Advancing explanatory and tonal dialectometry

Sung, H.W.M.

Samenvatting

Dialectometrie is een kwantitatieve tak van de dialectologie die gebruikmaakt van computationele en statistische methoden, toegepast op dialectdata, om ruimtelijke taalvariatie te begrijpen. Dit proefschrift beschrijft hoe dialectometrie ons begrip van de variatie van Yue-dialecten in Zuid-China kan verdiepen, en hoe Yue ons kan helpen de reikwijdte van computationele methoden die in dialectometrie worden gebruikt te verbreden, om tonale talen te verklaren die wereldwijd veel voorkomen, maar niet zo vaak als onderwerp binnen de dialectometrie worden bestudeerd.

In dit proefschrift worden een aantal onderzoeksvragen behandeld, die onder de volgende thema's vallen: 1) segmentale classificatie van Yue-dialecten, 2) identificatie van karakteristieke kenmerken van Yue-dialecten, 3) tonale classificatie van Yue-dialecten en 4) vergelijking tussen segmentale en tonale variatie van Yue-dialecten.

De dataset die in dit proefschrift wordt gebruikt, bestaat uit de IPA-transcriptie van ongeveer 130 woorden in 113 Yue-Pinghua-dialecten. Voor de segmentale classificatie werd de Levenshtein-afstand gebruikt om de fonetische afstanden te berekenen, gevolgd door multidimensionale schaling als techniek voor dimensiereductie en clusteranalyse om de interne structuur van het Yue-Pinghua-dialectlandschap te onderzoeken. Traditioneel Noordelijk Pinghua bleek buiten het Yue-continuüm te vallen, maar dat geldt niet voor traditioneel Zuidelijk Pinghua. Vervolgens werd een diepere analyse uitgevoerd met de resterende 104 dialecten (waarbij Noordelijk Pinghua als uitschieters werd verwijderd). Yue-dialecten blijken zich op segmentaal niveau op een groot continuüm te bevinden en de dialecten kunnen worden onderverdeeld in 2 tot

5 grote groepen, afhankelijk van het gewenste detailniveau.

Een dialectometrische classificatie wordt vaak bekritiseerd vanwege het gebrek aan details of uitleg over de geïdentificeerde dialectgroepen. Dit komt doordat classificaties gebaseerd waren op afstanden, en het moeilijk is om de kwalitatieve dialectkenmerken te achterhalen na de kwantificering in afstanden. Om deze reden werd Multiple Sequence Alignment (MSA) gebruikt voordat de dialectafstanden voor de segmentclassificatie van Yue werden berekend. MSA splitste de fonetische transcripties op in historisch verwante segmenten, en dit werd gelijktijdig voor alle dialecten gedaan. De transformatie van de transcriptiegegevens maakt klanksegmenten 1) historisch gezien nauwkeuriger uitgelijnd en 2) bruikbaarder voor post-hoc analyses, zoals automatische kenmerkextractie. Met behulp van genormaliseerde Pointwise Mutual Information, een associatiemaat die veel wordt gebruikt in natuurlijke taalverwerking, werden karakteristieke kenmerken die nauw verbonden zijn met elke dialectgroepen (geïdentificeerd met behulp van clusteranalyse) vastgesteld. Deze techniek heeft het verklarende aspect van dialectometrie vergroot, dat verder gaat dan alleen dialectclassificatie.

Aan de andere kant worden toontalen, ondanks dat ze vrij algemeen zijn binnen de talen van de wereld, niet veel bestudeerd met behulp van dialectometrische technieken. Een probleem is dat het niet meteen duidelijk is hoe toonafstanden gemeten kunnen worden. Er zijn verschillende benaderingen geweest, van simpelweg binaire overeenkomsten tot geavanceerde, op perceptie gebaseerde methoden voor afstandsberekening. In dit proefschrift wordt de Levenshtein-afstand toegepast op een representatie van tonen, hierna gemodificeerde Onset-Contour-Offset (gOCO), om dialectafstanden te verkrijgen en te onderzoeken hoe tonen in de ruimte variëren tussen verschillende dialecten. gOCO is een geschiktere toonrepresentatie voor dialectometrie vergeleken met andere, omdat het 72 van de 73 tonen in de dataset kan onderscheiden, met geleidelijke afstanden van de ene toon tot de andere, wat overeenkomt met de menselijke perceptie. Bij toepassing van gOCO op een dialectometrische analyse van Yue-dialecten op toonniveau verschilt het variatiepatroon van segmenten. Terwijl segmentale variatie een continuümpatroon vertoont, vertoont tonale variatie een meer categorisch dialectgebiedspatroon. Verdere analyse heeft aangetoond dat niet alle dialectgroepen hetzelfde vertonen in segmenten en tonen. Deze discrepanties zijn intrigerend en roepen verdere onderzoeksvragen op die wachten op verder onderzoek.