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

Optionele ondertitel.

Dylan Cluyse.

Scriptie voorgedragen tot het bekomen van de graad van
Professionele bachelor in de toegepaste informatica

Promotor: Mevr. L. De Mol
Co-promotor: J. Decorte; J. Van Damme; M. Dhondt
Academiejaar: 2022–2023
Eerste examenperiode

Departement IT en Digitale Innovatie .

HO
GENT

Woord vooraf

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Samenvatting

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Inhoudsopgave

Lijst van figuren

1

Inleiding

Het middelbaar onderwijs staat op springen. Dagelijks sneuvelen leerkrachten en leerlingen van het middelbaar onderwijs onder de te harde werkdruk. Daarnaast is taal vrijwel onmogelijk om aan te ontsnappen. Dagelijks komen mensen in aanraking met taal, van Nederlandse nieuwsartikelen tot de ondertiteling van Koreaanse Netflix-series, ongeacht de doelgroep. Het onderwijs richt zich de afgelopen tien jaar sterk op het gebruik van gevarieerde bronnen in lessen. De moeilijkheidsgraad van deze bronnen verandert echter niet, want de noodzaak aan verscheidenheid brengt ook de noodzaak aan uitdagingen met zich mee. STEM-leerkrachten in een derde graad middelbaar onderwijs moeten volgens het leerplan van zowel het katholiek[1] als het gemeenschapsonderwijs[2] hun theorielessen op een toegankelijke manier aanbieden, zodat iedereen betrokken is bij het verhaal.
Met een jaarlijks budget van 32 miljoen is België een pionier (Crevits, 2022) in het vakgebied kunstmatige intelligentie (AI) op de werkvloer. Zo zijn er verschillende projecten, om Vlaamse AI-ontwikkelingen in het onderwijs op te starten, uit de grond gestampt. Het amai!-project[3] brengt AI-softwarebedrijven samen uit verschillende domeinen. Dit project leidt tot het ontstaan van AI-toepassingen die processen automatiseren om de werkdruk te verminderen, zoals binnen het onderwijs *real-time* ondertiteling en een taalassistent voor leerkrachten in meertalige klasgroepen.

1.1. Probleemstelling

Scholieren met dyslexie in het middelbare onderwijs krijgen te maken met vele uitdagingen. Gelukkig worden ze niet aan hun lot overgelaten en kunnen ze rekenen op ondersteuning van coaches en beschikbare hulpmiddelen om hun ach-

[1]https://pro.katholiekonderwijs.vlaanderen/basisoptie-stem/ondersteunend-materiaal
[2]https://g-o.be/stem/
[3]https://amai.vlaanderen/

terstand te beperken. Het leerplan voor STEM-vakken stimuleert het gebruik van wetenschappelijke artikelen, maar houdt niet altijd rekening met de moeilijkheidsgraad ervan. De complexe woordenschat en zinsopbouw in deze artikelen vormen een barrière voor begrijpelijkheid, waardoor de scholieren de kerninhoud moeilijk kunnen doorgronden.
Het vereenvoudigen van wetenschappelijke artikelen vraagt tijd en inspanning van STEM-docenten in het derde graad middelbare onderwijs. Het onderwijs staat onder druk en docenten hebben al moeite om hun werklast aan te kunnen. Daarom is er behoefte aan software die wetenschappelijke artikelen automatisch kan vereenvoudigen, specifiek afgestemd op de behoeften van scholieren met dyslexie. Een dergelijke toepassing zou het routinematige werk van STEM-docenten verlichten en scholieren met dyslexie in het derde graad middelbare onderwijs de mogelijkheid bieden om de kern van een tekst beter te begrijpen.

1.2. Onderzoeksvraag

Dit onderzoek toont aan hoe de inhoud van wetenschappelijke artikelen met kunstmatige intelligentie automatisch vereenvoudigd kan worden, specifiek gericht op de noden van een scholier met dyslexie in het derde graad middelbaar onderwijs. Om een antwoord op deze onderzoeksvraag te vinden, moet het onderzoek eerst zeven fasen doorlopen.

- Wat is geautomatiseerde tekstvereenvoudiging? Allereerst moeten er een definitie worden gevormd wat geautomatiseerde tekstvereenvoudiging is en welke transformaties bijdragen tot een tekstvereenvoudiging. De nodige theoretische concepten om tekstvereenvoudiging mogelijk te maken, worden aangehaald.
- Wat zijn de voordelen van wetenschappelijke artikelen te vereenvoudigen bij scholieren met dyslexie in de derde graad middelbaar onderwijs? Waarom speelt tekstvereenvoudiging een rol bij wetenschappelijke artikelen?
- Wat zijn de struikelblokken bij het vereenvoudigen van wetenschappelijke artikelen?
- Welke fasen heeft een pipeline voor tekstvereenvoudiging bij wetenschappelijke artikelen? Welke modellen zijn er white-box of black-box?
- Aan welke metrieken moet een vereenvoudigd wetenschappelijk artikel voldoen? In welke mate kan de eindgebruiker hiervan op de hoogte gesteld worden?

1.3. Onderzoeksdoelstelling

Het resultaat van dit onderzoek is een vergelijkende studie en een prototype voor een toepassing die de tekstinhoud van een wetenschappelijke paper zal omzetten. De vergelijkende studie zal vereenvoudigde of samengevatte teksten van drie verschillende soorten programma's vergelijken:

- Toepassingen die momenteel in het onderwijs worden ingezet en waarvan licenties aan te vragen zijn voor scholieren in het derde graad van het middelbaar.
- Toepassingen die online terug te vinden zijn.
- Een zelfgemaakte prototype dat de inhoud van een wetenschappelijke paper automatisch zal vereenvoudigen met kunstmatige intelligentie.

Als tweede onderdeel wordt er een prototype ontwikkeld om wetenschappelijke artikelen automatisch te vereenvoudigen, specifiek gericht op de noden van een scholier in de derde graad middelbaar onderwijs. Het prototype houdt geen rekening met de transformatie van het bronbestand, bijvoorbeeld een PDF of een afbeelding, naar de tekstinhoud. Dergelijke AI-toepassingen of AI-modellen die tekst uit afbeeldingen of PDF-bestanden halen, bestaan al. De invoer van dit prototype is een wetenschappelijk artikel van 300 tot 500 woorden lang. De uitvoer van dit prototype is een vereenvoudigde versie van ditzelfde wetenschappelijk artikel. Metrieken, indien mogelijk per zin, worden weergegeven. Verdere concretisering volgt...

1.4. Opzet van deze bachelorproef

De rest van deze bachelorproef is als volgt opgebouwd:
In Hoofdstuk 2 wordt een overzicht gegeven van de stand van zaken binnen het onderzoeksdomein, op basis van een literatuurstudie.
In Hoofdstuk 3 wordt de methodologie toegelicht en worden de gebruikte onderzoekstechnieken besproken om een antwoord te kunnen formuleren op de onderzoeksvragen.

- Wat is geautomatiseerde tekstvereenvoudiging? Allereerst moeten er een definitie worden gevormd wat geautomatiseerde tekstvereenvoudiging is en welke transformaties bijdragen tot een tekstvereenvoudiging. De nodige theoretische concepten om tekstvereenvoudiging mogelijk te maken, worden aangehaald.
- Wat zijn de voordelen van wetenschappelijke artikelen te vereenvoudigen bij scholieren met dyslexie in de derde graad middelbaar onderwijs? Waarom speelt tekstvereenvoudiging een rol bij wetenschappelijke artikelen?

- Wat zijn de struikelblokken bij het vereenvoudigen van wetenschappelijke artikelen?
- Welke fasen heeft een pipeline voor tekstvereenvoudiging bij wetenschappelijke artikelen? Welke modellen zijn er white-box of black-box?
- Aan welke metrieken moet een vereenvoudigd wetenschappelijk artikel voldoen? In welke mate kan de eindgebruiker hiervan op de hoogte gesteld worden?

In Hoofdstuk 4, tenslotte, wordt de conclusie gegeven en een antwoord geformuleerd op de onderzoeksvragen. Daarbij wordt ook een aanzet gegeven voor toekomstig onderzoek binnen dit domein.

2

Stand van zaken

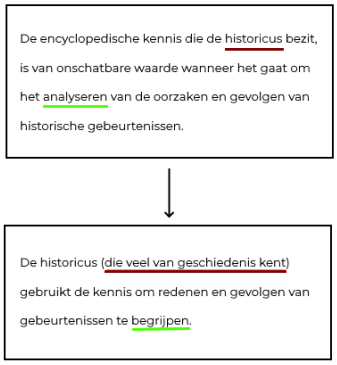
2.1. Tekstvereenvoudiging

Tekstvereenvoudiging is het proces waarin het technisch leesniveau en/of woordgebruik van een geschreven tekst wordt verminderd. Belangrijk hierbij is dat de vereenvoudiging geen effect mag hebben op de kerninhoud. Een complete vereenvoudiging van een tekst bestaat uit minstens drie transformaties (Siddharthan, 2014). Daarnaast is tekstvereenvoudiging een taalbewerking dat geautomatiseerd kan worden. Tekstvereenvoudiging is namelijk een zijtak van natuurlijke taalverwerking.

2.1.1. Natural Language Processing

Natuurlijke taalverwerking of NLP is een brede term die zich richt op het verwerken en analyseren van menselijke taal door computers en andere technologieën. Het omvat verschillende technieken, zoals tekstanalyse, taalherkenning en -generatie, spraakherkenning en -synthese, en semantische analyse. Computers zijn ertoe in staat om op een menselijke manier te communiceren en begrijpen wat er wordt gezegd. Vooraleer het onderzoek zich verdiept in hoe teksten worden vereenvoudigd, moeten er eerst begrippen worden aangehaald die noodzakelijk zijn om de volgende fasen te kunnen uitleggen. Sohom (2019) haalt de volgende begrippen aan.

- **Tokenisatie** splitst de stam of basisvorm van woorden in een tekst. Gebruikelijk zetten ontwikkelaars deze stap in om een woordenschat voor een taalmodel op te bouwen. Bij tokenisatie wordt er geen rekening gehouden met de betekenis achter ieder woord.
- **Lemmatiseren** in NLP bouwt verder op *stemming*, maar de betekenis van ieder woord wordt in acht genomen. Voor het lemmatiseren bestaan er Ne-



Figuur (2.1)
Voorbeeld van manuele tekstvereenvoudiging. Oorspronkelijke tekst uit Historia 5 bron toe te voegen

derlandstalige modellen, waaronder JohnSnow[1]. Bij **omgekeerd lemmatiseren** wordt er een afgeleide achterhaald vanuit de stam. Bijvoorbeeld voor het werkwoord 'zijn' zou dit 'is', 'was' of 'ben' zijn. Voor zelfstandige naamwoorden, zoals 'hond', is dit dan enkelvoud of meervoud.

- Bij een **parsing**-fase wordt er een label aan ieder woord of zinsdeel toegekend. Voorbeelden van labels zijn zelfstandig naamwoord, bijwoord, werkwoord, bijzin of stopwoord. Het herkennen van zinsdelen wordt *chunking* genoemd. Parsing heeft een dubbelzinnigheidsprobleem, want een 'plant' staat niet gelijk aan de vervoeging van werkwoord 'planten'.

Sequence Labeling

Volgens Eisenstein (2019) is *sequence labeling* essentieel tot het achterhalen van de structuur van een tekst met *supervised learning*. Elk woord in een tekst of zin wordt geclassificeerd met behulp van specifieke labels, zoals bijvoorbeeld een Part of Speech (PoS) label of een Named Entity Recognition label. De structuur van de tekst wordt achterhaald en informatie en patronen kunnen uit de tekst worden gehaald.

2.2. De verschillende soorten tekstvereenvoudiging

Tekstvereenvoudiging bestaat uit vier soorten transformaties: lexicale, syntactische en semantische vereenvoudiging en samenvatten.

2.2.1. Lexicale vereenvoudiging

Bij lexicale vereenvoudiging worden complexe woorden vervangen door eenvoudigere synoniemen. Bijvoorbeeld, het woord 'adhesief' kan worden vervangen door 'klevend'. De zinsstructuur verandert niet en er is garantie dat de kerninhoud en be-

[1]https://nlp.johnsnowlabs.com/2020/05/03/lemma_nl.html

nadrukking hetzelfde blijft. Het doel van lexicale vereenvoudiging is om de moeilijkheidsgraad van de woordenschat in een zin of tekst te verlagen. Dit is, volgens het aantal onderzoeken, de meest gekende vorm van vereenvoudiging en een noodzakelijke stap bij het vereenvoudigen van een tekst. Voor prevalente domeinen, zoals de onderwijs-, medische en financiële sector, zijn er onderzoeken vrij beschikbaar. In de medische sector haalt Kandula e.a. (2010) twee manieren aan om lexicale vereenvoudiging mogelijk te maken, namelijk het vervangen door een synoniem en het aanmaken of genereren van extra uitleg. Zij bouwden verder op een vorig onderzoek van Zeng e.a. (2005).

2.2.2. Syntactische vereenvoudiging

Syntactische vereenvoudiging transformeert de grammatica en zinsstructuur van een tekst om de complexiteit van een zin te verlagen. Bijvoorbeeld, twee afzonderlijke zinnen kunnen worden samengevoegd tot één eenvoudigere zin. Syntactische vereenvoudiging richt zich op het verminderen van complexe of onduidelijke zinsconstructies, terwijl de inhoud en betekenis van de tekst behouden blijft. Dergelijke transformaties zijn het vereenvoudigen van de syntax of door de zinnen korter te maken. Zinnen worden toegankelijker, zonder de kerninhoud of relevante inhoud te verliezen.
Het vereenvoudigen van medische journalen wordt besproken in het onderzoek van Kandula e.a. (2010). Zij ontwikkelden een toepassing om medische informatie te vereenvoudigen met beschikbare biomedische bronnen, door syntactische vereenvoudiging op zinniveau toe te passen. Zinnen met meer dan 10 woorden worden als complex beschouwd en worden verwerkt door drie modules. Op het einde van deze vereenvoudiging kan de oorspronkelijke zin ongewijzigd worden behouden of vervangen worden door twee of meer kortere zinnen. De architectuur van het model omvat drie onderdelen: een *Part of Speech (PoS) Tagger*, een *Grammar Simplifier* en een *Output Validator*.

- Voor de *PoS Tagger*-fase gebruikten de onderzoekers beschikbare functies uit het open-source pakket OpenNLP[2].
- De *Grammar Simplifier* module splitst de lange zin in twee of meer kortere zinnen door POS-patronen te identificeren en een set transformatieregels toe te passen.
- De *Output Validator* module controleert de output van de Grammar Simplifier op grammatica en leesbaarheid. Er zijn drie condities:
  -

[2]https://opennlp.apache.org/

2.2.3. Conceptuele vereenvoudiging

Conceptuele vereenvoudiging lost dit probleem op. Theoretische kennis hierover is schaars, maar Siddharthan (2006) bestudeerde dit concept verder. Dit type vereenvoudiging betreft het opdelen van complexe concepten in eenvoudigere delen, het gebruik van duidelijke en bondige taal en het vermijden van technische jargon en abstracte uitdrukkingen. Het doel is om de inhoud begrijpelijker te maken, zonder dat hierbij de betekenis of nauwkeurigheid wordt aangetast. Siddharthan (2006) noemt deze transformatie een vorm van elaboratie of het uiteenzetten van een begrip.

2.2.4. Semantische vereenvoudiging

2.2.5. Tekstvereenvoudiging automatiseren

Geautomatiseerde tekstvereenvoudiging is niets nieuws. Volgens het onderzoek van Canning e.a. (2000) en Siddharthan (2006) waren de eerste aanpakken op geautomatiseerde tekstvereenvoudiging gebouwd op rule-based modellen. Deze modellen bewerken de syntax door zinnen te splitsen, te verwijderen of de volgorde van de zinnen in een tekst aan te passen. Lexicale vereenvoudiging kwam hier niet aan de pas. Enkel bij recentere onderzoeken van Coster en Kauchak (2011) en Bulté e.a. (2018) werd het duidelijk hoe lexicale en syntactische vereenvoudiging gecombineerd kon worden.

2.2.6. Discourse edits

Discourse

2.2.7. Combineren tot het geheel van tekstvereenvoudiging

Het onderzoek van De Belder (2010) richt zich op tekstvereenvoudiging voor kinderen. De doelgroep ligt echter jonger dan deze casus, maar het onderzoek haalt aan hoe de onderzoekers een methode opzetten voor lexicale en syntactische vereenvoudiging.

Een onderzoek van Bulté e.a. (2018) ging met dit concept aan de slag. Het resultaat van hun onderzoek was een *pipeline* ontworpen om moeilijke woordenschat naar simpele synoniemen te vervangen. Eerst ging de tekstinhoud door een *preprocessing*-fase, samen met het uitvoeren van WSE. Daarna werd de moeilijkheidsgraad van ieder token overlopen. De moeilijkheidsgraad is gebaseerd op hoe vaak een woord voorkomt in SONAR500[3] een corpus met eenvoudige Nederlandstalige woorden. Synoniemen werden teruggevonden met Cornetto[4], een lexicale databank met Nederlandstalige woorden. Hiervoor gebruikten de onderzoekers een *reverse lemmatization* fase. Lexicale vereenvoudiging is ingewikkeld wanneer er geen eenvoudigere synoniemen zijn. In dat geval blijft een moeilijk woord voor wat

[3]https://taalmaterialen.ivdnt.org/download/tstc-sonar-corpus/
[4]https://github.com/emsrc/pycornetto

het is.

2.2.8. Samenvatten

Lexicale, conceptuele en syntactische vereenvoudiging is er geen garantie dat de tekstinhoud korter zal worden. Eenvoudigere woordenschat
El-Kassas e.a. (2021) deed verder onderzoek op geautomatiseerd samenvatten.

2.3. Voordelen van tekstvereenvoudiging

2.4. Struikelblokken

2.4.1. Evaluatie van de toepassing

2.4.2. Datasets

2.4.3. Meaning distortion

2.4.4. Word Ambiguity

Sequence Labeling voorziet labels aan tokens in een tekst. Homoniemen kunnen echter roet in het eten gooien, want .

2.4.5. Paternalisme

De doelstelling van assisterende software is om gelijke kansen te bieden aan iedereen. Zoals eerder vermeld, zorgt tekstvereenvoudiging voor een simpelere syntax en woordenschat in een tekst. Volgens Niemeijer e.a. (2010) zijn de ethische overwegingen die samenhangen met tekstvereenvoudiging via implicaties voor assistieve technologie niet gemakkelijk te scheiden van de technologie die wordt gebruikt om het resultaat te bereiken. Ontwikkelaars moeten, volgens deze auteur, rekening houden met de doelgroep waarvoor ze een toepassing maken.
Het onderzoek van Gooding (2022) richtte zich op dit probleem. Ontwikkelaars moeten zich meer bewust worden van de behoeften en verwachtingen van de eindgebruiker bij het ontwikkelen van een tekstvereenvoudigingstoepassing. Haar onderzoek benadrukt de paternalistische en afhankelijke aard van assisterende toepassingen. Tekstvereenvoudiging omvat drie transformaties, maar de moeilijkheidsgraad is niet statisch. Een adaptieve tekstvereenvoudigingstoepassing moet de eindgebruiker een keuze aanbieden om aan te passen wat vereenvoudigd wordt, afhankelijk van zijn of haar specifieke behoeften.
Volgens Sikka en Mago (2020), maken de meeste AI-toepassingen voor tekstvereenvoudiging gebruik van *black-box* modellen. Een *black-box* model maakt het onmogelijk om transparant te zijn over waarom bepaalde transformaties worden uitgevoerd, bijvoorbeeld het vervangen van een woord door een eenvoudiger synoniem. Het model kan dus niet aangeven waarom het juist dat woord heeft vervangen door dat specifieke synoniem. Deze AI-toepassingen vallen onder de categorie van *supervised learning* en het model leert handelingen uit de data waarop het is

getraind. Dit is echter problematisch, aangezien Xu e.a. (2015) benadrukt dat veel toepassingen voor tekstvereenvoudiging geen rekening houden met de doelgroep waarvoor ze zijn ontwikkeld.
Om dit probleem op te lossen, is het belangrijk om de eindgebruiker, in dit geval scholieren met dyslexie in het derde graad middelbaar onderwijs, de keuze te geven. Zoals beschreven in Gooding (2022), zijn er verschillende mogelijkheden. Bijvoorbeeld, de eindgebruiker moet de mogelijkheid hebben om te kiezen welke synoniemen de tekst lexicaal zullen aanpassen. Een alternatieve aanpak voor syntactische vereenvoudiging is om de scholier zelf zinnen te laten markeren die moeilijk te begrijpen zijn, zodat het systeem alleen de door de eindgebruiker aangegeven zinnen vereenvoudigt.

2.4.6. Problemen bij lexicale vereenvoudiging

- Acroniemen
- Homoniemen

2.4.7. Problemen bij syntactische vereenvoudiging

- Kerninhoud verliezen

2.5. Beschikbare software voor tekstvereenvoudiging

2.5.1. Toepassingen nu in het onderwijs beschikbaar
2.5.2. Online toepassingen

2.6. Pipeline voor tekstvereenvoudiging

2.6.1. Lexicale vereenvoudiging
2.6.2. Syntactische vereenvoudiging
2.6.3. Samenvatten

2.7. Evaluatiemetrieken

3

Methodologie

Etiam pede massa, dapibus vitae, rhoncus in, placerat posuere, odio. Vestibulum luctus commodo lacus. Morbi lacus dui, tempor sed, euismod eget, condimentum at, tortor. Phasellus aliquet odio ac lacus tempor faucibus. Praesent sed sem. Praesent iaculis. Cras rhoncus tellus sed justo ullamcorper sagittis. Donec quis orci. Sed ut tortor quis tellus euismod tincidunt. Suspendisse congue nisl eu elit. Aliquam tortor diam, tempus id, tristique eget, sodales vel, nulla. Praesent tellus mi, condimentum sed, viverra at, consectetuer quis, lectus. In auctor vehicula orci. Sed pede sapien, euismod in, suscipit in, pharetra placerat, metus. Vivamus commodo dui non odio. Donec et felis.

Etiam suscipit aliquam arcu. Aliquam sit amet est ac purus bibendum congue. Sed in eros. Morbi non orci. Pellentesque mattis lacinia elit. Fusce molestie velit in ligula. Nullam et orci vitae nibh vulputate auctor. Aliquam eget purus. Nulla auctor wisi sed ipsum. Morbi porttitor tellus ac enim. Fusce ornare. Proin ipsum enim, tincidunt in, ornare venenatis, molestie a, augue. Donec vel pede in lacus sagittis porta. Sed hendrerit ipsum quis nisl. Suspendisse quis massa ac nibh pretium cursus. Sed sodales. Nam eu neque quis pede dignissim ornare. Maecenas eu purus ac urna tincidunt congue.

Donec et nisl id sapien blandit mattis. Aenean dictum odio sit amet risus. Morbi purus. Nulla a est sit amet purus venenatis iaculis. Vivamus viverra purus vel magna. Donec in justo sed odio malesuada dapibus. Nunc ultrices aliquam nunc. Vivamus facilisis pellentesque velit. Nulla nunc velit, vulputate dapibus, vulputate id, mattis ac, justo. Nam mattis elit dapibus purus. Quisque enim risus, congue non, elementum ut, mattis quis, sem. Quisque elit.

Maecenas non massa. Vestibulum pharetra nulla at lorem. Duis quis quam id lacus dapibus interdum. Nulla lorem. Donec ut ante quis dolor bibendum condimentum. Etiam egestas tortor vitae lacus. Praesent cursus. Mauris bibendum pede at elit. Morbi et felis a lectus interdum facilisis. Sed suscipit gravida turpis. Nulla at

lectus. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Praesent nonummy luctus nibh. Proin turpis nunc, congue eu, egestas ut, fringilla at, tellus. In hac habitasse platea dictumst.

Vivamus eu tellus sed tellus consequat suscipit. Nam orci orci, malesuada id, gravida nec, ultricies vitae, erat. Donec risus turpis, luctus sit amet, interdum quis, porta sed, ipsum. Suspendisse condimentum, tortor at egestas posuere, neque metus tempor orci, et tincidunt urna nunc a purus. Sed facilisis blandit tellus. Nunc risus sem, suscipit nec, eleifend quis, cursus quis, libero. Curabitur et dolor. Sed vitae sem. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Maecenas ante. Duis ullamcorper enim. Donec tristique enim eu leo. Nullam molestie elit eu dolor. Nullam bibendum, turpis vitae tristique gravida, quam sapien tempor lectus, quis pretium tellus purus ac quam. Nulla facilisi.

4

Conclusie

Curabitur nunc magna, posuere eget, venenatis eu, vehicula ac, velit. Aenean ornare, massa a accumsan pulvinar, quam lorem laoreet purus, eu sodales magna risus molestie lorem. Nunc erat velit, hendrerit quis, malesuada ut, aliquam vitae, wisi. Sed posuere. Suspendisse ipsum arcu, scelerisque nec, aliquam eu, molestie tincidunt, justo. Phasellus iaculis. Sed posuere lorem non ipsum. Pellentesque dapibus. Suspendisse quam libero, laoreet a, tincidunt eget, consequat at, est. Nullam ut lectus non enim consequat facilisis. Mauris leo. Quisque pede ligula, auctor vel, pellentesque vel, posuere id, turpis. Cras ipsum sem, cursus et, facilisis ut, tempus euismod, quam. Suspendisse tristique dolor eu orci. Mauris mattis. Aenean semper. Vivamus tortor magna, facilisis id, varius mattis, hendrerit in, justo. Integer purus.

Vivamus adipiscing. Curabitur imperdiet tempus turpis. Vivamus sapien dolor, congue venenatis, euismod eget, porta rhoncus, magna. Proin condimentum pretium enim. Fusce fringilla, libero et venenatis facilisis, eros enim cursus arcu, vitae facilisis odio augue vitae orci. Aliquam varius nibh ut odio. Sed condimentum condimentum nunc. Pellentesque eget massa. Pellentesque quis mauris. Donec ut ligula ac pede pulvinar lobortis. Pellentesque euismod. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent elit. Ut laoreet ornare est. Phasellus gravida vulputate nulla. Donec sit amet arcu ut sem tempor malesuada. Praesent hendrerit augue in urna. Proin enim ante, ornare vel, consequat ut, blandit in, justo. Donec felis elit, dignissim sed, sagittis ut, ullamcorper a, nulla. Aenean pharetra vulputate odio.

Quisque enim. Proin velit neque, tristique eu, eleifend eget, vestibulum nec, lacus. Vivamus odio. Duis odio urna, vehicula in, elementum aliquam, aliquet laoreet, tellus. Sed velit. Sed vel mi ac elit aliquet interdum. Etiam sapien neque, convallis et, aliquet vel, auctor non, arcu. Aliquam suscipit aliquam lectus. Proin tincidunt magna sed wisi. Integer blandit lacus ut lorem. Sed luctus justo sed enim.

Morbi malesuada hendrerit dui. Nunc mauris leo, dapibus sit amet, vestibulum et, commodo id, est. Pellentesque purus. Pellentesque tristique, nunc ac pulvinar adipiscing, justo eros consequat lectus, sit amet posuere lectus neque vel augue. Cras consectetuer libero ac eros. Ut eget massa. Fusce sit amet enim eleifend sem dictum auctor. In eget risus luctus wisi convallis pulvinar. Vivamus sapien risus, tempor in, viverra in, aliquet pellentesque, eros. Aliquam euismod libero a sem. Nunc velit augue, scelerisque dignissim, lobortis et, aliquam in, risus. In eu eros. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Curabitur vulputate elit viverra augue. Mauris fringilla, tortor sit amet malesuada mollis, sapien mi dapibus odio, ac imperdiet ligula enim eget nisl. Quisque vitae pede a pede aliquet suscipit. Phasellus tellus pede, viverra vestibulum, gravida id, laoreet in, justo. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer commodo luctus lectus. Mauris justo. Duis varius eros. Sed quam. Cras lacus eros, rutrum eget, varius quis, convallis iaculis, velit. Mauris imperdiet, metus at tristique venenatis, purus neque pellentesque mauris, a ultrices elit lacus nec tortor. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent malesuada. Nam lacus lectus, auctor sit amet, malesuada vel, elementum eget, metus. Duis neque pede, facilisis eget, egestas elementum, nonummy id, neque.

A

Onderzoeksvoorstel

Samenvatting

Tekstvereenvoudiging helpt scholieren met dyslexie in het derde graad middelbaar onderwijs bij hun lees- en verwerkingssnelheid. Artificiële intelligentie kan dit proces automatiseren. Ingewikkelde woordenschat en een lange zinsopbouw hinderen scholieren met dyslexie van een derde graad middelbaar onderwijs bij het lezen van wetenschappelijke papers. Vlaamse middelbare scholen ontbreken de toepassingen om specifiek scholieren met dyslexie van een derde graad middelbaar onderwijs te ondersteunen bij het lezen van wetenschappelijke papers, ondanks de steun en initiatief van de Vlaamse Overheid. Dit onderzoek wijst aan hoe de inhoud van wetenschappelijke artikelen automatisch met kunstmatige intelligentie kan worden vereenvoudigd, specifiek gericht op de noden van een scholier met dyslexie in het derde graad middelbaar onderwijs. Een technische analyse van relevante vakgebieden wordt uitgevoerd, gevolgd door een veldonderzoek bij Belgische informaticabedrijven en de ontwikkeling van een AI-gestuurde pipeline voor tekstvereenvoudiging. De huidige toepassingen voor tekstvereenvoudiging in het onderwijs zijn niet verzorgd voor scholieren met dyslexie in het derde graad middelbaar onderwijs. Ze zijn bedoeld om door een breed publiek in het lager en middelbaar onderwijs gebruikt te worden. Internationale AI-toepassingen omvatten een vertaalfase, maar afwijkingen van de kerninhoud moeten in acht worden genomen door software-ontwikkelaars. De ontwikkeling van een pipeline voor tekstvereenvoudiging met standaard modellen staat nog in de beginfase. Software-ontwikkelaars moeten aangepaste transformers ontwikkelen om te voldoen aan deze noden.

A.1. Introductie

Het Vlaamse middelbaar onderwijs staat nu op barsten, aangezien leraren en scholieren worden overspoeld door werkdruk en stress. Bovendien is de derde graad van het middelbaar onderwijs een belangrijke mijlpaal voor de verdere loopbaan van leerlingen, hoewel deze het moeilijk hebben om grip te krijgen op de vaktiteratuur binnen STEM-vakken (Dapaah & Maenhout, 2022). Het STEM-agenda[1] van de Vlaamse Overheid bestaat uit aandachtspunten om het STEM-onderwijs tegen 2030 aantrekkelijker te maken, door de ondersteuning voor zowel leerkrachten als scholieren te verbeteren. Toch wordt het aanpakken van de steeds complexere wetenschappelijke taal, zoals beschreven in Barnett en Doubleday (2020), niet als prioriteit beschouwd binnen de STEM-agenda. Het vereenvoudigen van wetenschappelijke artikelen is tijd- en energie-intensief. Gelukkig biedt geautomatiseerde en adaptieve tekstvereenvoudiging een baanbrekende oplossing om deze last te verlichten.

Dit onderzoek achterhaalt hoe de inhoud van een wetenschappelijke artikel op een geautomatiseerde wijze vereenvoudigd kan worden, specifiek gericht op de verschillende behoeften van scholieren met dyslexie in de derde graad middelbaar onderwijs. Hierbij wordt gestart met een theoretische basis voor tekstvereenvoudiging en een literatuurstudie naar welke uitdagingen een dergelijke toepassing in acht moet nemen. In een vervolgstap wordt met een veldonderzoek gekeken naar bestaande AI toepassingen voor tekstvereenvoudiging in Nederlandstalige en Engelstalige teksten. Hierna beschrijft het onderzoek een pipeline voor geautomatiseerde tekstvereenvoudiging en staat het stil bij de verschillende metrieken om een vereenvoudigde tekst te beoordelen. Daarna vindt een vergelijkende studie plaats tussen de vereenvoudigde tekstinhoud van verschillende aangehaalde toepassingen, die beoordeeld wordt met behulp van enquêtes en statistische metrieken. Tot slot worden de resultaten van het onderzoek gebruikt om inzicht te krijgen in hoe wetenschappelijke artikelen op een geautomatiseerde en adaptieve manier vereenvoudigd kunnen worden, specifiek voor scholieren met dyslexie in het derde graad middelbaar onderwijs. Dit leidt tot verdere ontwikkeling voor AI-ontwikkelaars om een bruikbare toepassing te creëren voor gebruik in het onderwijs.

A.2. State-of-the-art

De voorbije tien jaar is kunstmatige intelligentie (AI) sterk verder ontwikkeld. De toename in kennis zorgde voor nieuwe toepassingen (Vasista, 2022). Tekstvereenvoudiging vloeide hier uit voort. Momenteel bestaan er al robuuste toepassingen

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

die teksten kunnen vereenvoudigen, zoals Resoomer[2], Paraphraser[3] en Prepostseo[4]. Binnen het kader van tekstvereenvoudiging is er bestaande documentatie beschikbaar waar onderzoekers het voordeel van toegankelijkheid aanhalen, maar volgens Gooding (2022) ontbreken deze toepassingen de extra noden die scholieren met dyslexie in het derde graad middelbaar onderwijs vereisen.
Shardlow (2014) haalt aan dat het algemene doel van tekstvereenvoudiging is om ingewikkelde bronnen toegankelijker te maken. Het zorgt voor verkorte teksten zonder de kernboodschap te verliezen. Siddharthan (2014) haalt verder aan dat tekstvereenvoudiging op één van drie manieren gebeurt. Er is conceptuele vereenvoudiging 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 volgens Siddharthan (2014) in staat om ingewikkelde woorden te vervangen door eenvoudigere synoniemen of zinnen te verkorten zodat ze sneller leesbaar zijn.
Tekstvereenvoudiging behoort tot de zijtak van natuurlijke taalverwerking (NLP) in kunstmatige intelligentie. NLP omvat methodes om, door machinaal leren, menselijke teksten om te zetten in tekst voor machines. Documenten vereenvoudigen met NLP kan volgens Chowdhary (2020) 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 tekstvereenvoudiging komt volgens (Sciforce, 2020) het meeste voor. 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.
Volgens Plavén-Sigray e.a. (2017) houden onderzoekers zich vaak op in hun eigen taalbubbel, wat negatieve gevolgen heeft voor de leesbaarheid van een wetenschappelijk artikel. Bovendien vormt de stijgende trend van het gebruik van acroniemen Barnett en Doubleday (2020) een extra hindernis. Donato e.a. (2022) wijst uit dat een van de redenen waarom scholieren met dyslexie in het middelbaar onderwijs van richting veranderen, te wijten is aan onbegrijpelijke teksten.
Het onderzoek van Franse wetenschappers
Gala en Ziegler (2016) illustreert dat manuele tekstvereenvoudiging 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 kin-

[2]https://resoomer.com/nl/
[3]https://www.paraphraser.io/nl/tekst-samenvatting
[4]https://www.prepostseo.com/tool/nl/text-summarizer

deren 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 visuele weergave van tekst beïnvloedt de leessnelheid bij scholieren met dyslexie. 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 verder aan dat teksten voor deze unieke noden aanpassen tijdrovend is, dus tekstvereenvoudiging door kunstmatige intelligentie kan een revolutionaire oplossing bieden. De Universiteit van Kopenhagen is met bovenstaande idee aan de slag gegaan. Onderzoekers Bingel e.a. (2018) hebben gratis software ontwikkeld, genaamd Hero[5], om tekstvereenvoudiging 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.
Roldós (2020) haalt aan dat NLP in de laatste decennia volop in ontwikkeling is, maar ontwikkelaars botsen nog op uitdagingen. Het gaat om zowel interpretatie- als dataproblemen bij AI machines. Het onderzoek haalt twee punten aan. 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.
Het onderzoek van Sciforce (2020) haalt aan dat het merendeel van NLP-toepassingen Engelstalige invoer gebruikt. 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. Hetzelfde onderzoek bewijst dat het vertalen van gelijkaardige talen, zoals Duits en Nederlands, een minimaal verschil opleverd.
Voor scholieren met dyslexie in het derde graad middelbaar onderwijs bestaan digitale hulpmiddelen die voor een betere visuele presentatie zorgen van teksten. De Vlaamse overheid leent gratis abonnementen uit voor voorlees- en schrijfsoftware. De voornaamste zijn SprintPlus[6], Alinea[7] en Kurzweil3000[8]. Vlaamse scholieren met dyslexie in het middelbaar onderwijs kunnen voor deze software een gratis abonnement of licentie aanvragen. Al bieden de vijf softwarepakketten elk een samenvattingsfunctie aan, echter ligt de focus op spreek- en luisterfuncties waarbij het samenvatten en markeren van tekst als extra wordt gehouden.

[5]https://beta.heroapp.ai/
[6]https://www.sprintplus.be/
[7]https://sensotec.be/product/alinea-suite/
[8]https://sensotec.be/product/kurzweil-3000/

ChatGPT[9] van OpenAI is een *chatbot* gebouwd op het GPT-3 model. Het GPT-3 model omvat meer dan vijf miljard verschillende woorden, wat het revolutionair maakt voor AI taaltoepassingen. Nadelig moet de *chatbot* via de online toepassing expliciet gevraagd worden om tekst te kunnen vereenvoudigen. Verhoeven (2023) haalt aan dat toepassingen zoals ChatGPT een wondermiddel zijn om de werklast van routinematig en boilerplate werk te verminderen in het onderwijs. Toepassingen ontwikkelen met het GPT-3 model is mogelijk, al is de API van GPT-3 enkel tegen betaling beschikbaar. Readable[10] is een Engelstalige AI toepassing dat zinnen beoordeeld met leesbaarheidsformules. Bij beide tools is het enkel mogelijk om tekst op de webpagina te plakken, dus er kunnen geen PDF-documenten of scans worden geüpload en eenzelfde werking verwachten.

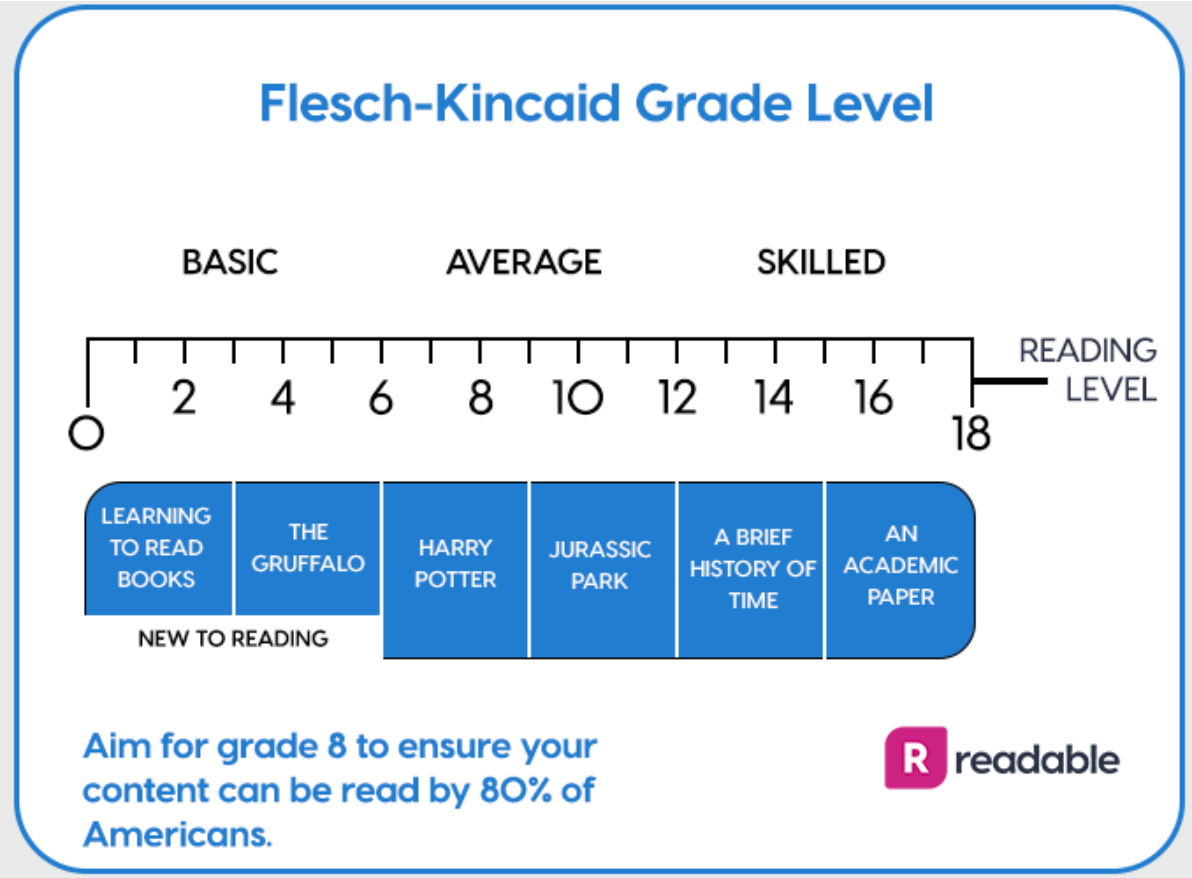
Vlaanderen heeft weinig zicht op de geïmplementeerde AI software in scholen. Dit werd vastgesteld door (Martens e.a., 2021a), een samenwerking tussen de Vlaamse universiteiten en overheid voor kunstmatige intelligentie. Vergeleken met andere Europese landen, maakt België het minst gebruik van leerling-georiënteerde hulpmiddelen. Degenen die wel gebruikt worden, zijn vooral 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).

Python staat bovenaan de lijst van programmeertalen voor NLP-toepassingen. Volgens het onderzoek van Thangarajah (2019) is dit te wijten aan de eenvoudige syntax, kleine leercurve en grote beschikbaarheid van kant-en-klare bibliotheken. Moeilijke wiskundige berekeningen of statistische analyses kunnen worden uitgevoerd door middel van één lijn code. Een artikel van Malik (2022) haalt de twee meest voorkomende aan, namelijk NLTK[11] en Spacy[12].

Iedere soort tekstvereenvoudiging omvat verschillende fases. Het onderzoek van Shardlow (2014) wijst uit dat een pipeline voor lexicale vereenvoudiging uit vier fases bestaat. Een *proof-of-concept* genaamd *Deep Martin*[13] bouwt verder op dit theoretisch concept. Hun pipeline maakt gebruik van *custom transformers* om invoertekst om te zetten naar een vereenvoudigde versie van de tekstinhoud.

Garbacea e.a. (2021) benadrukken dat AI ontwikkelaars te weinig aandacht besteden aan het achterhalen waarom een woord of zin moet worden aangepast. Zij halen twee ethische aspecten van AI taaltoepassingen aan de eindgebruiker moet worden meegegeven. Allereerst moet de toepassing meegeven waarom een zin of woord is aangepast. De moeilijkheidsgraad van de woord of de zin moet worden bewezen door het model. Iavarone e.a. (2021) haalt zo een methode aan om de moeilijkheidsgraad te bepalen. In dit onderzoek werden regressiemodellen ingezet om een gemiddelde moeilijkheidspercentage te berekenen per zin. Verder

[9] https://chat.openai.com/chat
[10] https://readable.com/
[11] https://www.nltk.org/
[12] https://spacy.io/
[13] https://github.com/chrislemke/deep-martin



Figuur (A.1)
(Readable, 2021)

haalt Garbacea e.a. (2021) om de complexe delen van een tekst te markeren. Hiervoor worden *lexical of deep learning* methoden aangehaald.
Er is een tactvolle aanpak nodig om een vereenvoudigde tekst met AI te beoordelen. 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. Daarnaast wijst het onderzoek van Readable (2021) uit dat de Flesch-Kincaid leesbaarheidstest een manier aanbiedt om vereenvoudigde tekstinhoud te beoordelen, zonder de nood van vooraf getrainde modellen. Met de Python-library *textstat*[14] kan deze score eenvoudig worden berekend.

A.3. Methodologie

Er wordt een *mixed-methods* onderzoek uitgevoerd om te bepalen of een AI toepassing de tekstinhoud van een wetenschappelijke paper op maat van de noden

[14]https://pypi.org/project/textstat/

voor een scholier met dyslexie in het derde graad middelbaar onderwijs kan vereenvoudigen. Het onderzoek houdt zes fases in.
De eerste fase is het proces van tekstvereenvoudiging beschrijven, waaronder een omschrijving van het begrip en de verschillende soorten van tekstvereenvoudiging met AI. 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 tekstvereenvoudiging bij scholieren met dyslexie van het derde graad middelbaar 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, aan de hand van een technische uitleg, mogelijke nadelen en tekortkomingen van AI software bij tekstvereenvoudiging aan het licht.
De vierde fase omvat een toelichting over beschikbare AI toepassingen voor tekstvereenvoudiging. 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. Het resultaat van deze fase is een longlist van alle beschikbare AI toepassingen die teksten kunnen vereenvoudigen.
De vijfde fase omschrijft de technische uitwerking van een pipeline voor tekstvereenvoudiging, alsook een shortlist van metrieken om de vereenvoudigde tekstinhoud te evalueren. Er zal een tekstvereenvoudigingspipeline worden ontwikkeld met beschikbare kant-en-klare bibliotheken, *transformers* en algoritmen. Het resultaat van deze fase is een pipeline opgebouwd in de programmeertaal Python.
De zesde fase bestaat uit een toelichting van de beschikbare evaluatiemetrieken om vereenvoudigde tekst te kunnen beoordelen. Het resultaat is een shortlist van alle evaluatiecriteria waaraan de uitvoertekst van een tekstvereenvoudigingstoepassing moet voldoen.
De zevende en laatste fase omvat een vergelijkende studie van de gevonden AI toepassingen die tekst vereenvoudigen en de pipeline. De tekstinhoud van wetenschappelijke papers, die in een derde graad middelbaar onderwijs worden gebruikt, dienen hier als invoertekst voor de evaluatie. De subjectieve test gebeurt aan de hand van een enquête 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 tekstvereenvoudiging.

A.4. Verwacht resultaat, conclusie

Er wordt verwacht dat de software, die nu in het onderwijs wordt ingezet, niet voldoet aan de noden van een scholier met dyslexie in het derde graad middelbaar onderwijs. Er wordt onvoldoende rekening gehouden met het adaptieve aspect. Bestaande internationale AI toepassingen bieden een gelijkwaardige oplossing, al steekt ChatGPT met het GPT-3 model boven de rest uit. Met dit model kan er een krachtige applicatie worden opgebouwd. Het vertalen van de vereenvoudigde tekstinhoud bij een internationale AI toepassing kan afwijken van de oorspronkelijke context.

Er zijn te weinig kant-en-klare algoritmen en modellen beschikbaar om een pipeline voor tekstvereenvoudiging op te zetten, gericht op scholieren met dyslexie in het middelbaar onderwijs. De pipeline is moeilijk af te stemmen op de specifieke noden van deze doelgroep. Er is een behoefte aan aangepaste transformers om bevredigende resultaten te bereiken. Bovendien is er een gebrek aan Nederlandstalige word embeddings die de complexiteit van elk woord kunnen bijhouden en aan kant-en-klare modellen die de inhoud van wetenschappelijke papers kunnen vereenvoudigen. Word embeddings uit een Germaanse taal gebruiken, gevolgd door vertaling naar het Nederlands is wel een acceptabel alternatief.

Bibliografie

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

Bulté, B., Sevens, L. & Vandeghinste, V. (2018). Automating lexical simplification in Dutch. *Computational Linguistics in the Netherlands Journal*, *8*, 24–48. https://clinjournal.org/clinj/article/view/78

Canning, Y., Tait, J., Archibald, J. & Crawley, R. (2000). Cohesive Generation of Syntactically Simplified Newspaper Text. In P. Sojka, I. Kopeček & K. Pala (Red.), *Text, Speech and Dialogue* (pp. 145–150). Springer Berlin Heidelberg.

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

Coster, W. & Kauchak, D. (2011). Learning to Simplify Sentences Using Wikipedia. *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, 1–9. https://aclanthology.org/W11-1601

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.

Dapaah, J. & Maenhout, K. (2022, juli 8). *Iedereen heeft boter op zijn hoofd* (D. Standaard, Red.). https://www.standaard.be/cnt/dmf20220607_97763592

De Belder, M.-F., Jan; Moens. Text simplification for children. eng. In: ACM; New York, 2010.

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.

Eisenstein, J. (2019). *Introduction to Natural Language Processing*. MIT Press. https://books.google.be/books?id=72yuDwAAQBAJ

El-Kassas, W. S., Salama, C. R., Rafea, A. A. & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, *165*, 113679. https://doi.org/https://doi.org/10.1016/j.eswa.2020.113679

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.

Garbacea, C., Guo, M., Carton, S. & Mei, Q. (2021). Explainable Prediction of Text Complexity: The Missing Preliminaries for Text Simplification. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1086–1097. https://doi.org/10.18653/v1/2021.acl-long.88

Gooding, S. (2022). On the Ethical Considerations of Text Simplification. *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, 50–57. https://doi.org/10.18653/v1/2022.slpat-1.7

Iavarone, B., Brunato, D. & Dell'Orletta, F. (2021). Sentence Complexity in Context. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 186–199. https://doi.org/10.18653/v1/2021.cmcl-1.23

Kandula, S., Curtis, D. & Zeng-Treitler, Q. (2010). A semantic and syntactic text simplification tool for health content. *AMIA annual symposium proceedings*, *2010*, 366.

Malik, R. S. (2022, juli 4). *Top 5 NLP Libraries To Use in Your Projects* (T. AI, Red.). https://towardsai.net/p/l/top-5-nlp-libraries-to-use-in-your-projects

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 30 maart 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 1 april 2022, van https://data-en-maatschappij.ai/nieuws/school-innovation-forum-2021

Niemeijer, A., Frederiks, B., Riphagen, I., Legemaate, J., Eefsting, J. & Hertogh, C. (2010). Ethical and practical concerns of surveillance technologies in residential care for people with dementia or intellectual disabilities: an overview of the literature. *Psychogeriatrics*, *22*(7), 1129–1142. https://doi.org/10.1017/S1041610210000037

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.

Readable. (2021). *Flesch Reading Ease and the Flesch Kincaid Grade Level*. https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/

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 1 april 2022, van https://monkeylearn.com/blog/natural-language-processing-challenges/

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

Shardlow, M. (2014). A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014*, *4*(1). https://doi.org/10.14569/SpecialIssue.2014.040109

Siddharthan, A. (2006). Syntactic Simplification and Text Cohesion. *Research on Language and Computation*, *4*(1), 77–109. http://oro.open.ac.uk/58888/

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

Sikka, P. & Mago, V. (2020). A Survey on Text Simplification. *CoRR*, *abs/2008.08612*. https://arxiv.org/abs/2008.08612

Sohom, G., Ghosh; Dwight. (2019). *Natural Language Processing Fundamentals*. Packt Publishing. https://medium.com/analytics-vidhya/natural-language-processing-basic-concepts-a3c7f50bf5d3

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

Thangarajah, V. (2019). Python current trend applications-an overview.

Vasista, K. (2022). Evolution of AI Design Models. *Central Asian Journal of Theoretical and Applied Science*, *3*(3), 1–4. https://www.cajotas.centralasianstudies.org/index.php/CAJOTAS/article/view/415

Verhoeven, W. (2023, februari 8). *Applaus voor de studenten die ChatGPT gebruiken* (Trends, Red.). https://trends.knack.be/economie/bedrijven/applaus-voor-de-studenten-die-chatgpt-gebruiken/article-opinion-1934277.html?cookie_check=1676034368

Xu, W., Callison-Burch, C. & Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, *3*, 283–297.

Zeng, Q., Kim, E., Crowell, J. & Tse, T. (2005). A Text Corpora-Based Estimation of the Familiarity of Health Terminology. In J. L. "Oliveira, V. Maojo, F. Martín-Sánchez & A. S. Pereira (Red.), *Biological and Medical Data Analysis* (pp. 184–192). Springer Berlin Heidelberg.