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

Tekstsimplificatie helpt scholieren met dyslexie in het derde graad middelbaar onderwijs bij hun lees- en verwerkingssnelheid. Artificiële intelligentie kan dit proces automatiseren. Vlaamse secundaire scholen ontbreken de toepassingen om scholieren met dyslexie van een derde graad secundair onderwijs te ondersteunen, ondanks de steun van de Vlaamse Overheid. Moeilijke woordenschat en zinsopbouw hinderen scholieren met dyslexie van een derde graad secundair onderwijs bij het lezen van wetenschappelijke papers. Het onderzoek achterhaalt welke toepassing scholieren met dyslexie in het derde graad secundair onderwijs de meest nauwkeurige en betekenisvolle tekst weergeeft bij het lezen van wetenschappelijke papers met tekstsimplificatie. Dit gebeurt via een technische analyse van de vakgebieden, samen met een veldonderzoek bij Belgische informaticabedrijven en de ontwikkeling van een tekstsimplificatiepipeline met bestaande kant-en-klare modellen. De huidige software in het onderwijs zal niet voldoen aan de behoeften van scholieren met dyslexie in het derde graad secundair onderwijs, omdat er onvoldoende rekening wordt gehouden met hun unieke uitdagingen. Internationale AI-tools moeten worden vertaald, maar de output zal afwijken van de oorspronkelijk context. Een eenvoudige tekstsimplificatiepipeline opbouwen met de beschikbare kant-en-klare algoritmen staat nog in de kinderschoenen, want er zijn *custom transformers* nodig. Het gebrek aan Nederlandstalige word embeddings en kant-en-klare algoritmen zal de nauwkeurigheid van het model verlagen en er is behoefte aan Nederlandstalige *word embeddings* die de complexiteit per woord bijhouden.

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). Soortgelijke technologieën worden amper toegepast in het derde graad van het secundair onderwijs, al zijn er wel taalgerelateerde AI-ontwikkelingen. Onder het amai!-project[1] 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. Volgens Martens e.a. (2021a) is er terughoudendheid door enerzijds ouders van leerlingen, anderzijds door de trage ontwikkeling in schoolgerelateerde AI-toepassingen.

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. Het onderzoek van Donato e.a. (2022) wijst aan dat er hierdoor meer scholieren met dyslexie binnen het secundair onderwijs uit hun richting vallen, wat voornamelijk bij STEM-richtingen het geval is. Het STEM-agenda[2] van de Vlaamse Overheid is een duidelijk initiatief om het STEM-

[1] https://amai.vlaanderen/

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

onderwijs tegen 2030 aantrekkelijker te maken en door leraren, opleiders en begeleiders te ondersteunen.

Dit onderzoek achterhaalt welke tekstsimplificatietoepassing scholieren met dyslexie in het derde graad secundair onderwijs de meest nauwkeurige en betekenisvolle tekst weergeeft bij het lezen van wetenschappelijke papers. Allereerst zal het onderzoek een technische omschrijving geven van wat tekstsimplificatie is. Aansluitend zal het onderzoek omschrijven uit welke verschillende soorten tekstsimplificatie wordt ingezet, inclusief een overtuiging over welke soort het beste aansluit voor deze casus.

Vervolgens bespreekt het onderzoek hoe tekstsimplificatie en taalverwerking met AI scholieren met dyslexie van het derde graad secundair onderwijs kan helpen. Nadien staat het onderzoek stil bij de struikelblokken op taalvlak waarmee een tekstsimplificatietoepassing rekening mee moet houden.

Tenslotte bespreekt het onderzoek welke software er momenteel wordt ingezet, alsook software die vrij beschikbaar is en eenzelfde tekstsimplificatiefunctie aanbiedt. Op basis van de onlinedocumentatie wordt er een van een tekstsimplificatiepipeline opgebouwd. Het resultaat van deze fase is een proof-of-concept pipeline waar de werking van tekstsimplificatie wordt weergegeven door middel van kant-en-klare *libraries* en modellen.

Als volgt geeft het onderzoek aan welke evaluatietechnieken er nodig zijn om de transformatie van een tekstsimplificatiepipeline te beoordelen. Tenslotte worden de momenteel ingezette applicaties, de absente applicaties en de tekstsimplificatiepipeline geëvalueerd, zo om aan te tonen welke toepassing de best mogelijke tekstinhoud als output geeft voor scholieren met dyslexie in het secundair onderwijs.

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 Hero, om tekstsimplificatie voor scholieren in het secundair onderwijs met dyslexie te automatiseren. De software bestudeert met welke woorden de gebruiker moeite heeft, en vervangt die door simpelere alternatie-

[1] https://beta.heroapp.ai/

ven. 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. Hero bevindt zich in beta-vorm en wordt enkel in het Engels en het Deens ondersteund.

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 maakten 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 voor deze software een gratis abonnement of licentie aanvragen. Al bieden de vijf softwarepaketten elk een eigen De focus van deze softwarepakketten ligt echter op spreek- en luistersoftware, waarbij het samenvatten van tekst als extra wordt gehouden.

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 zes 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.

[1] https://www.sprintplus.be/
[2] https://sensotec.be/product/alinea-suite/
[3] https://sensotec.be/product/kurzweil-3000/
[4] https://www.textaid-dyslexiesoftware.nl/textaid/
[5] https://intowords.nl/

De vijfde fase omschrijft de benodigde machineleertechnieken om zelf een tekstsimplificatiepipeline te maken. De pipeline wordt opgebouwd met beschikbare bibliotheken en algoritmes die vrij te vinden zijn. Het resultaat van deze fase is een pipeline opgebouwd in de Pythonprogrammeertaal.

In de laatste fase van het onderzoek worden de beschikbare Nederlandstalige AI-tools, alsook de zelfgemaakte tekstsimplificatiepipeline, tegenover elkaar geplaatst. Wetenschappelijke papers, die in een derde graad secundair onderwijs worden ingezet, dienen hier als inputdata. De complexiteit van de outputtekst wordt op drie objectieve factoren geëvalueerd: de *Flesch Reading Ease score*, de lengte van de outputzin ten opzichte van de oorspronkelijke zin en de complexiteit van de woorden. De teksten worden subjectief getest aan de hand van een survey of een *think-aloudtest*.

4. Verwacht resultaat, conclusie

Er wordt verwacht dat de software, die momenteel in het onderwijs wordt ingezet, niet voldoet aan de noden van een scholier met dyslexie in het derde graad secundair onderwijs. Dit is omdat er onvoldoende rekening wordt gehouden met hun unieke uitdagingen. De internationale AI-tools zijn meer gericht op het vereenvoudigen, al is er de compromis dat het vertalen mogelijks woorden uit context zal halen.

Er zijn onvoldoende bibliotheken en kant-en-klare algoritmen beschikbaar om een eenvoudige tekstsimplificatiepipeline te bouwen. De pipeline vergt zelfgemaakte transformers om betere resultaten te bekomen. 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.

Donato, A., Muscolo, M., Arias Romero, M., Caprì, T., Calarese, T., & Olmedo Moreno, E. M. (2022). Students with dyslexia between school and university: Post-diploma choices and the reasons that determine them. An Italian study. *Dyslexia*, *28*(1), 110–127. https://doi.org/https://doi.org/10.1002/dys.1692

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/