GEHOORD & BIJGEWOOND

Lunchlezing CLARIAH over transcriptie kranten

Suze Zijlstra

De Koninklijke Bibliotheek heeft veel zeventiende-eeuwse kranten gedigitaliseerd, maar de tekstherkenningstechniek OCR was niet goed genoeg om deze kranten goed doorzoekbaar te maken. Tijdens een CLARIAH lunchlezing op 8 april jongstleden spraken Nicoline van der Sijs (Radboud Universiteit) en Joris van Zundert (Huygens ING) over het project dat de transcriptie van deze kranten mogelijk maakte. Meer dan tweehonderd vrijwilligers werkten hieraan, wat een corpus van wel 20 miljoen woorden opleverde. Van Zundert lichtte eerst de achtergrond van het project binnen CLARIAH toe. CLARIAH heeft vanuit het werkpakket ‘tekst’ een oproep gedaan aan wetenschappers om pilotprojecten voor te stellen. Voor CLARIAH speelt de achterliggende vraag of het software kan ontwikkelen die breder toepasbaar is, of dat het kan leiden tot de ontwikkeling van bijvoorbeeld een tutorial over al bestaande methoden en technieken. Van der Sijs droeg haar krantentranscriptieproject aan. Dit project bood verschillende uitdagingen op het gebied van digitale tekst: niet alleen wat betreft de transcriptie, maar ook wat betreft metadata en annotatie. Nu de zeventiende-eeuwse kranten zijn getranscribeerd, gaan ze in de toekomst kijken naar de toepassing van programma’s als Transkribus of Calamari.

Van der Sijs presenteerde hoe ze met de vrijwilligers te werk is gegaan. Het project draaide om kranten tussen 1618 en 1700, waarvan al enige metadata beschikbaar waren. Vrijwilligers gingen aan de slag toen Transkribus nog niet ver was ontwikkeld. De bronnenset is voor taalwetenschappers en historici uniek: het is een longitudinale collectie met een mooi aaneengesloten corpus. Je kan hier zowel op micro- als op macroniveau onderzoek naar doen. Na de transcriptie zijn de metadata nog uitgebreid verbeterd. In de toekomst willen ze er uitgebreider onderzoek mee doen naar de herkomst van nieuws. Vanwege verschillende spellingsvarianten van plaatsnamen en het feit dat correspondenten niet altijd duidelijk waren over welke locatie ze schreven, is dit nog een mooie uitdaging.

CLARIAH.nl

Oprechte Haerlemse Saturdaegse Courant. N°. 1

SPANGIEN.

ITALIEN.

POLEN, PRUYSSEN, &c.

Duytslant en d'aengrensende Rijcken.

***Fragment uit Oprechte Haerlemsche courant, Haerlem,** 1694/01/02 00:00:00, p. 1., urn=ddd:011227040, Delpher.*

***Tijdens de meeting werden de mogelijkheden voor geo-analyses met CBS microdata verkend.** Credits: Daria from TaskArmy.nl via Unsplash.*

ODISSEI workshop Integrating Microdata with Geospatial Approaches

Eva Heitbrink

Onderzoek met CBS-microdata biedt veel mogelijkheden voor interessant onderzoek, maar wat zijn de mogelijkheden voor geo-analyses met CBS microdata? Dat is waar de ODISSEI workshop van 22 april over ging.

De middag begon met een introductie van geo-analyse door Marco Helbich (UU), die zelf onderzoek doet naar de invloed van omgevingsfactoren op mentale gezondheid met behulp van CBS microdata gekoppeld aan gps-data. Wat onderzoek met geo-data aantrekkelijk maakt, vertelt Helbich, is dat, vooral in Nederland, er vele geografische datasets gratis te gebruiken zijn. Deze datasets, gecombineerd met CBS microdata, kunnen voor de sociale wetenschappen nieuw en belangrijk onderzoek teweeg brengen. Hierna presenteerde Hilde van Oirschot (CBS) en Deirdre Bosch (CBS) over de mogelijkheden van CBS microdata, de geografische datasets die CBS zelf beschikbaar heeft, en hoe met deze data aan de slag te gaan. Bij deze datasets blijft de bescherming van privacy de hoogste prioriteit voor CBS. Geografische datasets met daarin gebieden met weinig datapunten zijn dan ook niet beschikbaar.

Tijdens het volgende deel van de middag presenteerden onderzoekers hun projecten uitgevoerd met CBS microdata en geografische data. Als eerst presenteerde Bastian Ravesteijn (EUR) zijn Kansenkaart-project, waarin Ravesteijn en zijn team met gepseudonimiseerde data van CBS de plek waarin mensen opgroeien verbindt aan hun huidige socio-economische status. Zo kan men op de kaart zien dat mensen die opgroeien in een relatief arm huishouden, nu vaak ook een lagere socio-economische status hebben. Na Ravesteijn presenteerde Nienke Boderie (Erasmus MC) haar project *Do neighbourhoods affect health?*, waarin zij en haar team onderzoeken of de buurt waarin iemand woont effect heeft op zijn of haar (mentale) gezondheid. Hiervoor kijkt het team naar verhuisbewegingen binnen Rotterdam en gezondheidsgegevens, waarvoor zij gebruik maken van gepseudonimiseerde CBS microdata. De middag sloot af met break-out rooms waarin de aanwezigen vragen konden stellen aan de presentatoren, onder andere over hun eigen ideeën voor onderzoek met microdata en geo-analyse.

odissei-data.nl

NPSO-lezingenmiddag over onzekerheden

Ricarda Braukmann

Op 6 mei organiseerde het Nederlandstalig platform voor surveyonderzoek (NPSO) een digitale lezingenmiddag over de presentatie van onzekerheden.

De middag trok zo’n 120 deelnemers en werd geopend door journalist Sanne Blauw. Als Correspondent Ontcijferen en in ‘Het beste verkochte boek ooit’ helpt Blauw het algemeen publiek om cijfers in de juiste context te plaatsen. Op deze bijeenkomst presenteerde zij haar visie op onzekerheid. Volgens Blauw worden we tijdens een pandemie continue geconfronteerd met onvolledige kennis en kunnen dit soort complexe problemen moeilijk juist worden ingeschat. Blauw citeert onderzoek naar het voorspellen van uitkomsten en pleit voor het durven twijfelen. Onzekerheden erkennen, jezelf bewust worden van je eigen vooroordelen en het veranderen van mening, leiden volgens haar tot de beste uitkomsten.

Edwin de Jonge, methodoloog bij het CBS, benadrukt dat onzekerheden een essentieel onderdeel zijn van onderzoek. Aan de hand van twee voorbeelden toont De Jonge hoe onzekerheden gepresenteerd kunnen worden. Ook benadrukt hij het belang van communicatie. “We denken vaak dat mensen betrouwbaarheidsintervallen niet kunnen begrijpen maar uit onderzoek blijk dat ook leken onzekerheden best goed kunnen interpreteren.”

De laatste presentatie door Marko Roos van het CBS ging verder in op de mogelijkheden van visualisaties. Roos en zijn team kregen in hun onderzoek naar mobiliteit te maken met grote hoeveelheden data. Door verplaatsingen tussen steden samen te vatten in zogenaamde ‘donut-maps’ kon het team van Roos op een slimme manier meer informatie weergeven door minder data te tonen.

De namiddag werd afgerond met een discussie rondom een aantal stellingen. De sprekers waren het eens dat het op de juiste manier communiceren over onzekerheden kan helpen om het begrip ervan te vergrootten. “Bij het weerbericht vinden we het ook niet meer gek als iemand zegt dat er morgen 80% kans is op regen”.

npso.net

Vervolg van pagina 1

Data LISS panel inzetten

verzamelde gegevens kunnen vervolgens gekoppeld worden aan andere data in het LISS Data Archive. Dit levert een zeer rijke dataset op waarmee vele onderzoeksvragen beantwoord kunnen worden. Zo onderzoekt Nynke van der Laan (Tilburg University) in opdracht van het Ministerie van VWS het gebruik van en opvattingen over de CoronaMelder app. Voor deze vragenlijst werden panelleden geselecteerd die eerder hadden meegedaan aan de kernvragenlijst *Health* en aan een onderzoek naar de gevolgen van het coronavirus, zodat extra achtergrondinformatie gebruikt kon worden. De voorlopige resultaten laten zien dat jongeren en mensen met basisschool of VMBO als hoogste opleidingsniveau de app naar verhouding minder gebruiken. De overheid wil daarom meer aandacht gaan besteden aan deze doelgroepen.

centerdata.nl/nl/databank/liss-panel-data

Vele mogelijkheden
Wil je meer informatie of ben je benieuwd of ook jouw vraagstuk beantwoord kan worden met gegevens uit het LISS panel? Neem dan contact op via info@lissdata.nl.

Open ODISSEI eScience Call ’21

In mei lanceerden ODISSEI en het Netherlands eScience Center een open call voor de sociale wetenschappen. Deze call ondersteunt sociale wetenschappers die digitale technieken in hun onderzoek willen gebruiken en hulp nodig hebben bij de toepassing hiervan. De call biedt uren van Research Software Engineers (RSEs) van het eScience Center als in-kind ondersteuning. Er is deze ronde ruimte voor zes projecten van elk maximaal drie RSE-persoonsmaanden. Wetenschappers die werken bij een ODISSEI deelnemer kunnen een aanvraag indienen voor 30 juli 2021 om 14 uur. (SZ)

odissei-data.nl

COLOFON Uitgever: E-data & Research. **Redactieadres:** Anna van Saksenlaan 51, 2593 HW Den Haag, 070-3494450, edata@dans.knaw.nl, edata.nl, @eDataResearch. **Hoofd-/eindredacteur**: Heidi Berkhout. **Redactie**: Ricarda Braukmann, Evi de Cock, Maarten Heerlien, Mathilde Jansen, Erica Renckens, Suze Zijlstra. **Redactiesecretariaat**: Ingrid van der Elst. **Aan dit nummer werkten mee**: Lamyk Bekius, Stephanie Bosschaert, Elly Dijk, Valentijn Gillissen, Eva Heitbrink, Hella Hollander, Ellen Leenarts, Josje Spinhoven, Nart Wielaard. **Opmaak**: Colette Sloots, Haarlem. **Productie**: Amsterdam University Press. **Druk**: Ten Brink, Meppel. **Webmaster**: Sonja Duijkers. **Oplage**: 4800 papier, 4800 digitaal. **ISSN**: 1872-0374. We hebben getracht alle belanghebbenden met betrekking tot het gebruikte beeldmateriaal te benaderen. Degenen die menen rechten te kunnen doen gelden, kunnen zich tot ons wenden. Toezending papieren en/of digitale versie is kosteloos aan relaties van de stakeholders en studenten in de alfa- en gammarichtingen. E-data & Research liever (tijdelijk) op een ander adres ontvangen? Geef wijzigingen aan ons door via edata@dans.knaw.nl.