wordt dat de verteller aan een verhaal begint. De beste voorspellers van broodjeaapverhalen plaatsen het verhaal in een spatio-temporeel bepaalde setting en doen een beroep op de geloofwaardigheid van het verhaal.

Kwispels worden gekenmerkt door veel onbepaalde voorspellers, zowel in tijd als in ruimte:

(6) *opdracht, oplossing, is _ iets, er ligt een man, deze _ is, er _ man, er _ een man, er ligt, een _ man, bord, de _ was, in _ een, ligt _ een, er ligt een _, als hij, van groot, deze _ een, er _ dood, een man _ dood, _ man _ dood, er _ een _ dood, een _ dood, tijdens de _ was, tijdens de, en _ zijn, er _ een, rijd, toen _ de, haar _ is*

Het schema *_ man _ dood* wordt regelmatig ingevuld met het onbepaalde lidwoord *een* en het werkwoord *liggen*. Veel andere schema's hebben betrekking op zowel een onbepaalde persoon als onbepaalde plaats delict. De prominentste formule begint met *opdracht* en valt onder de categorie 'contactformules' met een haast symbolische functie voor kwispels. Hierbij moet echter opgemerkt worden dat de meeste kwispels in de verhalenbank uit de schriftelijke overlevering afkomstig zijn waarbij heel frequent het woord *opdracht* op de eerste positie staat. Een ander voorbeeld van een contactopening vinden we in 'Deze is iets schokkender ...' waarmee de verteller de luisteraars voorbereidt op het te vertellen verhaal.

De beste voorspellers van legendes zijn de volgende kenmerken:

(7) *Willebrord, Sint, de heilige, Brunssum, Radboud, St., ol vrouw, relieken, Lambertus, Civitavecchia, heilige, hostie, visioen, het visioen, Rome, Roermond, mirakel, Gerlacus, mariabeeldje, Onze Lieve, Civitavecchia het, eligius, St. Lambertus, Servaas, het graf, het wonder, Onze Lieve Vrouwe, Lieve Vrouwe, Onze Vrouwe, heiligenlegenden, kent heiligenlegenden, het mariabeeldje, Johannes het, Schiedam, Amandus, het _ der, verhaalt, wonderen, Servatius, de _ klokken, klokken, Kemp, Marie, onder Schaesberg*

De lijst bestaat uit plaatsnamen en veel (heilige) personen die het verhaal in een spatio-temporeel bepaalde context plaatsen. De lijst bestaat voornamelijk uit enkele woorden. Bi- of meergrammen worden nauwelijks genoemd evenals *skip*-grammen. De personen en plaatsen zijn discriminerend genoeg om het genre te herkennen en er zijn weinig formulaïsche openingen.

Goede voorspellers voor moppen zijn welbekende moppersonages als *Jantje*, *Sam* en *Moos* die ofwel alleen ofwel als duo direct de verwachting van een mop oproepen. Sommige personages zijn tijdsgebonden zoals Clinton, Saddam en Dutroux. Verder zien we bekende entiteiten als *blondjes* en *Belgen.* Werkwoorden op de eerste positie zijn veelvuldig kenmerkend:

(8) *vraagt, zit, komt, klant, Jantje, Moos, er _ twee, loopt, Sam, er is _ een, Clinton, Saddam, zegt de, dokter, in Enschede, Belg, Petrus, grappen, een pastoor, een _ komt, Hans, een staat, dat dan, een poep, stapt, een loopt, Belgische, pastoor, Temel, een _ en een _, lopen, zegt, toentertijd, een mop, belt, mop, Belgen, op school, Dutroux, Turk, Belgisch, twee, Grolsch, Bill, ja jongens, blondje, helpdesk, gisteren*

De personages zijn (meestal) onbepaald maar worden niet zoals sprookjes expliciet in een onbepaalde tijd geplaatst (niet: 'Er was eens een man in een bar ...', maar: 'Een man in een bar ...'). Enkele opvallende moppenschema's zijn:

- *er _ twee* zoals in 'er lopen twee meisjes' of 'er waren eens twee weilanden';
- *een _ komt* zoals in 'een Belg komt een café binnen', 'een dom blondje komt ...' en 'een Surinamer komt werkschoenen kopen ...';
- *een _ en een _* zoals in 'een man en een vrouw', 'een beer en een konijn' en 'een slagwerker en een altviolist'.

Een groot aantal moppen begint met het telwoord *twee*. Enkele voorbeelden: *twee politieagenten staan ..., twee soepkommen liggen in bed, twee eieren komen elkaar tegen.* Met enige frequentie wordt gevraagd of de toehoorder de te vertellen mop al kent met als contactformule: 'ken je die (mop) van ...'.

Persoonlijke narratieven benoemen bestaande locaties waarmee het verhaal aan een plaats verbonden wordt (*Flevoland, Zuid-Holland, Groningen*):

(9) *blauwdruk, rapport, haarteuten, via het, Flevoland, graancirkel, alweer, geliefde, op meldde, heb zelf, stadskanaal Groningen, bandopname, formatie, Mia, Lakeland, in Lakeland, als ik, wat ik me, wat me, Brink, religieuze, sjamanisme, je vertelde, Gerrits, Zuid-Holland, schele, op kreeg, telefoon, hoe kijkt, meldde, op _ juli, zelf wel, noord, kent gebruiken, stoelenmatters*

De persoonlijke aard van het verhaal wordt vaak expliciet gemaakt door frasen te gebruiken als *wat ik me*, *wat me, als ik*, *zelf wel* waarmee de geloofwaardigheid bij de verteller wordt gelegd.

De volgende kenmerken hebben grote voorspellende kracht voor raadsels:

(10) *waarom, hoeveel, waarvoor, hoe, wat _ als, toppunt, wat _ een, wat zijn, wat, toppunt van, het loopt, wat _ doen, hoe _ een, is _ rood, het _ rood, het is _ rood, is en, het zit, wat hebben, het _ en _, waar ligt, groen, het _ vliegt, het is _ en, wanneer is, weet _ waar, hoe komt, wat zei, het _ ligt, wat is, het _ door, weet je, is groen, wat eten, zie je, vraag, Diana, betekent, wat betekent, mop, wat vond, hoe _ de, hoeveel _ van, wanneer weet, wie _ het, ligt, het _ zit in, welke*

De lijst bestaat vrijwel alleen uit (schema's met) vragende voornaamwoorden. Zo kan het schema *wat _ een* ingevuld worden met 'wat is een vuilnisbelt met' of 'wat doet een officier met'. En ander frequent gebruikt schema is *het _ en _* dat ingevuld kan worden door 'het is groen en skiet ...' of 'het is rood/zwart/bruin en het'. Indirect dienen deze openingen een contactfunctie: de conventionele status van het vragende voornaamwoord maakt het genre vrijwel eenduidig kenbaar.

De topvoorspellers van sagen zijn de volgende:

(11) *een _ uit, bekend, heksen, witte, in _ staat, ik _ eens, onderaardse, weerwolf, spokerijen, een _ heks, heden, men, naam, diezelfde, me, geesten, dwaallichtjes, tovenaars, kabouters, gehucht, zowat, was _ een, bokkenrijders, volgens de, kasteel, avonds, weerwolven, het _ als, de _ der, tovenaar, heks, betoverde, Bartje, ik werkte, molenaar, op _ in, ons vader, tv, weert, eeuw, hij _ komt, in een _ van*

Anders dan wordt beweerd in de *Enzyklopädie des Märchens* zijn er zowel veel bepaalde als onbepaalde entiteiten. We zien vaak plaatsnamen in het schema *in _ staat* (Oostermeer, Arnhem, Deventer, Hoogeveen etc) maar ook onbepaalde plaatsen zoals *kasteel, onderaardse* of *gehucht*. De genoemde personages hebben meestal geen naam en behoren tot abstracte, onbepaalde entiteiten (*heks, weerwolf, tovenaar, molenaar, kabouter*). Verhalen worden vaak in een onbepaald punt in de tijd geplaatst, bijvoorbeeld met het generieke schema *was _ een* zoals in 'er was eens een koning' en 'te Scheveningen was een bad'. Soms wordt er een expliciet beroep gedaan op de geloofwaardigheid van het verhaal in de opening zoals in 'volgens de buren', 'volgens de overlevering', 'volgens de buurtbewoners' en 'volgens de verhalen'.

Tot slot de beste kenmerken van sprookjes:

(12) *Anansi, sprookje, sprookjes, Kantjil, daar was, Keutje, van _ er, een sprookje, van den, Kancil, jan had, ga, op _ ga, leefden, reis, er was eens een, vos, Griet, van _ drie, het verhaaltje, Pieterke, vader had, mannetje, broertje, kaaiman, een kort, de vertelling, heer spin, was eens een, van den _ die, den _ die, was _ winter, het mannetje, Boontje, Kooltje, wolf, den _ met, er was eens, veegden*

Er was eens is een van de frequentste openingen voor sprookjes. En, hoewel het ook geattesteerd is bij andere genres is de frequentie hoog genoeg om discriminerend te zijn. Andere gerelateerde openingen die het verhaal in een onbepaalde tijd en ruimte plaatsen, zijn: *was eens een* en *van van _ die*. Verder zien we een aantal namen van personages die vaak in sprookjes in de verhalenbank voorkomen, zoals

Anansi, Kantjil, (Koffie)Boontje, Kooltje en *Griet.* Een enkele keer wordt expliciet verteld dat de vertelling een sprookje is. Een voorbeeld van een dergelijke contactopening is 'een sprookje is mij bekend ...'.

Ter afsluiting geef ik in onderstaande tabel een overzicht van de verschillende openingscategorieën die gevonden zijn per genre:

Tabel 6 Overzicht van de gevonden openingen per genre geclassificeerd naar communicatieve functie voor Nederlandse verhalen.

	contact	*(in)credibilis*	*ruimte en tijd*
Broodjeaapverhaal	X	credibilis	bepaald
Kwispel	X	–	onbepaald
Legende	–	–	bepaald
Mop	X	–	onbepaald
persoonlijke narratief	–	credibilis	bepaald
Raadsel	X	–	–
Sage	–	credibilis	onbepaald / bepaald
Sprookje	X	incredibilis	onbepaald

4.2.2 Fries

Evenals voor het Nederlands bevatten de openingen van Friese moppen personages die vaak in moppen voorkomen. Dokkummers en de plaats Dokkum zijn populair om grappen over te maken (in de verhalenbank), dominees spelen vaak een rol evenals poepen (Buben, Duitse seizoensarbeiders). Een groot verschil met de Nederlandse moppen is dat de Friese moppen niet of nauwelijks met een werkwoord op de eerste positie beginnen. Een frequent gebruikt schema is der wie _ in _, als in *der wie in boerefaem dy* en *der wie ris in jager.* Dergelijke openingen zijn onbepaald wat betreft tijd en ruimte en behoren tot de corresponderende openingscategorie. De resultaten maken ook enkele oneffenheden in de verhalenbank zichtbaar. Grappende raadsels beginnend met een vragend voornaamwoord zijn voor het Nederlands ondergebracht bij het genre raadsels, terwijl ze voor het Fries bij de moppen zijn geplaatst (bijvoorbeeld *hwat is* _). Dit zijn de beste kenmerken:

(13) *doomny, ulespegel, sei _ de, poepen, Dokkum, poep, meester, dominy, no, pastoar, Dokkumers, Mücnchhausen, Hitler, hwat is, de Dokkumers, Feitse, un, dat die, teltsje fan, yn _ skoalle, yn oarlochstiid, thyl, hoe, twa froulju, joaden, Amsterdam, in poep, dat woarde, omauke, mieren, in _ Jan, Jan en _, hy _ it, yn Amsterdam, in dominy, der wie _ in _, in _ en _, hwatte, symen, moos, sille, de rover, april, de faem, oarloch, Koudum, boer en _*

Friese sagen bevatten net als Nederlandse sagen veel generieke personages, zoals *rovers, feinten, arbeiders* en *reuzen*:

(14) *der wie _ yn, der _ in yn, earne op, feinten, reuzen, der wie in _ yn, arbeiders, ek _ in, der wienen _ reuzen, twa reuzen, der _ twa man, der waar, nachtmerje, wie in _ yn, koe, wienen man, har, der wie froeger, der _ ien, de _ sei, woonde, skodde, der wienen _ twa man, Harkema, in Tsjoenster, in nachtmerje, ut in, mar _ gong, Tsjoensters, der*

kommen, doe hie, mat, der _ hwat, doarpen, guon, ergens _ stie, dy man, bruorren, men, ergens yn, der kom _ ris _, der wienen _ op

Veel van de voorspellers vallen onder de categorie van 'ruimte en tijd', zoals *der wie _ yn, ergens _ stie, der _ in yn, ek _ in, der wienen _ reuzen, der _ twa man* en *ergens yn, der wie froeger.* Naast deze onbepaalde openingen benoemen openingen als *Harkema, In Tsjoenster* etc ruimtelijk bepaalde entiteiten. De betrouwbaarheid van het verhaal wordt soms aangegeven door een familielid op te voeren van wie de verteller het verhaal zou hebben gehoord, zoals *mem fortelde der wie in ....*

Tot slot zijn hier de meest discriminerende kenmerken voor Friese sprookjes:

(15) *foks, bear, raef, liuw, Blauburd, der _ raef, der _ in raef, in raef, _ oege en _, de _ snoade, snoade, de _ foks, Ychelbaerch, de Ychelbaerch, de bear, sa wie, in koaning, winterkoaninkje, it winterkoaninkje, antsjemuoi, der koaning, der _ in koaning, boeredochter, der leefde, de liuw, in beantsje, beantsje, Blauburd in, der _ ris in koaning, ris in koaning, der _ ris koaning, ris _ koaning, moard komt, in moard komt, in ezel, oege, Ruslân, de fûgels, leefde, fokse, fûgels, greate oege, sa wie _ der, wier, en de, in moard*

Ook de Friese sprookjes openen met het noemen van een aantal frequent gebruikte personages in sprookjes, zoals *bear, fokse, liuw, hazze, Blauburd,* en *raef*. Het schema *der _ in*, ingevuld door bijvoorbeeld *der wie in*, is een openingsfrase in de categorie van 'tijd en ruimte' en laat daarmee zowel de tijd als de ruimte onbepaald. Andere soortgelijke schema's zijn: *der _ ris in koaning, sa wie _ der, der _ in raef*.

De volgende tabel geeft een overzicht van de verschillende openingscategorieën gevonden in de Friese verhalen:

Tabel 7 Overzicht van de gevonden openingen per genre geclassificeerd naar communicatieve functie voor Friese verhalen.

	contact	*(in)credibilis*	*ruimte en tijd*
Mop	–	–	onbepaald
Sage	–	credibilis	onbepaald / bepaald
Sprookje	X	incredibilis	onbepaald

5 Besluit

De verschillende genres die te onderscheiden zijn binnen het domein van volksverhalen vertonen grote verschillen in de formules die aangewend worden om een verhaal te openen. In dit artikel heb ik met computationele methoden laten zien dat de openingen van de verschillende genres genoeg discriminerende kenmerken vertonen om met behoorlijke zekerheid zowel Nederlandse als Friese verhalen naar hun genre te kunnen classificeren. Uit de experimenten bleek dat het slot van een verhaal een veel slechtere voorspeller is, wat vanuit communicatief oogpunt weinig verbazingwekkend is.

Ik heb laten zien dat *skip*-grammen zowel voor het Fries als voor het Nederlands betere classificatieresultaten opleveren dan reguliere *n*-grammen. Het pres-

tatieverschil is echter wat teleurstellend gezien de veel krachtigere set van kenmerken die *skip*-grammen voortbrengen. De *skip*-grammen blijven interessant, omdat ze ons naast lexicaal gefixeerde openingsformules ook openingsschema's bieden op verschillende niveaus van abstractie.

Met behulp van een kwalitatieve analyse van de experimentele resultaten heb ik laten zien welke openingen kenmerkend zijn voor de verschillende volksverhaalgenres. Hiermee werden verschillende meer en minder lexicaal gespecificeerde patronen zichtbaar. Volksverhaalgenres tonen een grote verscheidenheid in de manier waarop een verhaal kan beginnen. Op een abstracter niveau, het niveau van communicatieve functies, zijn er echter duidelijke overeenkomsten te herkennen tussen de openingen van een genre. Sommige genres geven de voorkeur aan contactformules (broodjeaapverhalen en kwispels) of zetten in op de betrouwbaarheid van het verhaal (broodjeaapverhalen, persoonlijke narratieven en sagen).

Bibliografie

Bennis 2001 – H. Bennis, '"Lopen twee mannen in de Damstraat. Zegt die Turk tegen die Marokkaan ..." V1: de plaats van het werkwoord in het Lombokse moppencorpus'. In: Theo Meder (red.), *'Er waren een Marokkaan, een Turk en een Nederlander ...' Volkskundige en taalkundige opstellen over het vertellen van moppen in de multiculturele wijk Lombok*. Amsterdam: Stichting beheer IISG, 2001, p. 103-114.

EM 1975 – Kurt Ranke (Begr.), Rolf Wilhelm Brednich, e.a. (Hrsg.), *Enzyklopädie des Märchens. Handwörterbuch zur historischen und vergleichenden Erzählforschung*. Berlin: De Gruyter, 1975.

Meder 2010 – T. Meder, 'From a Dutch Folktale Database Towards an International Folktale Database'. In: *Fabula 51* (2010) 1–2, p. 6-22.

Meder & Burger 2006 – T. Meder & P. Burger, '"A rope breaks. A bell chimes. A man dies." The kwispel: a neglected international narrative riddle genre'. In: P. Catteeuw, M. Jacobs & S. Rieuwerts [e.a.] [red.], *Toplore. Stories and Songs*. Trier: WVT-Verlag, 2006 p. 28-38.

Nguyen e.a. 2012 – D. Nguyen, D. Trieschnigg, T. Meder & M. Theune, *Automatic classification of folk narrative genres* First International Workshop on Language Technology for Historical Text(s) at KONVENS, 2012, p. 379-382.

Pedregosa e.a. 2011 – F. Pedregosa, G. Varoquaux, A. Gramfort e.a., 'Scikit-learn: Machine Learning in Python'. In: *JMLR* 12 (2011), p. 2825-2830.

Van Rijsbergen 1979 – C. Van Rijsbergen, *Information Retrieval*. London/Boston (Butterworth), 1979. 2e editie.

Adres van de auteur

Meertens Instituut
Joan Muyskenweg 25
1096 CJ Amsterdam
folgert.karsdorp@meertens.knaw.nl