

ONDERZOEKSVOORSTEL

Scholieren met dyslexie van het derde graad middelbaar 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. Moeilijke woordenschat en zinsopbouw hinderen scholieren met dyslexie van een derde graad middelbaar onderwijs bij het lezen van wetenschappelijke papers. Vlaamse middelbare scholen ontbreken de toepassingen om scholieren met dyslexie van een derde graad middelbaar onderwijs te ondersteunen, ondanks de steun van de Vlaamse Overheid. Dit onderzoek achterhaalt welke toepassing het meest conform is aan de noden van een scholier met dyslexie in het derde graad middelbaar onderwijs bij het lezen van wetenschappelijke papers. Dit gebeurt via een technische analyse van de vakgebieden, samen met een veldonderzoek bij Belgische informaticabedrijven en de ontwikkeling van een tekstsimplificatiepipeline met kant-en-klare AI-modellen. De huidige tekstsimplificatiefuncties in het onderwijs zijn niet gebouwd voor een specifieke doelgroep en dienen om een zo breed mogelijk publiek te dienen binnen het lager en middelbaar onderwijs. Internationale AI-toepassingen hebben een vertaalfase, maar er is geen garantie dat de oorspronkelijke context wordt behouden in de getransformeerde tekst. Het ontwikkelen van een tekstsimplificatiepipeline met kant-en-klare algoritmen staat nog in de kinderschoenen en er is nood aan aangepaste *transformers*.

Keuzerichting: AI & Data Engineering

Sleutelwoorden: Machineleertechnieken en kunstmatige intelligentie, tekstsimplificatie, dyslexie.

Inhoudsopgave

1	Introductie	1
2	State-of-the-art	2
3	Methodologie	3
4	Verwacht resultaat, conclusie	4
	Referenties	4

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). Zo zijn er verschillende projecten uit de grond gestampt om taalgerelateerde AI-ontwikkelingen op te starten. Vooral het amai!-project¹ geeft de aanzet door twee toepassingen te ontwikkelen die nu in middelbare klassen terug te vinden zijn, namelijk *real-time* ondertiteling en een taalassistent voor leerkrachten bij meertalige klasgroepen.

¹<https://amai.vlaanderen/>

Het STEM-agenda² van de Vlaamse Overheid omvat aandachtspunten om het STEM-onderwijs tegen 2030 aantrekkelijker te maken door leraars, opleiders en begeleiders te ondersteunen. Het overbruggen van wetenschappelijke jargon is nergens in de aandachtspunten terug te vinden. Dit is echter spijtig, want de derde graad van het middelbaar onderwijs is een cruciale stap voor de verdere loopbaan van scholieren.

Dit onderzoek achterhaalt welke toepassing het meest conform is aan de noden van een scholier met dyslexie in het derde graad middelbaar onderwijs bij het lezen van wetenschappelijke papers. Eerst vat het onderzoek samen wat tekstsimplificatie inhoudt en uit welke soorten deze technologie bestaat. Vervolgens bespreekt het onderzoek hoe tekstsimplificatie en taalverwerking met AI scholieren met dyslexie van het derde graad middelbaar onderwijs kan helpen. Nadien staat het onderzoek stil bij de struikelblokken op

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

taalvlak waarmee een tekstsimplificatietoepassing rekening mee moet houden. Als volgt bespreekt het onderzoek de tekstsimplificatietoepassingen die momenteel in het onderwijs worden ingezet, internationale applicaties en een zelfgemaakte pipeline. Vervolgens geeft het onderzoek aan welke evaluatietechnieken er nodig zijn om de transformatie van een tekstsimplificatie te beoordelen. Het onderzoek eindigt met een vergelijkende studie van de aangehaalde toepassingen. De getransformeerde tekst wordt beoordeeld op basis van evaluatiecriteria.

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 middelbaar 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 eenvoudiger 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. Volgens het onderzoek van Chowdhary (2020) heeft deze vorm potentieel dankzij de

menselijke interpretatie, maar zit nog in de kinderschoenen.

Voor kinderen met dyslexie bestaan digitale hulpmiddelen die voor een betere visuele presentatie zorgen van teksten. Zo haalt het onderzoek van Rello e.a. (2012) tips aan waarmee teksten en documenten rekening moeten houden bij scholieren met dyslexie in het derde graad middelbaar onderwijs. Het gaat over speciale lettertypes, spreiding tussen woorden en het gebruik van inzoomen op aparte zinnen. Het onderzoek haalt aan dat teksten voor deze unieke noden aanpassen tijdovend is, dus 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 Hero³, om tekstsimplificatie voor scholieren in het middelbaar onderwijs met dyslexie te automatiseren. De software bestudeert met welke woorden de gebruiker moeite heeft, en vervangt die door simpelere alternatieven. Hero bevindt zich in beta-vorm en wordt enkel in het Engels en het Deens ondersteund.

Plavén-Sigra e.a. (2017) halen aan hoe onderzoekers in hun 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 erop dat ondoordringelijke teksten te wijten zijn aan scholieren met dyslexie in het middelbaar onderwijs die uit hun richting vallen, wat voornamelijk bij STEM-richtingen het geval is.

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).

³<https://beta.heroapp.ai/>

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). Hetzelfde onderzoek bewijst dat het vertalen van gelijkaardige talen, zoals Duits en Nederlands, een minimaal verschil oplevert.

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 SprintPlus⁴, Alinea⁵, Kurzweil3000⁶, TextAid⁷ en Intowords⁸. Middelbare scholieren met dyslexie in het middelbaar onderwijs in België kunnen voor deze software een gratis abonnement of licentie aanvragen. Al bieden de vijf softwarepakketten elk een eigen tekstsimplificatiefunctie, echter ligt de focus op spreek- en luistersoftware, waarbij het samenvatten en markeren 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. Vergelijken 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).

Het onderzoek van Vásquez-Rodríguez e.a. (2021) geeft een aanzet op de objectieve evaluatie van een transformatie bij tekstsimplificatie. De kerninhoud van een tekst dient te allen tijde behouden te blijven. Om dit te realiseren, worden er specifieke formules toegepast, waaronder de wet van Zipf. Daarnaast is er de Flesch-Kincaid leesbaarheidstest, wat de moeilijkheidsgraad van tekst bepaald door verschillende factoren, zoals zinlengte, woordfrequentie en complexiteit van de taalge-

bruik, 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 BERT⁹, XLNet¹⁰, GPT-3¹¹ en een open-source model genaamd TRU-NAJOD¹².

3. Methodologie

Het onderzoek houdt zeven fases in. De eerste fase is het proces van tekstsimplificatie beschrijven, waaronder een omschrijving van het begrip en de verschillende soorten van technologische tekstsimplificatie. Dit gebeurt via een grondige studie van vakliteratuur en wetenschappelijke teksten. Ook blogs van experts 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 middelbaar onderwijs. Hiervoor zijn geringe thesen beschikbaar, die zorgvuldigheid vragen tijdens interpretatie. De resulterende tekst bevat de voordelen samen met hun wetenschappelijke onderbouwing.

De derde fase is het verzamelen van alle nodige transformaties om een wetenschappelijke paper beter leesbaar te maken voor een scholier met dyslexie in het derde graad middelbaar onderwijs. Het resultaat is een shortlist van alle evaluatiecriteria waaraan de outputtekst van een tekstsimplificatietoepassing moet voldoen.

De vierde 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 vijfde fase omvat een toelichting en advies over de beschikbare Nederlandstalige AI-tools voor tekstsimplificatie. Aan de hand van een veldonderzoek op het internet en bij bedrijven wordt er op zoek gegaan naar dergelijke software. Er wordt niet gezocht naar vertaalsoftware of toepassingen die de inhoud van een afbeelding of tekstbestand omzet naar tekstinhoud.

De zesde fase omschrijft een shortlist van de benodigde machineleertechnieken om zelf een tekstsimplificatiepipeline op te bouwen, alsook een shortlist van metrieken om de transformatie te evalueren. Op basis van de metrieken wordt er een pipeline ontwikkeld met beschikbare kant-

⁴<https://www.sprintplus.be/>

⁵<https://sensotec.be/product/alinea-suite/>

⁶<https://sensotec.be/product/kurzweil-3000/>

⁷<https://www.textaid-dyslexiesoftware.nl/textaid/>

⁸<https://intowords.nl/>

⁹https://huggingface.co/docs/transformers/model_doc/bert

¹⁰https://huggingface.co/docs/transformers/model_doc/xlnet

¹¹<https://platform.openai.com/docs/>

¹²<https://trunajod20.readthedocs.io/>

en-klare bibliotheken en algoritmes. Het resultaat van deze fase is een pipeline opgebouwd in de programmeertaal Python.

De zevende en laatste fase omvat een vergelijkende studie met de gevonden applicaties en de zelfgemaakte pipeline. Wetenschappelijke papers, die in een derde graad middelbaar onderwijs worden ingezet, dienen hier als invoertekst. De transformatie wordt met zowel objectieve als subjectieve metrieken beoordeeld. De subjectieve test gebeurt aan de hand van een *survey* en een *think-aloudtest*. De objectieve testen gebeuren op basis van de shortlist uit de derde fase en de shortlist van metrieken uit de zesde fase. Ten slotte volgt er een persoonlijk advies over de nodige ontwikkelingen in het vak op vlak van Nederlandstalige tekstsimplificatie.

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 middelbaar onderwijs. Dit is omdat er onvoldoende rekening wordt gehouden met hun unieke uitdagingen. Het vertalen van de outputtekst bij een internationale AI-tool zal mogelijk afwijken van de oorspronkelijke context.

Er zijn onvoldoende kant-en-klare algoritmen en modellen beschikbaar om een tekstsimplificatiepipeline, waarvan de output verzorgd is aan de unieke noden van een scholier met dyslexie in het derde graad middelbaar onderwijs, te bouwen. De pipeline vergt *custom transformers* om nauwkeurige resultaten te bekomen, zodat de kerninhoud niet verloren raakt. Het vertalen van de zinnen verlaagt de nauwkeurigheid van het model, maar is een acceptabel alternatief. Er is nood aan Nederlandstalige *word embeddings* die de complexiteit per woord bijhouden, alsook meer kant-en-klare modellen die tekstsimplificatiefuncties aanbieden.

Referenties

- Barnett, A., & Doubleday, Z. (2020). Meta-Research: The growth of acronyms in the scientific literature (P. Rodgers, Red.). *eLife*, 9, e60080.
- Bingel, J., Paetzold, G., & Søggaard, 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., Capri, 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.
- 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.
- Rello, L., Kanvinde, G., & Baeza-Yates, R. (2012). Layout Guidelines for Web Text and a Web Service to Improve Accessibility for Dyslexics. *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*.
- 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/>
- Vásquez-Rodríguez, L., Shardlow, M., Przybyla, P., & Ananiadou, S. (2021). Investigating Text Simplification Evaluation. *CoRR*, abs/2107.13662.