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

Deze toepassing betreft de 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.

Dit project is een gevolg van een aantal gesprekken tussen het Instituut voor de Nederlandse Taal en de Taalunie naar aanleiding van voorgaande projectvoorstellen, 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 pilotproject 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., 2016), BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTscore (Zhang et al., 2020). Voor onderzoek naar Nederlandse vereenvoudiging zijn er nauwelijks referentiesets beschikbaar en als die er wel zijn, bestaan deze uit (automatische) vertalingen van Engelse testsets, zoals in Seidl and Vandeghinste (2024).

Om een grootschalige goudstandaard-testset te creëren, hebben we een crowdsourcing-applicatie opgezet waarmee gebruikers automatische vereenvoudigingen handmatig kunnen evalueren, op de volgende dimensies:

- Gebruikers wordt gevraagd de vloeiendheid van zinnen te beoordelen (is de zin in correct Nederlands geschreven?)
- Gebruikers wordt gevraagd de eenvoud van zinnen te beoordelen
- Gebruikers wordt gevraagd om bij een zinspaar te zeggen welke de eenvoudigste zin is
- Gebruikers wordt gevraagd de juistheid van de vereenvoudiging te beoordelen (heeft de vereenvoudigde zin dezelfde betekenis als het origineel?)

Om de gebruikersbetrokkenheid te stimuleren hebben we een score opgenomen op basis van inspanning, snelheid, overeenstemming en meerdaagse streaks en een gebruikersscorebord.

De website is beschikbaar op https://duidelijketaal.ivdnt.org/.

2 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 het grote 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 zin bestaan. 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.[1]:

> "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 zal beschikbaar worden gesteld in de CLARIN-infrastructuur van het Instituut voor de Nederlandse Taal.

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

De dataset is beschikbaar voor download op `https://hdl.handle.net/10032/tm-a2-y7`.

3 De applicatie

3.1 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

3.2 Beoordelingsmodule

Na het klikken op *MEEDOEN* kom je 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 3.2.1) en het beoordelen van welke zin het eenvoudigste is van een zinspaar (sectie 3.2.2).

Accuraatheid wordt beschreven in sectie 3.2.4 en vlotheid wordt beschreven in sectie 3.2.3.

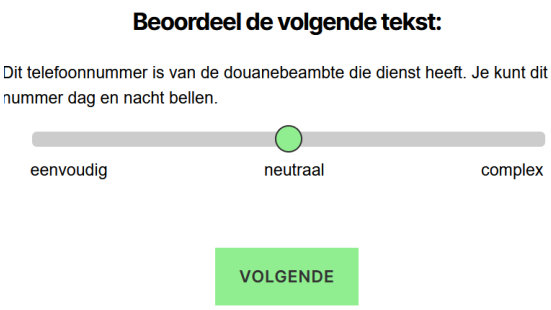
Er worden per sessie twintig vragen aangeboden, telkens vijf uit de vier verschillende taken.

3.2.1 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.

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

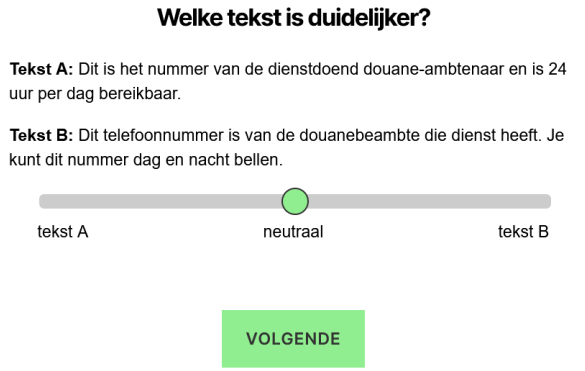


Figuur 2: Voorbeeld van een beoordelingsvraag met betrekking tot eenvoud

3.2.2 Paarsgewijze beoordeling van eenvoud

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 *teskt 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

3.2.3 Zinsbeoordeling op *fluency*

Bij deze beoordeling worden enkel zinnen uit de automatisch vereenvoudigde set aangeboden en wordt aan de gebruiker gevraagd in welke mate deze zinnen *goed* of *fout* Nederlands zijn.

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*

3.2.4 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.

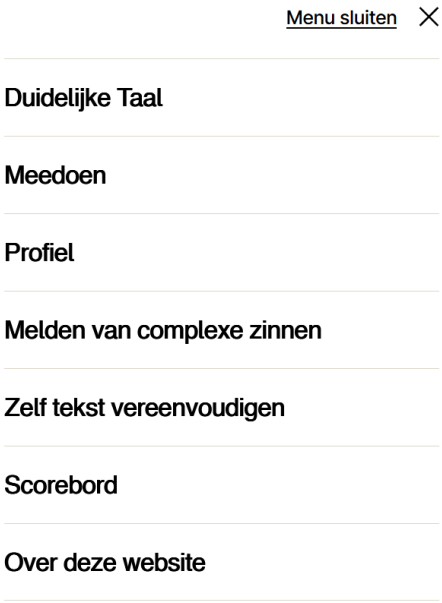
Zoals te zien in Figuur 5 krijgt de gebruiker nu een slider te zien op een as met als uitersten *andere betekenis* en *zelfde betekenis*. Net zoals bij de andere taken gaat het hier om een intervalschaal met 100 eenheden. De slider begint op de *neutrale* positie met waarde 50.

Figuur 5: Voorbeeld van een beoordeling over accuraatheid

3.3 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:

- Duidelijke Taal: een link naar de introductiepagina (Sectie 3.1)



Figuur 6: Het keuzemenu

- Meedoen: een link naar de beoordelingspagina (Sectie 3.2
- Profiel, zie Sectie 3.3.1
- Melden van complexe zinnen, zie Sectie 3.3.2
- Zelf tekst vereenvoudigen, zie Sectie 3.3.3
- Scorebord
- Over deze website

3.3.1 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.

3.3.2 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.

Daarnaast wordt nog de mogelijkheid geboden om een manuele vereenvoudiging te melden. De gemelde complexe zinnen kunnen later automatisch

Figuur 7: Melden van complexe zinnen

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.

3.3.3 Zelf tekst vereenvoudigen

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

4 Gamification

Om de crowd te motiveren zoveel mogelijk antwoorden te geven in de beoordelingsmodule (Sectie 3.2 werden een aantal gamification-elementen toegevoegd, zoals te zien in figuur 8.

Figuur 8: Gamification-elementen

4.1 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%.

4.2 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 kan gemeten worden 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 woorden per minuut (wpm) tot 350 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 tov 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 correctheid: in het geval dat er nog geen drie scores zijn voor deze zin of zinspaar, behandelen we de score alsof de score binnen 0,5 standaarddeviatie ligt van de scores die tot nu toe gegeven zijn. In dat geval geven we een correctheidsscore = 10. Voor afwijkingen groter verminderen we correctheidsscore gradueel met 1 per 0,2

4.2.1 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.

4.3 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.

5 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`.

6 Conclusie

Dit document beschreef het opzetten van een crowdsourcinginfrastructuur voor menselijke evaluatie van geautomatiseerde teksvereenvoudiging. Deze toepassing wordt gelanceerd op 27 juni 2024, en een prijs voor de best scorende gebruiker wordt toegekend eind september 2024.

De resulterende gegevens zullen publiek beschikbaar gemaakt worden via de onderzoeksinfrastructuur CLARIN.

References

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.

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

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.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. 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*.