Peter Wittenburg, grondlegger van The Language Archive:

‘De vraag van morgen moet je vandaag kunnen beantwoorden’

Eén jaar na lancering van ‘The Language Archive’ spreekt E-data & Research met Peter Wittenburg. Grondlegger van TLA, werkzaam bij het Max Planck Instituut voor Psycholinguïstiek vanaf de start en (op papier) met pensioen. Hij vertelt ons hoe hij de digitale geesteswetenschappen heeft zien opgroeien.

Heidi Berkhout en Erica Renckens

Peter Wittenburg: “Uiteindelijk lijkt het wel of alles makkelijker wordt” *foto Heidi Berkhout*

De methode

“Moet je je voorstellen. 35 jaar geleden wilden we handgebaren tijdens het spreken meten. We installeerden twee infraroodcamera’s en plakten infraroodlampjes op een hand. Samen maakten ze de driedimensionale observatie mogelijk van een bewegende hand. De methode was erg storingsgevoelig. Zo kon je de hand niet draaien, daarmee verbrak je de verbinding tussen licht en camera. Vervolgens had een computer een nacht nodig om de data te verwerken. Tegenwoordig is minder inspanning nodig, zijn er laboratoria gespecialiseerd in deze onderzoeksmethode. Zij gebruiken bijvoorbeeld zestien camera’s in plaats van twee. Uiteindelijk lijkt het wel of alles makkelijker wordt, waardoor meer en betere experimenten mogelijk zijn.”

De analyse

“In 1976, bij de start van het MPI, hadden we maar één PDP-11/55, toentertijd een supersnelle computer. Er werkte hier een Engelse onderzoeker die we ‘het spook’ noemden. Hij deed de zware berekeningen voor experimenten met de infraroodtechnologie ’s nachts, zodat de computer overdag beschikbaar was voor andere werkzaamheden. Als wij ’s ochtends binnenkwamen, ging hij weg, slapen. Dit heeft hij zeker een jaar volgehouden. Wat een pionierswerk.”

De inspiratie

“Deze zomer ontmoette ik een invloedrijke onderzoekster tijdens een workshop. Vol trots vertelde ze dat ze sneller en beter analyses kon uitvoeren op haar onderzoeksdata dankzij de mogelijkheid om in ELAN (een professionele *tool* voor het maken van annotaties bij video- en audio-opnames, red.) te zoeken met reguliere expressies. Er was een wereld voor haar opengegaan. Ik vind dat schattig en grappig. ELAN maakt dit type onderzoek al jaren mogelijk. Maar het duurt altijd vrij lang voor nieuwe technieken zijn doorgedrongen tot kleinere afdelingen op universiteiten. Onderzoekers moeten de tijd en rust vinden om zich erin te verdiepen, zodat ze zich het systeem eigen kunnen maken. De wijze waarop deze dame over haar recente ontdekking vertelde, enthousiasmeert anderen om deze applicatie ook te gebruiken. Zo verspreidt het gebruik zich verder. Zij verovert op dat moment de wereld. En dat doet mij goed.”

Peter Wittenberg

Peter Wittenberg werkt sinds de oprichting in 1976 bij het MPI. Peter was ruim een jaar hoofd van The Language Archive, dit najaar zal een opvolger gekozen worden en zal hij als adviseur bij TLA betrokken blijven. Het doel van TLA is om digitale taalbronnen op te slaan en te behouden, om onderzoekers en andere geïnteresseerde gebruikers daartoe toegang te verlenen en om nieuwe technologieën die taalonderzoek bevorderen te ontwikkelen en te integreren. Naast TLA is Peter actief binnen digitale infrastructuren zoals CLARIN (Common Language Resources and Tools Infrastructure) en EUDAT (European Data Infrastructure).

De wereld anno 2012

“Tegenwoordig beschikken we over betere technologieën waarmee we veel meer data kunnen analyseren. Dit zorgt ervoor dat, met name binnen observationeel onderzoek, een heel ander type onderzoek mogelijk is. Hoe hebben talen zich in de loop der tijd ontwikkeld? Zo’n onderzoeksvraag kunnen we sinds een jaar of vier dankzij de technologie beantwoorden door gebruik te maken van algoritmes uit de erfelijkheidsleer. We kunnen nu veel meer taalkundige kenmerken meenemen in onze kwantitatieve analyse.”

De uitdaging

“Eén van de grootste uitdagingen op dit moment is dat veel video- en audiomateriaal beschikbaar is, maar dat lang niet alles geannoteerd is. Begrijpelijk, er is sprake van een factor 1:35; het kost 35 uur om 1 uur audio handmatig te transcriberen. We hebben 80 terabyte in ons archief, dus het is praktisch onmogelijk om al het materiaal te annoteren. Maar zonder annotaties is geen taalkundig onderzoek mogelijk. De vraag is: hoe is dit materiaal toch te ontsluiten? We werken aan een technologie om in audio- en video-signalen semi-automatisch patronen te vinden. In een bepaalde taal worden bijvoorbeeld meerdere werkwoorden geclusterd. We segmenteren dan een typisch voorbeeld van zo’n clustering om kenmerkende *audiofeatures* eruit te halen. Vervolgens kijkt de software of dit patroon vaker voorkomt in de opname. Het zijn en blijven kansberekeningen die alsnog handmatig gecontroleerd moeten worden, maar onderzoekers kunnen zo veel makkelijker het complete materiaal doorzoeken. Wij ontwikkelen dit in samenwerking met andere gespecialiseerde instituten. Zo combineren we onze kennis over automatische spraakverwerking met kennis over audioverwerking (IAIS Institute Bonn) en kennis over videoverwerking (Heinrich Hertz Institute Berlijn).”

INTERVIEW

De toekomst

“Wat blijft, is dat het MPI alles zo open mogelijk wil maken. Zodat wetenschappers de kans krijgen van de nieuwste methoden en technologieën gebruik te maken. We zouden graag zien dat nog meer partijen hun technologieën beschikbaar stellen. Maar het is lastig, want het ‘open stellen van gebruik voor anderen’ betekent dat je moet investeren in iets waar je zelf al bij kan. Dat is niet eenvoudig, dat is missiewerk en kost geld.

Wij stellen op termijn onze patroonherkenningsalgoritmes als *web-services* beschikbaar. Om intensief rekenwerk op data door vele anderen mogelijk te maken, kopiëren we onze gegevens en de benodigde *services* naar een machine met grote computerkracht, denk bijvoorbeeld aan het computercentrum SARA. Andere onderzoekers kunnen hun data op dezelfde wijze ter beschikking stellen via het project EUDAT. Aan EUDAT doen alle ‘SARA’s’ van Europa mee, zodat de gebruiker uit een groot data-aanbod kan kiezen. Onderzoekers spreken dan vanuit hun eigen computer bijvoorbeeld in ELAN een *web-service* aan, die de opgevraagde resultaten berekent en de annotaties terugstuurt. Het gebruik van deze ‘supercomputers’ past ook goed in de CLARIN-gedachte: maak de *data* en *tools* beschikbaar, zodat ze optimaal door iedereen gebruikt kunnen worden.

Het data- en toollandschap is momenteel nog sterk gefragmenteerd, dat is historisch zo gegroeid. Er bestaan verschillen in formaten, in technologieën, in vragen. Er is dus ook niet één oplossing, één standaard. Verschillende organisaties, zoals MPI, DANS en het Meertens Instituut, werken daarom samen aan gemeenschappelijke antwoorden op syntactische en semantische vraagstukken. Op het gebied van tools en diensten mag er best wel competitie zijn, maar als je dezelfde gebruikers bedient, moet je tenminste wat standaarden en *best practices* betreft samenwerken om het leven voor gebruikers makkelijk te maken.”

De Wet van Wittenburg

“Wees pragmatisch, kijk naar oplossingen die op de vloer gebruikt kunnen worden. En wees proactief. Als je denkt dat de wetenschapper morgen met een vraag kan komen, moet je vandaag werken aan de oplossingen. Het kost vaak ruim vijf jaar om een nieuwe technologie of methode – het gaat hier dus niet om een snel te maken script – breed in te kunnen zetten. En dus moet je vijf jaar eerder zijn dan de wetenschapper. Dat betekent dat je risico’s moet nemen. Dat doen wij, wij kijken altijd naar wat in de toekomst zou kunnen gebeuren. Zo kunnen we de wetenschap competitief houden. En maken we de grote vloed aan data voor de wetenschap toegankelijk.”