ONTWERP AUTONOMATA GESPROKEN POI-CORPUS

Marijn Schraagen

Inleiding

Het Autonomata TOO gesproken POI-corpus maakt deel uit van het Nederlands-Vlaamse STEVIN-project "Autonomata TOO: Transfer of Output". Het doel van dit project is het verbeteren van multilinguale spraakherkenning. Voor evaluatiedoeleinden is een nieuw spraakcorpus ontwikkeld. Dit corpus bevat sprekers met verschillende taalachtergrond, en ook het lexicon is opgebouwd uit verschillende talen. Het Autonomata TOO-project is een vervolg op het Autonomata-project, waarin het Autonomata Spoken Name Corpus (ASNC) is ontwikkeld. Het ASNC is een spraakcorpus met namen van personen en plaatsen. Het nieuwe corpus heeft een aan het ASNC gerelateerd lexicaal domein, namelijk Points-Of-Interest (POI's) uit Nederland en België. De POI's zijn geselecteerd uit de database van TeleAtlas, een van de projectpartners. Het lexicon bevat namen van restaurants, hotels, campings, cafés en dergelijke.

Het ontwerp en de realisatie van het gesproken POI-corpus bestaat uit vijf fasen:
1. opstellen van het profiel van de sprekers
2. opstellen van de POI-lijst
3. contacteren van de sprekers
4. opnames maken
5. transcripties

Fase 1: opstellen van het profiel van de sprekers

Bij de recrutering van sprekers wordt rekening gehouden met de volgende factoren: moedertaal, geslacht, leeftijd, dialectgebied, opleiding. Hierbij wordt onderscheid gemaakt tussen vereiste factoren en wenselijke factoren. Vereiste factoren zijn moedertaal en geslacht, de sprekers worden volgens een strikte verdeling van deze factoren gekozen. Wenselijke factoren (overige) worden betrokken in de verdeling van de sprekers voor zover dit praktisch haalbaar is. Naast de selectiecriteria wordt ook enige aanvullende informatie over de sprekers opgeslagen. Het gaat om huidige woonplaats, thuistaal (moedertaalsprekers Nederlands), verblijfsduur in het Nederlandse taalgebied en taalniveau (moedertaalsprekers vreemde talen), talenkennis (alle sprekers).

Moedertaal

Het corpus bevat opnames van sprekers met als moedertaal Nederlands, Engels, Frans, Turks, en Marokkaans Arabisch/Berbers. De groep 'Nederlands' is verdeeld in sprekers uit Vlaanderen en Nederland. De groep 'Engels' bevat sprekers uit de Verenigde Staten, Canada, Australië, Groot-Brittannië en Hong Kong. De overige sprekers zijn afkomstig uit Frankrijk, Turkije en Marokko, respectievelijk. Het onderstaande overzicht bevat de verdeling van sprekers in aantallen. Moedertaalsprekers van het Nederlands zullen in het vervolg worden aangeduid als 'Nederlandse sprekers'. Dit verwijst naar de moedertaal, niet naar het land van herkomst. Moedertaalsprekers van vreemde talen worden 'buitenlandse sprekers' genoemd. Ook dit is een verwijzing naar de moedertaal, en niet naar de nationaliteit.

Moedertaal	**Aantal sprekers**
Nederlands (Nederland)	20
Nederlands (Vlaanderen)	20
Engels	10
Frans	10
Turks	10
Marokkaans Arabisch/Berbers	10
Totaal	*80*

Geslacht
De sprekers zijn gelijk verdeeld naar geslacht: 40 mannen en 40 vrouwen.

Leeftijd
Voor leeftijd zijn twee categorieën onderscheiden: jonger dan 40 jaar en ouder dan 40 jaar. De Nederlandse sprekers zijn per regio (Nederland/Vlaanderen) gelijk verdeeld over deze groepen. Voor de buitenlandse sprekers is om praktische redenen de verdeling naar leeftijd minder strikt gehanteerd.

Dialectgebied
Het dialectgebied is alleen van toepassing voor Nederlandse sprekers. Bepalend voor de indeling is het gebied waar de spreker tussen zijn derde en achttiende jaar het langst gewoond heeft. De dialectgebieden zijn gelijkgesteld aan de provincies van Nederland en Vlaanderen. Voor Nederland betreft dit Groningen, Friesland, Drenthe, Overijssel, Gelderland, Utrecht, Noord-Holland, Zuid-Holland, Zeeland, Noord-Brabant en Limburg. Voor Vlaanderen gaat het om West-Vlaanderen, Oost-Vlaanderen, Antwerpen, Vlaams Brabant en Limburg. Voor Vlaanderen zijn de sprekers gelijk verdeeld over alle provincies. Voor Nederland is deze verdeling minder strikt gehanteerd, wel is er gestreefd naar een zo goed mogelijke verdeling.

Opleiding
Het opleidingsniveau van de sprekers is verdeeld in twee categorieën. De eerste categorie betreft het hoger beroepsonderwijs (Nederland) en universitair onderwijs. De tweede categorie betreft al het overige onderwijs. Deze factor is niet strikt gehanteerd.

Thuistaal
Ook thuistaal is alleen van toepassing voor Nederlandse sprekers. Er zijn drie categorieën: standaard-Nederlands, dialect, of een tussenvorm. De definitie van elk van de categorieën is overgelaten aan de sprekers; er zijn geen criteria vastgesteld voor de grenzen tussen categorieën.

Verblijfsduur in het Nederlandse taalgebied en taalniveau
Voor buitenlandse sprekers is de verblijfsduur in het Nederlandse taalgebied vastgelegd. Ook het taalniveau is opgenomen. Indien beschikbaar is hiervoor het CEF-niveau aangehouden. Bij sprekers zonder CEF-niveau is scholing in de vorm van (niet-CEF) taalcursussen genoteerd.

Talenkennis
Van alle sprekers is de talenkennis opgenomen in het corpus. Dit is naar eigen beoordeling van de sprekers. Er zijn drie categorieën: matig, redelijk en goed.

Fase 2: opstellen van de POI-lijst

Iedere spreker dient een lijst van 200 POI's uit te spreken. De POI's zijn afkomstig uit de database van TeleAtlas, en betreffen werkelijk bestaande Points-of-Interest in Nederland en België. Een selectie is gemaakt uit twee categorieën: Hotel-Motel-Camping en Restaurant-Café-Nachtleven. De POI's zijn niet verdeeld naar type, wel is een verdeling gemaakt naar taalachtergrond. Het betreft Nederlandse, Engelse en Franse namen, en namen met een combinatie van talen (Nederlands en Engels, Nederlands en Frans). In het onderstaande overzicht staan de aantallen voor alle categorieën vermeld.

Alle buitenlandse sprekers spreken dezelfde lijst uit. Deze lijst bevat de Nederlandse POI's en de POI's met een taalcombinatie. Vanwege de richting van het onderzoek in Autonomata TOO was het niet nodig ook de Engelse en Franse POI's in deze lijst op te nemen.

De Nederlandse sprekers zijn verdeeld tussen vier lijsten. De lijsten bevatten elk een unieke selectie van Engelse en Franse POI's. Daarnaast bevat elke lijst een mutueel exclusieve selectie van Nederlandse- en combinatie-POI's van de lijst van de buitenlandse sprekers. Het onderstaande overzicht toont deze constructie met de bijbehorende aantallen. De kleur- en lettercodes geven aan welke POI's overeenkomen tussen de verschillende lijsten.

taal sprekers	aantal sprekers	aantal POI's		
		NL 50=30 NL+10 (NL+EN) + 10 (NL+FR)	EN	FR
NL groep 1	10	50 (A)	75	75
NL groep 2	10	50 (B)	75	75
NL groep 3	10	50 (C)	75	75
NL groep 4	10	50 (D)	75	75
niet-NL	40	50 (A) + 50 (B) + 50 (C) + 50 (D)	0	0

Door deze verdeling wordt elke Nederlandse (en combinatie-) POI 50 keer uitgesproken, en elke Engelse/Franse POI 10 keer. De totale lijst bevat dus 800 POI's, waarvan 200 Nederlandse (120 alleen Nederlands en 80 combinaties met Engels of Frans), 300 Franse en 300 Engelse.

Fase 3: contacteren van sprekers

Nederlandse sprekers
Voor Nederlandse sprekers is de via-via-methode geschikt: aanspreken van vrienden, familie en kennissen (via telefoon, mail, enz.), hen laten deelnemen aan de opnames en hen vervolgens vragen naar namen en adressen van andere personen die mogelijks willen deelnemen. Ook kan via universiteiten en verenigingen worden gewerkt. Er bestaan databases en contactwebsites om onderzoekers en potentiële proefpersonen bij elkaar te brengen.

Buitenlandse sprekers
Voor Engelse en Franse sprekers is veel gewerkt via de universitaire opleidingen van de respectievelijke talen (docenten en studenten) en migrantenverenigingen, zoals de International Neighbour Group of het Genootschap Nederland Engeland. Ook hier werkt de via-via-methode goed. Voor Turkse en Marokkaanse sprekers is contact met buurthuizen en goed geïntegreerde leden van de gemeenschap belangrijk. Ook adverteren op algemene plaatsen (bv. marktplaats.nl) kan iets opleveren.

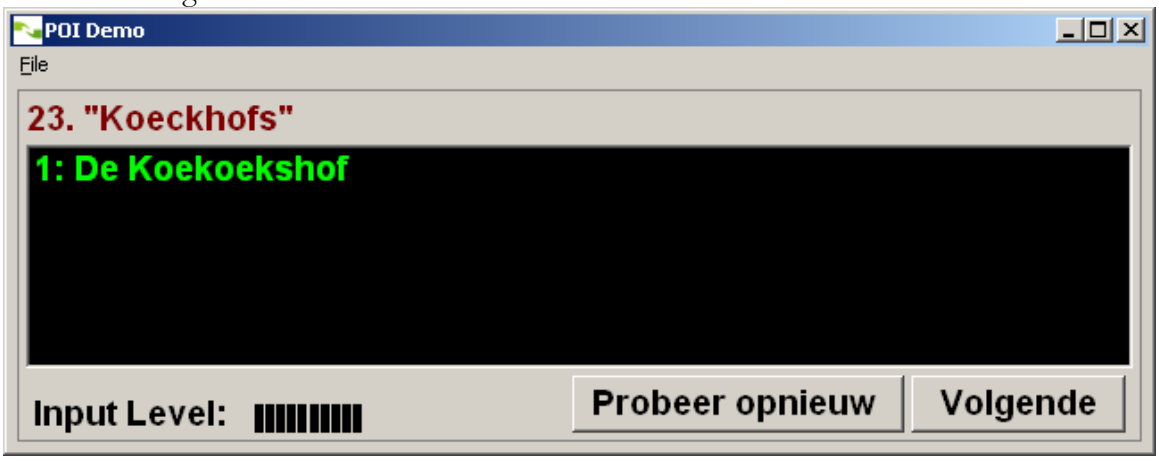
Sprekergegevens
Naast de inhoudelijke informatie die van belang is voor het onderzoek, zoals beschreven in fase 1, worden ook de naam en contactgegevens van de sprekers vastgelegd. Bij opname van gegevens in het corpus wordt uiteraard anonimiteit verzekerd (namen vervangen door sprekercode, geen contactgegevens).

Fase 4: maken van de opnames

Presentatie
De presentatie gebeurt met software die door Nuance is gemaakt voor de dataverzameling van Autonomata TOO. De software is een grafische gebruikersinterface rondom de Nuance VoCon 3200 spraakherkenner, versie 3 0F3. De spraakherkenning gebeurt met behulp van een Nederlandse grafeem-naar-foneem-omzetter (G2P) van Nuance, met als lexicon de complete POI-verzameling van TeleAtlas voor Nederland en België.

De spreker bedient de software zelf. Een POI wordt getoond op het scherm, die de proefpersoon gevraagd wordt uit te spreken. Het resultaat van de herkenning wordt getoond aan de gebruiker. Het systeem beoordeelt of de POI juist is herkend. Bij een fout herkende uiting wordt de proefpersoon gevraagd om de uiting nogmaals uit te spreken. Ook van deze tweede poging wordt het herkenningsresultaat getoond. Dit proces herhaalt zich totdat de uiting juist herkend is, of totdat de spreker besluit door te gaan naar de volgende POI. Het aantal pogingen na de eerste herhaling is een vrije keuze van de spreker. De lijst met POI's wordt bij elke opname in een andere volgorde getoond.

Onderstaand screenshot geeft een voorbeeld van de gebruikersinterface, bij een verkeerd herkende uiting.



De opnameleider zit naast de spreker om te assisteren met de bediening van de software, en om het proces van meerdere pogingen in goede banen te leiden. Eventuele ingrepen zijn erop gericht om de kwaliteit van de opnames te waarborgen (het voorkomen van half uitgesproken uitingen of opzettelijk onjuiste uitspraak), en om zinloze opnames te vermijden (bijvoorbeeld het meer dan 3 keer op dezelfde manier uitspreken van een uiting). De opnameleider beantwoordt geen vragen van sprekers bij twijfel over de uitspraak.

Nabewerking
Na afloop van de sessie worden de opnames gecontroleerd. Bij beschikbaarheid van meerdere opnames voor een POI worden onvolledige opnames en opnames met een bijzonder ernstige verspreking verwijderd. Spraak die niet bij de uiting hoort, zoals een spreker die een opmerking maakt die per ongeluk op de opname terecht is gekomen, wordt uit de uitingen verwijderd voordat deze worden opgenomen in het corpus.

Opnamelokaties
De opnames vinden in principe plaats in de opnamestudio's in Utrecht en Gent. Als het voor proefpersonen niet mogelijk of wenselijk is om naar de studio te komen, worden de opnames op lokatie uitgevoerd. In dit geval wordt aandacht besteedt aan het minimaliseren van achtergrondgeluiden en reverberatie (echo/galm).

Vergoeding van sprekers
Sprekers krijgen een vergoeding van 10 euro in contant geld (in Nederland) of als waardebon (in België). Indien gewenst worden ook reiskosten van sprekers vergoed.

Technische specificaties
De opnames gebeuren met een opnameplatform dat gebruik maakt van een laptop en een USB-headset. De digitale opnames worden rechtstreeks op de harde schijf van de laptop opgeslagen in 16 bit lineaire PCM (wav-formaat). De samplefrequentie is dezelfde voor alle opnames en bedraagt 16 kHz.
De specificaties van de gebruikte microfoon zijn als volgt:

Type: Electret condensator
Richtinggevoeligheid: Unidirectioneel
Frequentiebereik: 40 - 16k Hz
Impedantie: 2.2 kΩ
Signaalgevoeligheid: (1V/Pa - 1kHz): -36 +/-3dB

Testopnames
Alvorens de eigenlijke opnamesessie te beginnen, wordt eerst een aantal proefopnames gemaakt om na te gaan of alle scripts in orde zijn en of de kwaliteit van de opnames voldoet aan de verwachtingen.

Fase 5: transcripties
Van alle geluidsopnames is een handmatig geverifieerde fonetische transcriptie gemaakt. Deze transcripties zijn gemaakt in het LH+ fonetische alfabet. Meer informatie staat in het transcriptieprotocol, dat wordt beschreven in een afzonderlijk document.