Menselijke evaluatie van geautomatiseerde tekstvereenvoudiging door middel van crowdsourcing

Vincent Vandeghinste, Job van Doeslaar en Bram Vanroy
Instituut voor de Nederlandse Taal
Leiden, Nederland
voornaam.familienaam@ivdnt.org

Abstract

Dit document beschrijft de crowdsourcing applicatie die bij het Instituut voor de Nederlandse Taal (INT) gebouwd werd met als doel een goudstandaard referentieset te creëren die kan dienen om automatische tekstvereenvoudiginssystemen automatisch te evalueren.

1 Introduction

Dit document beschrijft de toepassing die gemaakt werd ter uitvoering van het pilootpproject "Duidelijke Taal", en past in het beleid dat de Taalunie voert voor begrijpelijk taal bij de overheid en in andere maatschappelijke sectoren (recht, zorg). Voor dat beleidsdoel werkt de Taalunie sinds 2016 intensief samen met het ministerie van Binnenlandse Zaken en Koninkrijksrelaties, de Vlaamse overheid en verschillende veldorganisaties in Nederland en Vlaanderen. Zie voor meer informatie https://taalunie.org/dossiers/44/begrijpelijke-overheidstaal.

Het project kwam tot stand na een aantal gesprekken tussen het Instituut voor de Nederlandse Taal en de Taalunie, waarbij vastgesteld werd dat er, om een datagedreven state-of-the-art benadering voor automatische tekstomzetting naar *duidelijke taal* voor het Nederlands te ontwikkelen, een duidelijk gebrek aan manueel gevalideerde vereenvoudigde data is, hetzij voor trainings- of voor evaluatiedoeleinden.

Tot nu toe was de aandacht gericht op het door de schrijvers zelf laten vereenvoudigen van overheidsteksten, zodat die voor iedereen toegankelijk en goed te begrijpen zijn. De huidige, snelle ontwikkelingen op het gebied van AI bieden mogelijkheden om dit werk (ook) door machines te laten uitvoeren. In dit pilootproject onderzoeken we in hoeverre dit nu kan en hoe goed dat gaat. We mikken hierbij hoofdzakelijk op overheidscommunicatie, omdat die in principe in duidelijke taal opgesteld moet worden, zodat zo veel mogelijk mensen de inhoud ervan kunnen begrijpen.

Metrieken voor de automatische evaluatie van tekstvereenvoudiging worden met verschillende uitdagingen geconfronteerd. Eén van deze uitdagingen is dat ze vaak referentievereenvoudigingen op basis van een goudstandaard vereisen waarmee de automatische vereenvoudigingen worden vergeleken. Dit is het geval voor statistieken zoals SARI (Xu et al., 2016a), BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BLEURT (Sellam et al., 2020) en BERTscore (Zhang et al., 2020). Voor onderzoek naar Nederlandse vereenvoudiging zijn er nauwelijks referentiesets beschikbaar en als die er wel zijn, bestaan deze vaak uit (automatische) vertalingen van Engelse testsets, zoals in Seidl and Vandeghinste (2024).

Gerelateerd werk over tekstvereenvoudiging, over Nederlandse datasets voor tekstvereenvoudiging en over menselijke beoordeling hiervan wordt beschreven in sectie 2.

Welke data we genomen hebben voor deze studie en hoe synthetische vereenvoudigingen gemaakt werden wordt beschreven in sectie 3.

Om een grootschalige goudstandaard-testset te creëren, hebben we een crowdsourcing-applicatie opgezet waarin gebruikers automatische vereenvoudigingen handmatig konden evalueren. Deze toepassing wordt beschreven in 4. Beoordeling gebeurde op deze dimensies:

- *Fluency*: Gebruikers wordt gevraagd tekst te beoordelen op een schaal met als uitersten `fout Nederlands` en `goed Nederlands`.
- *Eenvoud*: Gebruikers wordt gevraagd tekst te beoordelen op een schaal met als uitersten `eenvoudig` en `complex`.
- *Duidelijkheid*: Gebruikers wordt gevraagd om bij een tekstpaar aan te geven welke tekst duidelijker is op een schaal met als uitersten `tekst A` en `tekst B`.

- *Accuraatheid*: Gebruikers wordt gevraagd of de vereenvoudigde tekst dezelfde betekenis heeft als de originele tekst, op een shcaal met als uitersten `andere betekenis` en `zelfde betekenis`.

Om de gebruikersbetrokkenheid te stimuleren hebben we een aantal gamification elementen toegevoegd, zoals een score op basis van inspanning, snelheid en ingeschatte correctheid, en meerdaagse streaks en een gebruikersscorebord. Dit wordt in detail beschreven in sectie 5.

De toepassing was beschikbaar op `https://duidelijketaal.ivdnt.org/` maar is niet langer actief.

2 Gerelateerd Werk

2.1 Automatische tekstvereenvoudiging

Tot voor kort werd automatische tekstvereenvoudiging vaak uitgevoerd met technieken uit de machinevertaling (Xu et al., 2016b), waarbij de originele tekst als de bron werd gebruikt en de aangepaste tekst als doeltaal. Wubben (2013) heeft dit geprobeerd voor het Nederlands met statistische machinevertaling. Meer recentelijk zijn voor Nederlands grote taalmodellen, zoals het T5-model (Raffel et al., 2020), verfijnd met parallelle data voor tekstvereenvoudiging (Seidl and Vandeghinste, 2024).

Er zijn een aantal toepassingen voor automatische tekstvereenvoudiging voor het Nederlands op de markt. Dit houden we bij op `https://kdutch.ivdnt.org/wiki/Text_simplification/nl`.

2.2 Data voor Nederlandse tekstvereenvoudiging

We proberen bestaande datasets voor Nederlandse tekstvereenvoudiging bij te houden op de website van het CLARIN Knowledge Centre for Dutch, op `https://kdutch.ivdnt.org/wiki/Simplification_Data/nl`.

Handmatig gemaakte dataset Tot nu toe hebben we slechts één handmatige dataset voor Nederlands geïdentificeerd, gemaakt door Vlantis et al. (2024), met betrekking tot gemeentelijke teksten.

Automatisch vertaalde datasets Seidl and Vandeghinste (2024) heeft de ASSET dataset van Alva-Manchego et al. (2020a) en de WikiLarge dataset van Zhang and Lapata (2017) automatisch vertaald en deze gebruikt om het taalmodel te verfijnen. Een andere automatisch vertaalde dataset is de SimpleWiki dataset, gemaakt door het Nationaal Forensisch Instituut.

Synthetische datasets Nu we in het tijdperk van generatieve grote taalmodellen zijn, kunnen we ook dergelijke modellen vragen om aangepaste teksten te leveren. Dit is gedaan voor de UWV Leesplank NL wikipedia dataset,[1] en voor de data die gebruikt is in Van de Velde et al. (2023), en dat is ook wat we in dit onderzoek gedaan hebben.

Vergelijkbaar corpus van makkelijke en reguliere niveauteksten Er is een vergelijkbaar corpus beschikbaar van eenvoudige krantenartikels uit de Wablieft krant en artikels over dezelfde onderwerpen of gebeurtenissen uit de Standaard (Vanackere and Vandeghinste, 2022).

2.3 Menselijke beoordeling bij tekstvereenvoudiging

In hun beschrijving van de stand van zaken op dit gebied beschrijft Alva-Manchego et al. (2020b) hoe 3-punts of 5-punts Likert-schalen vaak worden gebruikt om grammaticale correctheid (ook wel *fluency* genoemd), betekenisbehoud (ook wel *adequacy* genoemd) en eenvoud te beoordelen. Fluency wordt alleen beoordeeld voor de output, terwijl adequacy en eenvoud worden beoordeeld op basis van het tekstpaar.

Deze aanpak wordt verfijnd in Alva-Manchego et al. (2020a) waar de beoordeling plaatsvindt op een continue schaal (0-100) met betrekking tot adequacy, fluency en eenvoud. Het gebruik van continue intervalschalen bij crowdsourcing van menselijke evaluaties is gangbare praktijk in machinevertaling (Bojar et al., 2018), omdat dit resulteert in hogere niveaus van interannotatorconsistentie dan ordinale Likert-schalen. In onze aanpak gebruiken we eveneens continue intervalschalen (zonder numerieke indicator).

3 De dataset

We hebben een testset van 6.986 zinsparen gemaakt, waarbij we de originele zinnen hebben geselecteerd uit de WRPEI-component van het SONAR-corpus (Oostdijk et al., 2013), die uit websites bestaat. We hebben voor dit onderdeel gekozen omdat we veronderstellen dat het taal bevat die gericht is op

[1] `https://huggingface.co/datasets/UWV/Leesplank_NL_wikipedia_simplifications`

het brede publiek en daarom wordt verwacht dat het duidelijke taal is.

We hebben enkel zinnen geselecteerd met meer dan 10 en minder dan 50 woorden, die minimaal één werkwoord bevatten, met een Leesindex-coëfficiënt (Brouwer, 1963) hoger dan 60, en die uit meer dan één *clause* bestaan. Hiermee bedoelen we dat de zin ook minstens één bijzin bevat. Dit gebeurde met de tooling zoals beschreven in Vandeghinste and Bulté (2019), om zinnen met in eerste instantie een redelijke complexiteit te vereenvoudigen.

Deze zinnen werden automatisch vereenvoudigd met behulp van GPT-4 met hetzelfde prompt als gebruikt in het UWV/Leesplank-project dat beschikbaar is op HuggingFace:

> "Simplify a Dutch paragraph directly into a single, clear, and engaging text suitable for adult readers that speak Dutch as a second language, using words from the 'basiswoordenlijst Amsterdamse kleuters.' Maintain direct quotes, simplify dialogue, explain cultural references, idioms, and technical terms naturally within the text. Adjust the order of information for improved simplicity, engagement, and readability. Attempt to not use any commas or diminutives."

De resulterende dataset wordt beschikbaar gesteld voor download in de CLARIN-infrastructuur van het Instituut voor de Nederlandse Taal, op https://hdl.handle.net/10032/tm-a2-y7.

Deze parallelle data kan via een dashboard, toegankelijk voor de administratoren, opgeladen worden in de applicatie in de vorm van een csv bestand.

3.1 Eigenschappen van de synthetische dataset

Uit de totale dataset werd een selectie gemaakt van 1071 parallelle paragrafen. Kenmerken van het bronmateriaal en de synthetische aanpassing worden gegeven in Tabel 1, met gemiddelde scores voor bron- en doelmateriaal evenals de significantiegraad van het verschil tussen de bron- en doeltaal aan de hand van de p-waarde bij een t-test. Voor de operationalisering van de kenmerken verwijzen we naar Vandeghinste and Bulté (2019).

Zoals we kunnen zien voldoen de synthetische *vereenvoudigingen* niet aan de standaardkenmerken die we zouden verwachten van vereenvoudigingen,

Kenmerk	Bron	Doel	p
tekstlengte	17.25	35.17	<.01
woorden/zin	16.53	20.22	<.01
gem. clauselengte	10.95	10.56	<.01
bijzinnen	0.81	1.58	<.01
bijzinnen/clause	0.35	0.31	<.01
bijzinlengte	7.53	7.25	>.05
woorden/finiet ww	8.92	8.63	<.05
NP-lengte	4.02	3.87	>.05
dep/hoofd	1.76	1.79	<.05
boomdiepte	7.17	7.53	<.01
type token ratio	0.91	0.72	<.01
Guiraud	3.68	3.39	<.01
Flesch Reading Ease	62.21	53.68	<.01
Flesch Douma	75.19	67.30	<.01
CLIB	75.47	83.11	<.01
CILT	78.00	79.06	<.01
Brouwer	60.77	49.52	<.01
karakters/w	4.74	4.78	>.05
lettergr/w	1.51	1.57	>.05
verwervingsleeftijd	9.78	8.61	>.05
inhoudswoorden/woord	0.54	0.53	>.05
concreetheid	2.89	2.85	>.05
77% ratio	78.57	84.15	<.01
lange woorden	0.05	0.05	>.05
vz ratio	0.14	0.12	<.01
spec ratio	0.02	0.02	>.05
vnw ratio	0.08	0.10	<.01
n ratio	0.21	0.20	<.01
tw ratio	.04	.04	>.05
lid ratio	0.11	0.10	<.05
vg ratio	0.05	0.05	>.05
bw ratio	0.05	0.05	>.05
adj ratio	0.06	0.06	>.05
let ratio	0.09	0.09	>.05
ww ratio	0.16	0.17	>.05

Tabel 1: Eigenschappen van de beoordeelde dataset

of van reguliere versus eenvoudige of duidelijke taal, zoals vastgesteld in Vandeghinste and Bulté (2019). We zien dat bvb. de gemiddelde tekstlengte langer is en dat het aantal woorden per zin ook hoger is. Dit valt vermoedelijk te verklaren doordat het prompt expliciet vraagt om bepaalde termen te verklaren. Wel zien we dat gemiddelde lengte van een clause lager is, en dit geldt ook voor het aantal woorden per finiet werkwoord. Verschillen in NP-lengte en bijzinlengte zijn niet significant. We zien dat voor de maten van lexicale densiteit (type-token-ratio, Guiraud), er wel een indicatie is voor significante vereenvoudiging.

Voor de leesbaarheidsmetrieken (Flesch, CLIB, CILT, Brouwer) stellen we merkwaardig genoeg vast dat de synthetische vereenvoudigingen als moeilijker (lager) gescoord worden.

Voor de woordsoortratio's stellen we vast dat in de vereenvoudigde teksten het aantal voorzetsels significant lager is, en het aantal zelfstandige naamwoorden en voornaamwoorden significant hoger is.

De 6 best scorende features in classificatie van teksten als regulier vs. eenvoudig uit Vandeghinste and Bulté (2019) scoren hier allemaal contra-intuïtief.

We willen hierbij graag benadrukken dat deze eigenschappen over de hele dataset berekend werden en er dus geen rekening gehouden werd met de kwaliteit (vooral adequacy) van de vereenvoudiging. Verder onderzoek moet uitwijzen of dit ook het geval zou zijn als we deze waarden berekenen op enkel die gevallen waarbij de vereenvoudiging als accuraat beoordeeld werd.

In de toekomst kan er ook geëxperimenteerd worden met andere prompts, waarbij minder de nadruk op het verklaren gelegd wordt. Dit zou mogelijk heel andere waarden voor de kenmerken van de synthetische tekst geven.

4 De applicatie

4.1 Introductiescherm

Het introductiescherm, getoond in Figuur 1 nodigt de gebruiker uit deel te nemen. Om de toepassing zo laagdrempelig mogelijk te houden wordt de gebruiker nog niet gevraagd om in te loggen, zodat gebruikers al kunnen antwoorden zonder verdere vragen te moeten beantwoorden. De gebruiker heeft wel de mogelijkheid om onmiddelijk in te loggen, zodat de score opgehaald kan worden en er op basis van het profiel vragen aangeboden worden die de gebruiker nog niet eerder beantwoord heeft.

Figuur 1: Introductiescherm

4.2 Beoordelingsmodule

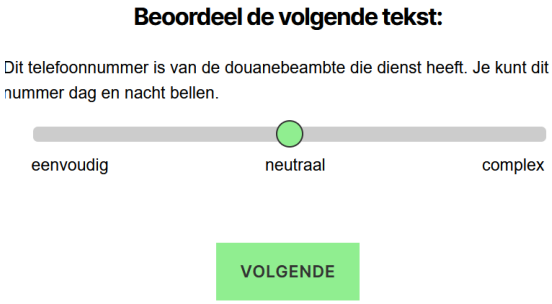
Na het klikken op *MEEDOEN* komt de gebruiker in de beoordelingsmodule terecht. Manuele beoordeling van automatisch gegenereerde vereenvoudigde zinnen kan best op een aantal dimensies gebeuren, zoals *eenvoud* (simplicity), *accuraatheid* (accuracy) en vlotheid (*fluency*). Voor de beoordeling van eenvoud hebben we twee taken gedefinieerd: de beoordeling van een zin op eenvoud (sectie 4.2.1) en het beoordelen van welke zin van een zinspaar duidelijker is (sectie 4.2.2). Accuraatheid wordt beschreven in sectie 4.2.4 en vlotheid wordt beschreven in sectie 4.2.3.

Er worden per sessie twintig vragen aangeboden, telkens vijf uit de vier verschillende taken.

4.2.1 Beoordelen van zinnen op hun eenvoud

In deze taak worden zinnen uit de dataset willekeurig gekozen en wordt de gebruiker gevraagd om de zin te beoordelen op eenvoud. Deze zinnen kunnen zowel uit de oorspronkelijke zinnen als uit de autmomatisch vereenvoudigde zinnen komen en garanderen een conditieblinde evaluatie.

Figuur 2 toont de slider die de gebruiker kan verschuiven op een as met als uitersten *eenvoudig* en *complex*. We vragen de gebruiker dus om een gradueel oordeel, die achter de schermen gemeten wordt op een intervalschaal met 100 eenheden. De slider begint op de *neutrale* positie met waarde 50.

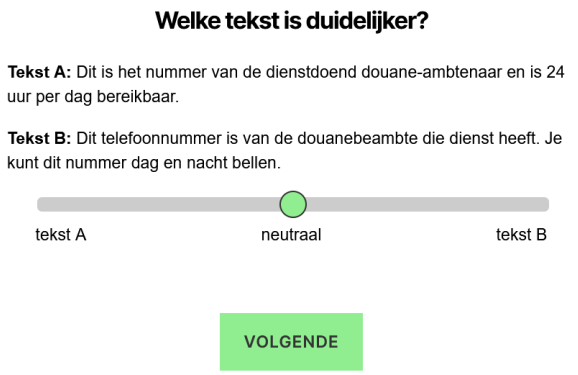


Figuur 2: Voorbeeld van een beoordelingsvraag met betrekking tot eenvoud

4.2.2 Paarsgewijze beoordeling van duidelijkheid

Bij de paarsgewijze beoordeling wordt een zinspaar uit de databank van zinnen met hun automatische vereenvoudiging gekozen en in willekeurige volgorde aan de gebruiker ter beoordeling aangeboden. Dit garandeert dat er geen volgorde-effect optreedt en de gebruiker conditieblind oordeelt.

Zoals te zien in Figuur 3 krijgt de gebruiker nu een slider te zien op een as met als uitersten *tekst A* en *tekst B*. Net zoals bij de andere taken gaat het hier om een intervalschaal met 100 eenheden. De slider begint op de *neutrale* positie met waarde 50.



Figuur 3: Voorbeeld van een paarsgewijze beoordelingssvraag

4.2.3 Zinsbeoordeling op *fluency*

Bij deze beoordeling worden zowel zinnen uit de oorspronkelijke als uit de automatisch vereenvoudigde set aangeboden en wordt aan de gebruiker gevraagd in welke mate deze zinnen *goed* of *fout* Nederlands zijn. Het gaat dus opnieuw om een conditieblinde beoordeling.

Zoals te zien in Figuur 4 krijgt de gebruiker nu een slider te zien op een as met als uitersten *fout Nederlands* en *goed Nederlands*. Net zoals bij de andere taken gaat het hier om een intervalschaal met 100 eenheden. De slider begint op de *neutrale* positie met waarde 50.

Figuur 4: Voorbeeld van een beoordeling omtrent *vlotheid*

4.2.4 Accuracy

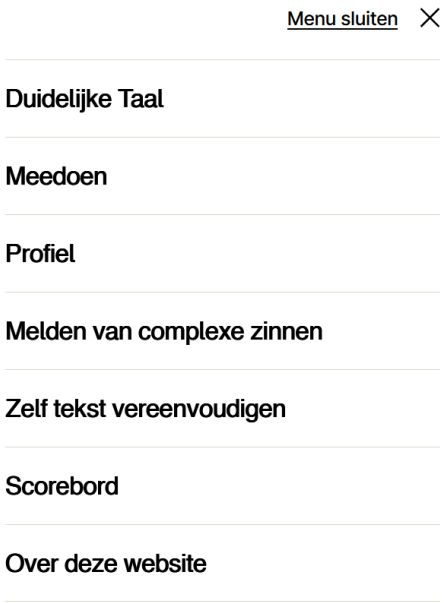
Bij deze beoordeling wordt het zinspaar aangeboden zodat de gebruiker kan oordelen of de inhoud en betekenis van de doelzin overeenkomt met die van de bronzin.

Zoals te zien in Figuur 5 krijgt de gebruiker nu een slider te zien op een as met als uitersten *andere betekenis* en *zelfde betekenis*. Net zoals bij de andere taken gaat het hier om een intervalschaal met 100 eenheden. De slider begint op de *neutrale* positie met waarde 50.

Figuur 5: Voorbeeld van een beoordeling over accuraatheid

4.3 Andere menu-opties

Op het introductiescherm kan de gebruiker ook klikken op het menu-icoon rechtsboven (*hamburgermenu*). Zoals Figuur 6 toont bestaat het menu bestaat uit:



Figuur 6: Het keuzemenu

- Duidelijke Taal: een link naar de introductiepagina (Sectie 4.1)
- Meedoen: een link naar de beoordelingspagina (Sectie 4.2
- Profiel, zie Sectie 4.3.1

- Melden van complexe zinnen, zie Sectie 4.3.2
- Zelf tekst vereenvoudigen, zie Sectie 4.3.3
- Scorebord
- Over deze website

4.3.1 Profiel

Als de gebruiker nog niet ingelogd was, wordt dit gevraagd als op *Profiel* geklikt wordt. Een gebruikersprofiel bevat enkel de volgende velden:

- de gebruikersnaam
- het emailadres van de gebruiker
- het wachtwoord van de gebruiker

Deze kunnen aangepast worden. Daarnaast heeft de gebruiker nog de optie om het account te verwijderen en om uit te loggen.

4.3.2 Melden van complexe zinnen

Er wordt de gebruiker de mogelijkheid geboden om complexe zinnen te melden. Zoals Figuur 7 toont krijgt de gebruiker dan een vrij veld om een complexe zin te melden.

Figuur 7: Melden van complexe zinnen

Daarnaast wordt nog de mogelijkheid geboden om een manuele vereenvoudiging te melden. De gemelde complexe zinnen kunnen later automatisch vereenvoudigd worden en de vereenvoudigingen, zowel de automatische als de manuele, kunnen in een latere beoordelingsronde door de crowd beoordeeld worden.

Om spam te vermijden wordt nog een vraagje gesteld, zodat de formulieren niet door bots ingevuld kunnen worden.

4.3.3 Zelf tekst vereenvoudigen

Bij deze optie krijgt de gebruiker hetzelfde scherm te zien als in Figuur 7, maar is de complexe zin al ingevuld. De manuele vereenvoudiging is hier niet langer optioneel.

5 Gamification

Om de crowd te motiveren zo veel mogelijk antwoorden te geven in de beoordelingsmodule (Sectie 4.2 werden een aantal gamification-elementen toegevoegd, zoals te zien in figuur 8.

Figuur 8: Gamification-elementen

5.1 Streak

De *streak* staat op *actief* als de gebruiker de afgelopen 30 uur een eerdere sessie heeft afgewerkt. Als de streak actief is verhoogt de score in de sessie met 10%.

5.2 Score

Er werd een scoresysteem toegevoegd aan de applicatie om gebruikers te motiveren. De samenstelling van de score bestaat uit een combinatie van de geleverde inspanning, de snelheid, en de correctheid van de antwoorden. Deze factoren worden op volgende wijze geoperationaliseerd:

- De inspanning wordt gemeten aan de hand van het aantal te lezen woorden
- Snelheid van het antwoord wordt gemeten door de tijd te nemen per woord en die af te wegen tegenover een schaal. Te snel antwoorden kan geen grondige beoordeling zijn.

  We gaan hierbij uit van een gemiddelde normale leessnelheid van 220 tot 350 woorden per minuut (wpm) of lengte per woord tussen 0,17 s/w en 0,27 s/w. Als de gebruiker binnen die grenzen antwoordt, is er niks aan

de hand ($s = 1$). Als de gebruiker sneller antwoordt dan is die misschien niet grondig aan het beoordelen. Daarom verkleinen we de score door te vermenigvuldigen met de ratio van versnelling t.o.v. maximale snelheid van 0,17 s/w. Dus als de gebruiker bijvoorbeeld antwoordt in 0,15 s/w, dus in 88% van de minimumtijd, dan vermenigvuldigen we de score met een factor van $s = 0.88\ (0,15/0,17)$. Als de gebruiker trager antwoordt (bvb. aan 0,40 s/w), berekenen we de factor door het percentage overschrijding van de maximale duur $0.27/0,40 : s = 0.675$.

- Wat betreft de veronderstelde correctheid c: by default stellen we $c = 10$, wat wil zeggen dat we de score met deze factor vermenigvuldigen. Als we al drie of meer oordelen hebben op een bepaalde vraag, veronderstellen we dat het nieuwe antwoord idealiter maar beperkt afwijkt van de eerdere oordelen. Als de nieuwe beoordeling binnen 0,5 standaarddeviatie daarvan ligt vermenigvuldigen we dan blijft $c = 10$. Bij hogere standaardeviaties veminderen we c met 1 eenheid per verdere afwijking van 0,2.

5.2.1 Scorebord

Op de website wordt een scorebord bijgehouden. De drie hoogste scores tot nog toe worden op elk scherm getoond. Via een aparte pagina kan de volledige ranking gezien worden.

Tijdens het beantwoorden laat het derde icoontje in de reeks gamification-elementen uit Figuur 8 zien hoeveel punten je nog nodig hebt om een plaats te stijgen in de ranking. In Figuur 8 wordt getoond wat de leider te zien krijgt.

5.3 Level

Het level geeft aan of de gebruiker zich bij de beste 10% deelnemers bevindt (*pro*), of bij de beste 66% (*gevorderd*). Anders wordt de gebruiker als *beginner* geklasseerd.

6 De antwoorden

De antwoorden worden opgeslagen in een databank en zijn downloadbaar in csv formaat. Zowel de individuele antwoorden als de gemiddeldes worden ter beschikking gesteld voor download op `https://hdl.handle.net/10032/tm-a2-y8`.

In totaal ontvingen we 24.745 antwoorden op de zes vragen die we per tekstpaar konden stellen. Tabel 2 toont de distributie van de antwoorden over de zes vragen en de 1071 zinsparen.

Hierbij valt op dat voor de vragen die parallelle tekstparen betreffen (duidelijkheid en accuraatheid) we de hoogste frequentie hebben voor vragen die zes maal beantwoord zijn. Het systeem was zo ingesteld om eerst vragen te selecteren uit een beperkte pool van 500 vragen, en als die vragen voor een gebruiker op waren, werd een nieuwe pool van vragen geactiveerd. Voor de vragen waar slecht één kant van de parallelle paren werd bevraagd geldt dat we het meeste antwoorden hebben voor teksten die door 3 personen beantwoord zijn. Voor tekstparen waarvoor alle vragen beantwoord zijn zien we de hoogste frequentie bij 2 personen die alles beantwoord hebben.

6.1 Analyse van de antwoorden

Een eerste analyse van de antwoorden wordt getoond in Tabel 3. Als we de drempelwaarde voor een accurate vereenvoudiging leggen op 50% *Accuraatheid 50*, dan zien we dat 83% van de vereenvoudigingen als accuraat beoordeeld worden. Als we de drempel op 70% leggen, dan zien we dat dit slechts in 49% van de gevallen geldt. Qua fluency kunnen we stellen dat er zo goed als geen verschil in beoordeling is tussen de oorspronkelijke zinnen en de synthetische kant. De oorspronkelijke zinnen worden in 69% van de gevallen als complexer beoordeeld, en qua duidelijkheid wordt de synthetische variant in 80% van de gevallen verkozen.

Als we naar ruwe gemiddeldes kijken (Tabel 4), dan zien we dat de accuraatheid van de simplificatie een beoordeling van 67 krijgt. De complexiteit van de oorspronkelijke zin (A) krijgt een gemiddelde beoordeling van 50, en de complexiteit van de synthetische zin (B) krijgt een beoordeling van 41. Dit verschil is statistisch significant ($p < .01$). Het verschil in fluency tussen zin A en zin B is verwaarloosbaar. De synthetische zin wordt in 80% van de gevallen als duidelijker beoordeeld.

Voor een verdere en meer gedetailleerde analyse verwijzen we naar toekomstig onderzoek.

7 Conclusie

In het voorgaande beschreven we een stand van zaken in verband met automatische tekstvereenvoudiging voor het Nederlands. Daaruit blijkt dat er weinig tot geen parallelle data zijn waarbij we erop kunnen vertrouwen dat de vereeenvoudigde kant ook effectief een vereenvoudiging inhoudt. Om

aantal	duidelijkheid	accuraatheid	fluency A	fluency B	eenvoud A	eenvoud B	alles
0	58	71	327	321	325	322	333
1	131	104	51	44	45	40	146
2	81	80	123	95	106	93	209
3	43	46	148	139	113	112	155
4	10	18	122	113	107	147	104
5	1	4	92	83	108	106	67
6	523	514	81	231	220	212	57
7	44	46	53	25	16	17	
8	36	42	36	9	18	6	
9	32	28	11	8	9	9	
10	29	32	11	1	3	4	
11	26	29	13	1		1	
12	17	20	3	1	1	1	
13	16	20				1	
14	9	7					
15	6	6					
16	3	3					
17	3	1					
18	2						
20	1						

Tabel 2: Distributie van de antwoorden over de zes verschillende vragen. De *aantal*-kolom geeft het aantal verschillende antwoorden op dezelfde vraag weer. De *alles*-kolom geeft het aantal zinsparen weer waarvoor alle vragen het aantal keer beantwoord zijn dat in de *aantal*-kolom staat.

Conditie	**% Teksten**
Accuraatheid 50%	83.00
Accuraatheid 70%	48.60
Complexiteit A > B	69.44
Fluency B > A	50.54
Duidelijkheid B > A	79.57

Tabel 3: Paarsgewijze vergelijkingen. Tekst A is de oorspronkelijke tekst. Tekst B is de synthetische tekst.

Conditie	**Gem**	**Stdev**
Accuraatheid	67.08	19.59
Complexiteit A	50.18	14.12
Complexiteit B	40.95	15.87
Fluency A	62.73	19.50
Fluency B	63.65	18.56
Duidelijkheid	64.19	19.47

Tabel 4: Ruwe gemiddeldes van de beoordelingen

aan menselijke beoordelingen te komen hebben we een crowdsourcinginfrastructuur infrastructuur opgezet en de mening van de crowd gevraagd over een aantal zinsparen.

Uit de antwoorden van de crowd blijkt alvast dat er best een onderscheid gemaakt kan worden tussen *duidelijkheid* en *eenvoud*. Zinnen waarin de moeilijkste concepten verduidelijkt worden zijn bvb. niet altijd eenvoudiger op syntactisch vlak.

De dataset kan helpen bij het bouwen of finetunen van een herschrijfsysteem, zoals bvb. gebeurt in Seidl and Vandeghinste (2024). Daarnaast kunnen we uit deze dataset een subset samplen waarbij we enkel die parallelle teksten overhouden waarbij de inter-annotatorovereenkomst het hoogst is en waarbij de accuraatheid hoog ligt en de vereenvoudiging of verduidelijking substantieel is, en die apart beschikbaar maken voor de research community, waarbij die dan bvb. gebruikt kan worden als evaluatieset.

Op basis van de beschikbare data kan ook gewerkt aan automatische evaluatiemetrieken voor tekstvereenvoudiging in het Nederlands die optimaal correleren met menselijke beoordeling, in de lijn van hoe de COMET metriek in automatische vertaling getraind is (Rei et al., 2020).

De data kan ook gebruikt worden voor het bouwen van classificatiesystemen die een gegeven tekst beoordelen op fluency, eenvoud en duidelijkheid. We denken hierbij aan integratie in tools zoals bvb Schrijfassistent.[2]

Daarnaast kan er ook onderzoek gebeuren naar de impact van bepaalde taalkundige kenmerken op bvb. de duidelijkheid van taal.

References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020a. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia

[2] https://schrijfassistent.be/index.php

Specia. 2020b. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

R.H.M. Brouwer. 1963. Onderzoek naar de leesmoeilijkheden van Nederlands proza. *PedagogischeStudiën*, 40:454–464.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. In *Essential speech and language technology for Dutch*, pages 219–247. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Theresa Seidl and Vincent Vandeghinste. 2024. Controllable sentence simplification in dutch. *Computational Linguistics in the Netherlands Journal*, 13:31–61.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Charlotte Van de Velde, Bram Vanroy, and Vincent Vandeghinste. 2023. Automatic sentence-level simplification for dutch. In *Computational Linguistics in the Netherlands. Abstracts*.

Nick Vanackere and Vincent Vandeghinste. 2022. Building a comparable corpus between easy-to-read Dutch Wablieft and De Standaard. Master's thesis, KU Leuven. Faculteit Ingenieurswetenschappen.

Vincent Vandeghinste and Bram Bulté. 2019. Linguistic proxies of readability: Comparing easy-to-read and regular newspaper dutch. *Computational Linguistics in the Netherlands Journal*, 9:81–100.

Daniel Vlantis, Iva Gornishka, and Shuai Wang. 2024. Benchmarking the simplification of Dutch municipal text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2217–2226, Torino, Italia. ELRA and ICCL.

S. Wubben. 2013. *Text-to-text generation by monolingual machine translation*. Ph.D. thesis, Tilburg University. Series: TiCC Ph.D. Series Volume: 26.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016a. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016b. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.