Het Asta-project: automatische spraakherkenners voor Nederlandse dialecten

Martijn Bentum, *CLS/CLST, Radboud Universiteit*
Eric Sanders, *CLS/CLST, Radboud Universiteit*
Antal van den Bosch, *Universiteit Utrecht*
Henk van den Heuvel, *CLS/CLST, Radboud Universiteit*

Het Meertens instituut heeft in de tweede helft van de 20ste eeuw in heel Nederland opnames gemaakt van verschillende dialecten en ongeveer driehonderd uur is handmatig getranscribeerd. Dit materiaal lijkt de ideale basis voor het ontwikkelen van *dialect-specifieke spraakherkenners*, maar er zijn ook aanzienlijke uitdagingen bij het verwerken van deze data.

De handmatige transcripties zijn gekoppeld aan de spraakopnames via een metadata-bestand. Helaas zijn de transcripties niet opgelijnd met de audio waardoor het onduidelijk is wanneer er wat gezegd wordt. Verder zijn de transcripties uitgeschreven in semi-conventionele spelling die is aangepast om de uitspraak in het dialect weer te geven. Hierdoor zijn de handmatige transcripties helaas niet altijd consistent en zijn ze moeilijker te koppelen aan automatische transcripties voor oplijning tussen audio en transcriptie.

Om een eerste oplijningsbenadering te maken hebben we gebruik gemaakt van automatische spraakherkenning door een Nederlands Wav2vec2 model in combinatie met het Needleman-Wunch algoritme. Dit algoritme benadert een optimale oplijning tussen twee sequenties in dit geval tussen de handmatige en automatische transcripties. Deze oplijning wordt gecontroleerd met een hiervoor ontwikkeld webgebaseerde annotatietool. Met de resultaten van deze annotatie kan de spraakherkenner verbeterd worden voor specifieke dialecten.

Tot nu toe kan geconcludeerd worden dat het oplijnen van de oorspronkelijke handmatige dialecttranscripties met behulp van een standaard Nederlands Wav2vec2 model en het Needleman-Wunsch algoritme goed werkt, maar dat er handmatige filtering van de data nodig is om het materiaal geschikt te maken voor het trainen van dialect-specifieke spraakherkenners.

ASTA is een subproject van Werkpakket 3 (“Linguistics”) van het CLARIAH-PLUS Grootschalige Wetenschappelijke Infrastructuurproject, en wordt gefinancierd door NWO (projectnummer 184.034.023).