Measuring Dialect Pronunciation Differences using Levenshtein Distance

Heeringa, Wilbert Jan

Samenvatting

Inleiding

Volgens het bewustzijn van dialectsprekers bestaan er dialectgrenzen in het dialectlandschap. Dit blijkt uit de dialectkaart van Daan, waarop grenzen zijn getekend op basis van het dialectbewustzijn van de sprekers. Het dialectlandschap kan echter ook als een continuüm beschouwd worden. Wanneer we langs een rechte lijn reizen van dorp naar dorp, bemerken we slechts geleidelijke veranderingen. Om dialectgrenzen en dialectcontinua te verkennen op elke niveau van gedetailleerdheid, hebben we een 'liniaal' nodig waarmee de taalkundige afstand voor een willekeurig dialectpaar op een objectieve manier gevonden kan worden.

De eerste die een methode ontwikkelde voor het meten van dialectafstanden was Jean Séguy. Hij berekende de afstand tussen twee dialecten als het aantal keren dat de twee dialecten voor een bepaald item verschilden. Het aantal verschillende items werd uitgedrukt in een percentage. Een vergelijkbare aanpak werd ook toegepast door Hans Goebl.

De gebroeders Hoppenbrouwers introduceerden in 1988 twee frequentie-gebaseerde methoden waarmee dialectafstanden gevonden kunnen worden op basis van fonetische teksten. Bij de eerste methode worden per tekst de frequenties van de klanken bepaald en die frequenties worden gedeeld door het totale aantal klanken in de tekst. De afstand tussen twee variëteiten is gelijk aan de som van de frequentieverschillen. Bij de tweede methode worden frequenties van features (kenmerken van klanken) bepaald. De afstand tussen twee variëteiten is in het eenvoudigste geval opnieuw gelijk aan de som van de frequentieverschillen. Beide methoden duiden we aan als varianten van de *corpus-frequentie-methode*.

De beide frequentie-gebaseerde methoden onderscheiden geen woorden in de tekst. Dit kan opgelost worden door foon- of featurefrequenties per woord te bepalen. De afstand tussen twee woorduitspraken, corresponderend met twee dialecten, is opnieuw gelijk aan de som van de frequentieverschillen. De afstand tussen twee dialecten is gelijk aan de som van de woordafstanden. We noemen deze aanpak de *frequentie-per-woord-methode*.

In 1995 gebruikte Kessler de *Levenshtein afstand* voor het bepalen van taalkundige afstanden tussen dialecten. Met deze afstandsmaat wordt de afstand tussen twee woorduitspraken bepaald door de kosten te bepalen van de minimaal vereiste

verzameling van toevoegingen, verwijderingen en vervangingen die nodig is om de ene uitspraak te veranderen in de andere. Kessler paste de afstandsmaat toe op Ierse dialecten. Dit bleek succesvol. Wij gebruikten deze afstandsmaat eveneens omdat de methode objectief is, graduele woordafstanden berekent, woorden als taalkundige eenheden verwerkt, en de volgorde van klanken in een woord in beschouwing neemt. De Levenshtein-afstand staat centraal in dit proefschrift.

Het meten van segmentafstanden

Als we taalvariëteiten op basis van woordtranscripties willen vergelijken, moeten vooraf de afstanden tussen de segmenten bekend zijn. Deze afstanden zijn afhankelijk van de manier waarop spraaksegmenten zijn gerepresenteerd. We onderzochten de foonrepresentatie, de featurerepresentatie en de akoestische representatie.

In het eenvoudigste geval is een spraaksegment of foon niet verder gedefinieerd: twee fonen zijn gelijk of verschillend. Nadeel van de foonrepresentatie is dat bijvoorbeeld de afstand tussen de [ɪ] en de [e] even groot is als de afstand tussen de [ɪ] en de [ɒ]. Dit probleem wordt opgelost door klanken te representeren door een reeks van onderscheidende kenmerken oftwel features. Featurewaarden representeren de mate waarin een feature geldig is. Bijvoorbeeld een feature *lang* is 0 voor een korte klank, 0.5 voor een halflange klank en 1 voor een lange klank.

We experimenteerden met drie featuresystemen. Het eerste werd in 1988 ontwikkeld door de gebroeders Hoppenbrouwers (H & H). Het betreft een articulatie-gebaseerd systeem dat de auteurs gebruikten voor het vergelijken van dialecten in het Nederlandse dialectgebied. Het tweede systeem is gebaseerd op twee andere systemen. Het ene systeem werd ontwikkeld door Vieregge in 1984. Vieregge ontwikkelde zijn systeem voor de controle van de kwaliteit van transcripties. Dit systeem is gedeeltelijk gebaseerd op metingen van perceptieve klankafstanden. Het andere systeem werd ontwikkeld door Cucchiarini 1993. Het systeem van Cucchiarini is een aangepaste versie van het systeem van Vieregge. We definieerden de klinkers in de lijn van Vieregge, en de medeklinkers in de lijn van Cucchiarini. Het derde systeem in ons onderzoek is ontwikkeld door Almeida en Braun in 1986 (A & B). Evenals het tweede systeem is ook dit systeem bedoeld voor de controle van de kwaliteit van transcripties. In het systeem worden op een heel directe manier de afstanden afgeleid uit het IPA-systeem.

Featuresystemen zijn vaak niet gebaseerd op fysische metingen. Alleen het systeem van V & C is gedeeltelijk gebaseerd op afstanden die gemeten werden in een perceptieëxperiment. We hebben daarom ook klankafstanden gemeten op basis van akoestische representaties van samples van de IPA klanken. We gebruikten samples van de geluidsband *The Sounds of the International Phonetic Alphabet* waarop alle IPA klanken uitgesproken worden door twee sprekers.

De klinkers werden geïsoleerd uitgesproken, en de medeklinkers knipten we uit de context waarin ze werden uitgesproken.

We experimenteerden met twee spectrogram-gebaseerde representaties en met een representatie door formantsporen. Een spectrogram is een grafiek waarin de frequentie gerepresenteerd wordt door de x-as en de tijd door de y-as, en waarin de grijswaarde voor ieder punt in de grafiek de intensiteit representeert. We gebruikten niet de standaard-spectrogrammen, maar meer perceptief gemotiveerde modellen: het Barkfilter en het cochleagram. Essentieel voor de waarneming van klinkers is dat spectrale pieken door het oor worden herkend. Hetzelfde geldt voor sonorante medeklinkers. Deze pieken heten *formanten*, en een reeks van formanten in het verloop van de tijd heet een *formantspoor*. We experimenteerden ook met de formantsporenrepresentatie.

Zowel op basis van feature-representaties als op basis van akoestische representaties berekenden we de segmentafstanden. Omdat in onze perceptie kleine klankverschillen soms een relatief sterke rol spelen ten opzichte van grote klankverschillen, experimenteerden we ook met een aanpak waarbij de logaritmen van de klankafstanden gebruikt werden. Omdat de logaritme van 1 gelijk is aan 0, berekenden we die als: $ln(\text{afstand} + 1)$.

Het meten van dialectafstanden

Wanneer de afstanden tussen spraaksegmenten vastgesteld zijn, kunnen we de afstanden tussen woorduitspraken bepalen en vervolgens de afstanden tussen taalvariëteiten. We bepaalden de afstand tussen een woorduitspraak uit de ene variëteit en de corresponderende woorduitspraak uit de andere variëteit met de Levenshtein-afstand. Dit algoritme bepaalt hoe zo eenvoudig mogelijk het ene woord kan worden veranderd in het andere woord door klanken toe te voegen, te verwijderen of te vervangen. Aan de operaties worden gewichten toegekend. In de eenvoudigste vorm van het algoritme hebben alle operaties hetzelfde gewicht, bijvoorbeeld 1. We illustreren het gebruik van de gewichten met een voorbeeld. Het woord *konijn* wordt uitgesproken als [kənɛːn] in het dialect van Amsterdam, en als [kniːnə] in het dialect van Zwollekerspel.[1] Het veranderen van de ene variant in de andere gaat als volgt:

kənɛːn	verwijder ə	1
knɛːn	vervang ɛː door iː	1
kniːn	voeg toe ə	1
kniːnə		
		3

[1] De woorduitspraken werden opgenomen en transcribeerd in 2000 door Renée van Bezooijen, Katholieke Universiteit Nijmegen.

Voor de bepaling van deze afstand met het Levenshtein-algoritme worden beide woorden onder elkaar gezet, waarbij een keuze gemaakt wordt welke segmenten uit de ene variant corresponderen met welke segmenten uit de andere variant. Met andere woorden: de varianten worden *opgelijnd*. De kracht van het Levenshtein-algoritme is nu dat dit algoritme de woordafstand altijd berekent op basis van de oplijning waarin klankcorrespondenties *zodanig* zijn gekozen, dat de som van de operaties minimaal is. In ons voorbeeld ziet de oplijning er als volgt uit:

k	ə	n	ɛː	n	
k		n	iː	n	ə
0	1	0	1	0	1

Wanneer woorduitspraken op deze manier met elkaar vergeleken worden, zal de afstand tussen langere woorden gemiddeld genomen groter zijn dan de afstand tussen kortere woorden. Hoe langer een woord is, hoe groter de kans dat er verschillen zijn ten opzichte van het corresponderende woord in een andere taalvariëteit. Omdat dit niet overeenstemt met het idee dat een woord een taalkundige eenheid is, ongeacht het aantal segmenten waaruit het bestaat, wordt de Levenshtein-afstand gedeeld door de lengte van de oplijning (de gecombineerde woordlengte). In ons voorbeeld is deze gelijk aan 6. De woordafstand, genormaliseerd over de lengte, is nu gelijk aan $3/6 = 0.5$.

Bij gebruik van de foonrepresentatie zijn de gewichten van de operaties gelijk aan 1. Gebruiken we echter een featurerepresentatie of een akoestische representatie, dan zullen de gewichten gradueel variëren.

Op basis van de gemiddelde Levenshtein-afstanden tussen variëteiten kunnen de variëteiten geclassificeerd worden. We maakten gebruik van *cluster-analyse* en *multidimensionale schaling*, twee technieken die elkaar aanvullen. Het resultaat van cluster-analyse is een dendrogram, een boom waarin de variëteiten de bladeren zijn. Het resultaat van multidimensionale schaling is een plot waarop sterk verwante verwante variëteiten dicht bij elkaar zijn geplaatst, en sterk verschillende variëteiten juist ver uit elkaar. We schaalden zowel naar twee als naar drie dimensies. In de plot worden de eerste en tweede dimensie gerepresenteerd door respectievelijk de x-as en de y-as, en de derde dimensie door de grijswaarde van de stippen.

Validatie

In een validatie-onderzoek vergeleken we de corpus-frequentie-methode, de frequentie-per-woord-methode en de Levenshtein-afstand met elkaar. Voor elk van de drie methoden testten we de verschillende segmentrepresentaties: de foonrepresentatie en de featurerepresentatie. Voor de Levenshtein-afstand testten we

ook de akoestische segment-representaties. Verder werden voor de Levenshtein-afstand zowel lineaire als logaritmische segment-afstanden in beschouwing genomen.

Het validatie-onderzoek voerden we uit op basis van 15 Noorse dialecten. Digitale opnamen en transcripties werden gemaakt door Jørn Almberg. De opnamen bestaan uit vertalingen van de fabel 'De noordenwind en de zon'. De tekst bestond (gewoonlijk) uit 58 woorden.[2] Op basis van de opnamen voerde Charlotte Gooskens een perceptie-experiment uit in de lente van 2000. In elk van de 15 plaatsen beluisterde een groep leerlingen op de middelbare school een band met daarop de opnames van alle 15 dialecten. Voor elk van de dialecten moesten de leerlingen op een schaal van 1 tot en met 10 de mate van verwantschap met hun eigen dialect geven, waarbij 1=gelijk aan eigen dialect en 10=ongelijk aan eigen dialect. De gemiddelde scores van de leerlingen in een plaats geven de afstanden van de 15 dialecten op de band ten opzichte van het dialect in die plaats. Omdat het experiment in elk van de 15 plaatsen werd uitgevoerd, kregen we een afstandenmatrix van 15×15 afstanden. We correleerden de resultaten van onze methoden (of varianten daarvan) met deze perceptieve afstanden. Hoe hoger de correlatie, hoe beter de methode de perceptie benadert.

De methoden op basis van de foonrepresentatie bleken het sterkste te correleren met de perceptieve afstanden, direct gevolgd door de Levenshtein-afstanden op basis van de akoestische segmentrepresentatie. De methoden op basis van featuresystemen waren beduidend slechter. Bekijken we resultaten per representatie, dan zien we zowel bij de foonrepresentatie als bij de featurerepresentatie dat de frequentie-per-woord-methode even goed is als, of beter is dan de corpusfrequentie-methode, de Levenshtein-afstand altijd beter is dan de frequentie-per-woord-methode, en de Levenshein-afstand op basis van logaritmische segmentafstanden even goed is als, of beter is dan de Levenshtein-afstand op basis van lineaire segmentafstanden. Dit is ook wat we op methodologische gronden verwachtten. Het feit dat de foonrepresentatie erg goed werkt (voor alle drie methoden), en dat logaritmische segmentafstanden vaak betere resultaten geven dan lineaire segmentafstanden, lijkt erop te wijzen dat het in de perceptie vooral belangrijk is *dat* twee segmenten verschillend zijn, en dat *de mate* waarin ze van elkaar verschillen veel minder belangrijk is. Van de drie akoestische segmentrepresenaties blijkt het Barkfilter beter te zijn dan de twee andere representaties wanneer lineaire segmentafstanden worden gebruikt. Bij gebruik van logaritmische afstanden is er geen verschil.

Resulaten

Hoewel de Levenshtein-afstand op basis van de foonrepresentatie iets sterker correleerde dan de Levenshtein-afstand op basis van de logaritmische akoes-

[2]De opnamen en transcripties zijn gratis beschikbaar via `http://www.ling.hf.ntnu.no`.

tische segmentafstanden, genereerden we de resultaten toch met de variant van de Levenshtein-afstand die gebruik maakt van logaritmische akoestische segmentafstanden. Voor een kleine gegevensverzameling van 15 dialecten werkt de foonrepresentatie-gebaseerde aanpak weliswaar goed, maar bij gebruik van een dichter net van plaatsen kunnen kleinere verschillen een sterkere rol spelen. Met de akoestische maat worden die verschillen in sterkere mate verwerkt. We pasten de afstandsmaat toe op dialecten in het Noorse en het Nederlandse dialectgebied.

De 15 Noorse dialecten van het perceptie-onderzoek maken deel uit van een grotere gegevensverzameling. We berekenden afstanden tussen 55 Noorse dialecten. Afgezien van enkele taaleilanden kregen we op basis van cluster-analyse een hoofdindeling bestaande uit zes groepen: noord, centraal, west, oost, zuidwest en zuidoost. Op basis van multidimensionale schaling kregen we een indeling bestaande uit ruwweg 5 groepen: noord, centraal, west, oost, zuid. De laatste indeling komt iets beter overeen met de traditionele indeling van Skjekkeland. Verschillen kunnen verklaard worden door beperkingen van onze woordenlijst enerzijds, en de keuze van de isoglossen door Skjekkeland anderzijds.

De *Reeks Nederlandse Dialectatlassen* werd samengesteld in de periode 1925–1982 door E. Blancquaert and W. Pée. Van de 1956 beschikbare dialecttranscripties kozen we er 360. We berekenden de afstanden tussen de dialecten op basis van 125 woorden. Met cluster-analyse kregen we een indeling in Fries, Nedersaksisch, Nederfrankisch en Limburgs. We onderzochten elk van de vier groepen meer gedetailleerd en vergeleken de indeling met de kaart van Daan. Verschillen konden soms verklaard worden uit notatieverschillen van de verschillende transcribenten in de RND, en een enkele keer uit een (vermoedelijke) tekortkoming van de kaart van Daan. Op basis van multidimensionale schaling kregen we de klassieke indeling in Fries, Nedersaksisch en Nederfrankisch. We vergeleken de dialecten ook ten opzichte van het Standaard Nederlands. Het dialect van Haarlem bleek het sterkst verwant, en de Friese variëteiten bleken het meest afwijkend.

Conclusie

In dit proefschrift ontwikkelden we verschillende varianten van de Levenshtein-afstand en onderzochten of deze afstandsmaat bruikbaar is voor het berekenen van afstanden tussen taalvariëteiten. Uit validatie-onderzoek bleek dat de Levenshtein-afstand betere resultaten geeft dan de corpus-frequentie-methode en de frequentie-per-woord-methode. Ook bij toepassing van de Levenshtein-afstand op Noorse en Nederlandse gegevens bleek de methode een geschikt gereedschap voor het vinden van afstanden tussen taalvariëteiten.