ONDERZOEKSVOORSTEL

Scholieren met dyslexie van het derde graad secundair onderwijs ondersteunen bij het lezen van wetenschappelijke papers via tekstsimplificatie.

Bachelorproef, 2022-2023

Dylan Cluyse

E-mail: dylan.cluyse@student.hogent.be
Co-promotors:

- J. Decorte (Hogeschool Gent, johan.decorte@hogent.be)
- J. Van Damme (Hogeschool Gent, jana.vandamme@hogent.be)
- M. Dhondt (Gelukstraat marloes.dhondt@gelukstraat.be)

Samenvatting

Complexe woordenschat en zinsopbouw vormen een obstakel bij het lezen van wetenschappelijke papers voor scholieren met dyslexie van een derde graad secundair onderwijs. Tekstsimplificatie helpt scholieren met dyslexie van een derde graad secundair onderwijs bij hun lees- en verwerkingssnelheid. Artificiële intelligentie kan dit proces automatiseren. Nochtans missen Vlaamse secundaire scholen de toepassingen hiervan om specifiek scholieren met dyslexie van een derde graad secundair onderwijs beter te ondersteunen. Dit onderzoek evalueert de implementatie van tekstsimplificatiesoftware voor scholieren met dyslexie in het derde graad secundair onderwijs door technische analyse van de relevante vakgebieden en het ontwikkelen van een theoretische basis voor de tekstsimplificatie-pipeline. Er is een gebrek aan Nederlandstalige tekstsimplificatie-toepassingen die specifiek gericht zijn op Nederlandstalige scholieren met dyslexie in het derde graad secundair onderwijs, dus er is meer inzet nodig van zowel de overheid als bedrijven om ondersteunende software voor deze groep te ontwikkelen. Off-the-shelf modellen maken dit mogelijk, al moeten er compromises op taalvlak worden gemaakt.

Keuzerichting: AI & Data Engineering
Sleutelwoorden: Machineleertechnieken en kunstmatige intelligentie, tekstsimplificatie, dyslexie.

Inhoudsopgave

1. Introductie

België is een koploper in het gebruik van artificiële intelligentie (AI) op de werkvloer. Jaarlijks investeert de Vlaamse overheid 32 miljoen in het vakgebied (Crevits, 2022). Er verschijnen alsmaar meer kant-en-klare pakketten die complexe wiskundige berekeningen vereenvoudigen, zodanig dat ontwikkelaars sneller aan de slag kunnen om complexe problemen op te lossen. Soortgelijke technologieën worden amper toegepast in het derde graad van het secundair onderwijs, al zijn er wel taalgerelateerde AI-ontwikkelingen. Onder het amai!-project zijn er twee applicaties ontwikkeld die momenteel in het basis en secundair onderwijs worden ingezet, waaronder real-time ondertiteling in de les en My Speech, een taalassistent voor leerkrachten bij meertalige klasgroepen. Er is terughoudendheid door enerzijds ouders van leerlingen (Martens e.a., 2021a), anderzijds door de trage ontwikkeling in schoolgerelateerde AI-software. Toch zijn er reeds bewijzen dat artificiële intelligentie ook op school nuttig kan zijn.

Het STEM-agenda van de Vlaamse Overheid is een duidelijk initiatief om het STEM-onderwijs tegen 2030 aantrekkelijker te maken en door leraren, opleiders en begeleiders te ondersteunen. Plavén-Sigray e.a. (2017) halen aan hoe onderzoekers in hun complexe taalgeoriënteerde taalbubbel blijven, wat gevolgen voor de lezers met zich meebrengt. Daarnaast brengt de stijging aan het gebruik van acroniemen volgens Barnett en Doubleday (2020) een extra obstakel met zich mee. Dyslexiestudenten ondervinden hier uitermate veel last van, met een toename van *dropouts* binnen STEM-richtingen ten gevolg.

Dit onderzoek richt zich op hoe het secundair onderwijs tekstsimplificatie kan bieden voor leerlingen met dyslexie in het derde graad. Het be-

[1] https://amai.vlaanderen/

[2] https://www.vlaanderen.be/publicaties/stem-agenda-2030-stem-competenties-voor-een-toekomst-en-missiegericht-beleid

gint met het uitleggen van wat tekstsimplificatie is, bespreekt daarna de uitdagingen en voordelen van het lezen van wetenschappelijke teksten en hoe tekstsimplificatie en taalverwerking met AI kan helpen. Tenslotte bespreekt het de beschikbare software die momenteel in het secundair onderwijs wordt gebruikt en vermeldt de nodige machineleertechnieken en word embeddings in een proof-of-concept.

2. State-of-the-art

De voorbije tien jaar is artificiële intelligentie sterk verder ontwikkeld. De toename in kennis zorgde voor nieuwe toepassingen. Tekstsimplificatie vloeide hier uit voort. Momenteel bestaan er al robuuste applicaties voor tekstsimplificatie. Toch houdt de meerderheid niet genoeg rekening met het menselijk aspect van taalverwerking. Binnen het kader van tekstsimplificatie is er bestaande documentatie beschikbaar waar onderzoekers het voordeel van toegankelijkheid aanhalen, maar deze toepassingen ontbreken de extra noden die scholieren met dyslexie in het derde graad secundair onderwijs vereisen.

Het algemene doel van tekstsimplificatie is om ingewikkelde bronnen toegankelijker te maken. Het zorgt voor verkorte teksten zonder de kernboodschap te verliezen. Tekstsimplificatie gebeurt doorgaans op één van drie manieren. Er is conceptuele simplificatie waarbij documenten naar een compacter formaat worden getransformeerd. Daarnaast is er uitgebreide modificatie die kernwoorden aanduidt door gebruik van redundantie. Als laatste is er samenvatting die documenten verandert in kortere teksten met alleen de topische zinnen. Met deze concepten zijn ontwikkelaars in staat om ingewikkelde woorden te vervangen door eenvoudigere synoniemen of zinnen te verkorten zodat ze sneller leesbaar zijn (Siddharthan, 2014).

Tekstsimplificatie behoort tot de zijtak van natuurlijke taalverwerking (NLP) in artificiële intelligentie. NLP omvat methodes om, door machinaal leren, menselijke teksten om te zetten in tekst voor machines. Documenten vereenvoudigen met NLP kan op twee manieren: extract of abstract. Bij extractieve simplificatie worden zinnen gelezen zoals ze zijn neergeschreven. Vervolgens bewaart een document de belangrijkste taalelementen om de tekst te kunnen hervormen. Deze vorm van tekstsimplificatie komt het meeste voor (Sciforce, 2020). Daarnaast is er abstracte simplificatie die de kernboodschap van de zin bewaart en daarmee een nieuwe zin opbouwt. Deze vorm heeft potentieel dankzij de menselijke interpretatie, maar zit nog in de kinderschoenen (Chowdhary, 2020).

Voor kinderen met dyslexie bestaan digitale hulpmiddelen die voor een betere visuele presentatie zorgen van teksten. Het gaat over speciale lettertypes, spreiding tussen woorden en het gebruik van inzoomen op aparte zinnen. Weinig aandacht wordt besteed aan het veranderen van de tekst zelf, want dit kost tijd. Tekstsimplificatie door artificiële intelligentie kan een revolutionaire oplossing bieden.

Het onderzoek van Franse wetenschappers Gala en Ziegler (2016) illustreert dat manuele tekstsimplificatie schoolteksten toegankelijker maakt voor kinderen met dyslexie. Dit deden ze door simpelere synoniemen en zinsstructuren te gebruiken. Verwijswoorden werden vermeden en woorden kort gehouden. De resultaten waren veelbelovend. Het leestempo lag hoger en de kinderen maakten minder leesfouten. Ook bleek er geen verlies van begrip in de tekst bij geteste kinderen. Resultaten van de studie werden gebundeld voor de mogelijke ontwikkeling van een AI-hulpmiddel.

De Universiteit van Kopenhagen is met bovenstaande idee aan de slag gegaan. Onderzoekers Bingel e.a. (2018) hebben gratis software ontwikkeld, genaamd Lexi, om tekstsimplificatie voor mensen met dyslexie te automatiseren. De software bestudeert met welke woorden de gebruiker moeite heeft, en vervangt die door simpelere alternatieven. Hoe meer de software gebruikt wordt, hoe beter hij op maat van de gebruiker zal werken. Dit is de eerste en momenteel enige software van zijn soort. Voorheen bestond alleen generieke AI-software voor tekstsimplificatie. Lexi is beschikbaar als een browserextensie en tot nu toe enkel in het Deens.

NLP is de laatste decennia volop in ontwikkeling, maar ontwikkelaars botsen nog op uitdagingen. Het gaat om zowel interpretatie- als dataproblemen bij AI-machines. Allereerst is het voor een machine moeilijk om de context van homoniemen te achterhalen. Bijvoorbeeld bij het woord 'bank' is het niet duidelijk voor de machine of het gaat over de geldinstelling of het meubel. Daarnaast zijn synoniemen geen probleem voor tekstverwerking (Roldós, 2020).

Het merendeel van NLP-toepassingen maakt gebruik van Engelstalige invoer. Niet-Engelstalige toepassingen zijn zeldzaam. De opkomst van AI-technologieën die twee datasets gebruiken, biedt een oplossing voor dit probleem. De software vertaalt eerst de oorspronkelijke tekst naar de gewenste taal, voordat de tekst wordt herwerkt (Sciforce, 2020).

Om tekstsimplificatiemethoden te beoordelen, is er een tactvolle aanpak nodig. De studie van Swayamdipta (2019) haalt aan dat er extra nood is aan NLP-modellen waarbij de tekst zijn kernboodschap behoudt. Samen met Microsoft Research bouwden ze NLP-modellen die gericht waren op de bewaring van zinsstructuur en -context door *scaffolded learning*. Hiervoor maak-

ten de onderzoekers gebruik van een voorspellingsmethode die de positie van woorden en zinnen in een document beoordeelde.

De Vlaamse overheid leent gratis abonnementen uit voor voorlees- en schrijfsoftware, zoals SprintPlus, Alinea, Kurzweil3000, TextAid en Intowords. Middelbare scholieren met dyslexie in het secundair onderwijs in België kunnen een gratis Alinea-account aanvragen. Alinea is een software suite die hen ondersteunt bij het efficiënter lezen en schrijven van teksten, waardoor ze sneller en foutloos kunnen lezen zonder de kern van een artikel te verliezen.

Vlaanderen heeft weinig zicht op de geïmplementeerde AI-software in scholen. Dit werd geconstateerd door (Martens e.a., 2021a), een samenwerking tussen de Vlaamse universiteiten en overheid voor artificiële intelligentie. Vergeleken met andere Europese landen, maakt België het minst gebruik van leerling-georiënteerde hulpmiddelen. Degenen die wel gebruikt worden, zijn voornamelijk online leerplatformen voor zelfstandig werken. Ook maakt België amper gebruik van beschikbare software die de leermethoden en -noden van leerlingen evalueert (Martens e.a., 2021b).

Er zijn specifieke formules in de wiskunde die gebruikt worden om de complexiteit van teksten te meten, met de Flesch-Kincaid leesbaarheidstest als het meest prominente voorbeeld. Deze test bepaalt de moeilijkheidsgraad van tekst door verschillende factoren, zoals zinlengte, woordfrequentie en complexiteit van de taalgebruik, in aanmerking te nemen. De uitslag is een score die aangeeft hoe toegankelijk en begrijpelijk de tekst is. Bovendien zijn er kant-en-klare modellen die de complexiteit van tekst kunnen bepalen, hoewel deze beperkt zijn en vooral gericht zijn op Engelse teksten, zoals BERT, PaLM, XLNet en GPT-3.

De kerninhoud van een tekst dient te allen tijde behouden te blijven. Om dit te realiseren, worden er specifieke formules toegepast, waaronder de bekende Zipf's wet. Deze wet beschrijft de frequentie van woorden in een tekst in verhouding tot elkaar en stelt dat het meest voorkomende woord twee keer zo vaak aanwezig is als het tweede meest voorkomende woord, en zo verder.

3. Methodologie

Het onderzoek houdt vijf fases in. De eerste fase is het proces van tekstsimplificatie beschrijven. Dit gebeurt via een grondige studie van vakliteratuur en wetenschappelijke teksten. Ook blogs van experten komen hier aan bod. Na het verwerven van de nodige inzichten wordt er een verklarende tekst opgesteld.

De tweede fase bestaat uit het analyseren van wetenschappelijke werken over de bewezen voordelen van tekstsimplificatie bij scholieren met dyslexie van het derde graad secundair onderwijs. Hiervoor zijn geringe thesissen beschikbaar, die zorgvuldigheid vragen tijdens interpretatie. De resulterende tekst bevat de voordelen samen met hun wetenschappelijke onderbouwing.

De derde fase is opnieuw een beschrijving. Hier worden de valkuilen bij taalverwerking met AI-software nagegaan. Deze fase van het onderzoek brengt mogelijke nadelen en tekortkomingen van AI-software bij tekstsimplificatie aan het licht. Dit gebeurt aan de hand van een technische uitleg.

De vierde fase omvat een toelichting en advies over de beschikbare Nederlandstalige AI-tools voor tekstsimplificatie. Aan de hand van een kort veldonderzoek op het internet wordt er op zoek gegaan naar dergelijke software. Het opzoekingswerk leidt uiteindelijk tot testen van de applicaties. Ten slotte volgt er een persoonlijk advies over de nodige ontwikkelingen in het vak op vlak van Nederlandstalige tekstsimplificatie.

In de laatste fase van de ontwikkeling wordt er een proof-of-concept (POC) ontwikkeld voor een tekstsimplificatiepipeline. Deze POC maakt gebruik van Python en is gericht op het verzamelen van machineleertechnieken die enerzijds de inhoud van wetenschappelijke artikelen vereenvoudigen voor scholieren met dyslexie in het derde graad secundair onderwijs, alsook het evalueren van het model. Het bevat de nodige en bestaande machineleertechnieken zoals *word embeddings* en *libraries* die ontwikkelaars nodig hebben om de teksten op de noden van deze groep scholieren aan te passen.

Uit dit onderzoek moet duidelijk blijken of het mogelijk is om een Nederlandstalig tekstsimplificatiemodel aan de hand van *off-the-shelf* softwarepakketten op te zetten en te evalueren.

4. Verwacht resultaat, conclusie

Er wordt verwacht dat de longlist van software, dat momenteel in het onderwijs wordt ingezet, nog niet voldoet aan de noden van een scholier met dyslexie in het derde graad secundair onderwijs. Dit is omdat er onvolodende rekening wordt gehouden met de unieke uitdagingen omtrent leerstoornissen. De POC moet zich toespitsen op de noden van een scholier met dyslexie van een derde graad secundair onderwijs en moet de aanzet geven aan ontwikkelaars om hierop verder te bouwen. Het vertalen van de zinnen, mede door het gebrek aan Nederlandstalige *word embeddings* en off-the-shelf modellen, verlaagt de nauwkeurigheid van het model. Er is nood aan Nederlandstalige *word embeddings* die de complexiteit per woord bijhouden.

Referenties

Barnett, A., & Doubleday, Z. (2020). Meta-Research: The growth of acronyms in the scientific literature (P. Rodgers, Red.). *eLife*, *9*, e60080. https://doi.org/10.7554/eLife.60080

Bingel, J., Paetzold, G., & Søgaard, A. (2018). Lexi: A tool for adaptive, personalized text simplification. *Proceedings of the 27th International Conference on Computational Linguistics*, 245–258.

Chowdhary, K. (2020). *Fundamentals of Artificial Intelligence*. Springer, New Delhi.

Crevits, H. (2022, maart 13). *Kwart van bedrijven gebruikt artificiële intelligentie: Vlaanderen bij beste leerlingen van de klas* (Persbericht). Vlaamse Overheid Departement Economie, Wetenschap en Innovatie.

Gala, N., & Ziegler, J. (2016). Reducing lexical complexity as a tool to increase text accessibility for children with dyslexia. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 59–66.

Martens, M., De Wolf, R., & Evens, T. (2021a). *Algoritmes en AI in de onderwijscontext: Een studie naar de perceptie, mening en houding van leerlingen en ouders in Vlaanderen*. Kenniscentrum Data en Maatschappij. Verkregen maart 30, 2022, van https://data-en-maatschappij.ai/publicaties/survey-onderwijs-2021

Martens, M., De Wolf, R., & Evens, T. (2021b, juni 28). *School innovation forum 2021*. Kenniscentrum Data en Maatschappij. Verkregen april 1, 2022, van https://data-en-maatschappij.ai/nieuws/school-innovation-forum-2021

Plavén-Sigray, P., Matheson, G. J., Schiffler, B. C., & Thompson, W. H. (2017). Research: The readability of scientific texts is decreasing over time (S. King, Red.). *eLife*, *6*, e27725. https://doi.org/10.7554/eLife.27725

Roldós, I. (2020, december 22). *Major Challenges of Natural Language Processing (NLP)*. MonkeyLearn. Verkregen april 1, 2022, van https://monkeylearn.com/blog/natural-language-processing-challenges/

Sciforce. (2020, februari 4). *Biggest Open Problems in Natural Language Processing*. Verkregen april 1, 2022, van https://medium.com/sciforce/biggest-open-problems-in-natural-language-processing-7eb101ccfc9

Siddharthan, A. (2014). A survey of research on text simplification. *ITL - International Journal of Applied Linguistics*, *165*, 259–298.

Swayamdipta, S. (2019, januari 22). *Learning Challenges in Natural Language Processing*. Verkregen april 1, 2022, van https://www.microsoft.com/en-us/research/video/learning-challenges-in-natural-language-processing/