Results of Crowdsourcing Human Evaluation of Synthetic Simplification

Vincent Vandeghinste, Bram Vanroy, Job van Doeslaar
Instituut voor de Nederlandse Taal, Netherlands
first.lastname@ivdnt.org

Abstract

This paper describes the creation of a set of crowd-sourced human evaluations of automated simplifications. We created a synthetic simplification dataset which was assessed by the crowd on dimensions of *fluency, clarity, accuracy* and *complexity*. We briefly describe the crowdsourcing environment and some techniques to keep the crowd engaged. This resulted in a dataset of nearly 25,000 responses from 384 respondents on synthetically simplified Dutch data. In order to mitigate the low inter-annotator agreement we investigate the effect of outlier removal techniques on Krippendorf α. The main conclusions stay the same: the synthetic texts are not simpler according to the linguistic proxies investigated, but the crowd judged that in about 80% of the cases the synthetic text was more clear than the original and less complex in about 70% of the cases. The fluency was the same for original sentences versus synthetic sentences, and in the case of accuracy the numbers changed most after outlier removal. A medium accuracy was achieved in about 80%, a high accuracy in 50% before outlier removal and 60% after outlier removal.

Both the synthetic parallel data as well as the crowd assessments are made available to the CLARIN infrastructure.

1 Introduction

Automatic text simplification (ATS) is the somewhat unfortunate term used in Natural Language Processing for the domain that works on automated text difficulty adaptation with the goal to enhance text accessibility for different groups of users. In government communication, there is particular emphasis on *plain language*, as evidenced by national campaigns in Belgium (Vandeghinste et al., 2021) and the Netherlands, and the ISO Standard on Plain Language (ISO, 2023). Clear and comprehensible communication is essential to ensure that all citizens, including those with limited literacy, can access and understand government information.

Recent advancements in artificial intelligence (AI) have enabled machine-based text simplification methods. However, a major challenge remains: the lack of reliable automated evaluation metrics, which depend on high-quality reference data. For most languages, including Dutch, such data is scarce and the development of robust training and evaluation datasets is still a key hurdle.

This work describes the pilot project *Duidelijke Taal* [EN: Clear language] which aligns with the policy of the Nederlandse Taalunie [EN: Dutch Language Union] for plain and understandable language in government and other societal sectors (law, healthcare). The Taalunie has been working closely with the Dutch and Flemish governments and various field organizations in the Netherlands and Flanders since 2016 to achieve this goal.[1]

The project originated from discussions between K-Dutch, the CLARIN Knowledge Centre for Dutch[2], hosted at the Instituut voor de Nederlandse Taal [EN: Dutch Language Institute][3] and the Taalunie, where it was established that in order to develop and properly evaluate a data-driven, state-of-the-art

[1] https://taalunie.org/dossiers/44/begrijpelijke-overheidstaal
[2] https://kdutch.ivdnt.org
[3] https://www.ivdnt.org

approach for automatic text conversion to *plain language* in Dutch, there is a significant lack of manually validated simplified data for both training and evaluation purposes.

Until now, plain text writing (or text adaptation) of government texts was mainly done manually. However, recent rapid advancements in AI offer the possibility of having machines perform this task as well. This pilot project explores the current feasibility of this and assesses its effectiveness. We primarily target government communication, as it is supposed to be written in plain language so that as many people as possible can understand its content.

We developed a crowdsourcing approach to manually assess and validate AI-generated simplifications. The evaluation focused on four dimensions: *fluency*, which measures grammatical correctness; *complexity*, which assesses ease (or rather difficulty) of comprehension; *clarity*, which evaluates comparative clarity in text pairs; and *accuracy*, which determines how well the meaning is preserved. To enhance engagement of the crowd, gamification elements were incorporated in the crowdsourcing environment, encouraging participants to contribute consistently.

Related work on text simplification, Dutch datasets for text simplification, and human evaluation of simplifications are described in Section 2. The data used for this study and how synthetic simplifications were created are described in Section 3.1. To create a large-scale gold-standard test set, we developed a crowdsourcing application in which users manually evaluated automatic simplifications. This application is described in Section 3.3. The results of the crowd sourcing and analysis of inter-annotator agreement and outlier removal are presented in Section 4, and Section 5 concludes the paper.

2 Related Work

In Section 2.1, we review existing applications for Dutch text simplification. Section 2.2 outlines parallel datasets available for this task, and Section 2.3 discusses the metrics used for the automated evaluation of simplified texts.

2.1 Automatic Text Simplification

Until recently, automatic text simplification was often performed using machine translation techniques (Xu et al., 2016), where the original text was treated as the source and the simplified text as the target language. Wubben (2013) attempted this approach for Dutch using statistical machine translation. More recently, large language models for Dutch, such as (multilingual or Dutch-specific variants of) the T5 model (Raffel et al., 2020), have been fine-tuned with parallel data for text simplification (Seidl & Vandeghinste, 2024).

Several applications for automatic text simplification in Dutch are available on the market. We try to keep track of these at the CLARIN Knowledge Centre for Dutch (K-Dutch).[4]

The most notable is the *Lees Simpel* [EN: Read Simple] mobile app developed by Benedictus (2025), which allows users to take a picture of a complex government letter and provides a simplified version of the letter, i.e. a number of bullets as to what the letter contains, with a special focus on actions expected from the user. The app is completely free and quite popular. It is wrapped around existing commercial generative AI models (currently with extensive prompt engineering). No systematic methodologically stringent evaluations have been performed in order to measure the quality of the app (Benedictus, 2025).

A related app with a similar approach is the *Schrijf Simpel* [EN: Write Simple] app,[5] which is from the same creators and takes the text writer's point of view by providing suggestions for simplifying letters.

2.2 Data for Dutch Text Simplification

At K-Dutch,[6], the CLARIN Knowledge Centre for Dutch, we aim to maintain an overview of existing datasets for Dutch text simplification. We list the most important ones here.

[4]https://kdutch.ivdnt.org/wiki/Text_simplification

[5]https://schrijfsimpel.nl/

[6]https://kdutch.ivdnt.org/wiki/Simplification_Data

Manually Created Datasets So far, we have identified only one manually created dataset for Dutch, compiled by Vlantis et al. (2024), which pertains to municipal texts.

Automatically Translated Datasets Seidl and Vandeghinste (2024) translated the ASSET dataset from Alva-Manchego, Martin, et al. (2020) and the WikiLarge dataset from X. Zhang and Lapata (2017) into Dutch and used them for fine-tuning a language model. Another automatically translated dataset is the SimpleWiki dataset,[7] created by the National Forensic Institute.

Synthetic Datasets With the advent of generative large language models, we can now prompt such models to generate simplified texts. This approach has been used for the UWV Leesplank NL Wikipedia dataset,[8] and for the data employed in Van de Velde et al. (2023). It is also the method used in our research.

Comparable Corpus of Easy and Regular-Level Texts A comparable corpus in this context is a corpus of regular text on the one hand and easy or plain language texts on the other hand, with an indication of how similar their topic and content is.

A comparable corpus is available containing simplified news articles from the Wablieft newspaper and articles on the same topics or events from De Standaard (Vanackere & Vandeghinste, 2022).

2.3 Evaluation of Text Simplification

Metrics for the automatic evaluation of text simplification face several challenges. One of these challenges is that they often require reference simplifications based on a gold standard to compare against automatic simplifications. This applies to metrics such as SARI (Xu et al., 2016), BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BLEURT (Sellam et al., 2020), and BERTscore (T. Zhang et al., 2020). For research on Dutch simplification, there are hardly any reference sets available, and if they exist, they often consist of (automated) translations of English test sets, as in Seidl and Vandeghinste (2024).

For manual evaluation, Alva-Manchego, Scarton, and Specia (2020) describe how 3-point or 5-point Likert scales are often used to assess grammatical correctness (also referred to as fluency), meaning preservation (adequacy), and simplicity. Fluency is evaluated solely on the output, while adequacy and simplicity are generally assessed based on the text pair.

This approach is refined in Alva-Manchego, Martin, et al. (2020), where evaluation is conducted on a continuous scale (0-100) for adequacy, fluency, and simplicity. The use of continuous interval scales in crowdsourcing human evaluations is common practice in machine translation (Bojar et al., 2018), as it results in higher levels of inter-annotator consistency compared to ordinal Likert scales. In our approach, we also use continuous interval scales (without numerical indicators).

3 Methodology

3.1 The Dataset

We created a test set of 6,986 sentence pairs, selecting the original sentences from the WRPEI component of the SONAR corpus (Oostdijk et al., 2013),[9] which consists of texts from websites. We chose this component because we assume that it contains language intended for a broad audience and is therefore expected to be clear and accessible.

We selected only sentences containing between 10 and 50 words, including at least one verb, with a Readability Index coefficient (Brouwer, 1963) higher than 60, and consisting of more than one clause. This means that the sentence must include at least one subordinate clause. The selection was performed using the tools described in Vandeghinste and Bulté (2019) in order to only simplify sentences that initially exhibit a reasonable level of complexity.

These sentences were automatically simplified using GPT-4 with the same prompt as used in the UWV/Leesplank project, available on HuggingFace:

[7] https://huggingface.co/datasets/NetherlandsForensicInstitute/simplewiki-translated-nl

[8] https://huggingface.co/datasets/UWV/Leesplank_NL_wikipedia_simplifications

[9] Available through the CLARIN infrastructure at http://hdl.handle.net/10032/tm-a2-h5.

> "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."

The resulting dataset is made available for download in the CLARIN infrastructure of the Instituut voor de Nederlandse Taal, at https://hdl.handle.net/10032/tm-a2-y7. This parallel dataset, or parts thereof, could be uploaded into the application described in section 3.3, in the form of a CSV file, via a dashboard accessible to administrators.

Characteristics of the Synthetic Dataset

From the total dataset, a random selection of 1,071 parallel text snippets was made. In order to get multiple assessments for as many sentence pairs as possible. and have not too many sentence pairs with only a single user assessment, we started with a limited set of 500 sentence pairs and gradually increased the number of sentence pairs because the most frequent users had consumed nearly all available sentence pairs in the application.

Characteristics of the source material and the synthetic adaptation are provided in Table 1, showing average scores for the source and target material, as well as the significance of differences between the source and target language using the p-value from a t-test. For details on the operationalisation of these characteristics, we refer to Vandeghinste and Bulté (2019).

Characteristic	Source	Target	p	Characteristic	Source	Target	p
Text length	17.25	35.17	$<.01$	Syllables/word	1.51	1.57	$>.05$
Words/sentence	16.53	20.22	$<.01$	Age of acquisition	9.78	8.61	$>.05$
Avg. clause length	10.95	10.56	$<.01$	Content words/word	0.54	0.53	$>.05$
Subordinate clauses	0.81	1.58	$<.01$	Concreteness	2.89	2.85	$>.05$
Subordinate clauses/clause	0.35	0.31	$<.01$	77% ratio	78.57	84.15	$<.01$
Subordinate clause length	7.53	7.25	$>.05$	Long words	0.05	0.05	$>.05$
Words/finite verb	8.92	8.63	$<.05$	Preposition ratio	0.14	0.12	$<.01$
NP length	4.02	3.87	$>.05$	Specificity ratio	0.02	0.02	$>.05$
Dependency/head ratio	1.76	1.79	$<.05$	Pronoun ratio	0.08	0.10	$<.01$
Tree depth	7.17	7.53	$<.01$	Noun ratio	0.21	0.20	$<.01$
Type-token ratio	0.91	0.72	$<.01$	Numeral ratio	0.04	0.04	$>.05$
Guiraud	3.68	3.39	$<.01$	Article ratio	0.11	0.10	$<.05$
Flesch Reading Ease	62.21	53.68	$<.01$	Conjunction ratio	0.05	0.05	$>.05$
Flesch Douma	75.19	67.30	$<.01$	Adverb ratio	0.05	0.05	$>.05$
CLIB	75.47	83.11	$<.01$	Adjective ratio	0.06	0.06	$>.05$
CILT	78.00	79.06	$<.01$	Letter ratio	0.09	0.09	$>.05$
Brouwer	60.77	49.52	$<.01$	Verb ratio	0.16	0.17	$>.05$
Characters/word	4.74	4.78	$>.05$				

Table 1: Characteristics of the evaluated dataset

As we can see, the synthetic simplifications do not align with the standard characteristics expected from the comparison between regular and easy-to-read language, as established in Vandeghinste and Bulté (2019). For example, the average text length of the alleged simplification is longer, and the number of words per sentence is also higher. This is likely due to the fact that the prompt explicitly instructs the model to explain certain terms. However, we do observe that the average clause length is shorter, as is the number of words per finite verb. Differences in NP length and subordinate clause length are not significant. In terms of lexical density measures (type-token ratio, Guiraud), there is a significant indication of simplification.

Surprisingly, for readability metrics (Flesch, CLIB, CILT, Brouwer), the synthetic simplifications are rated as more difficult (lower scores). This may be largely due to the fact that nearly all these metrics use words/sentence as one of their features, and, as Table 1 shows, sentences in our targets are on average nearly 4 words longer.

For part-of-speech ratios, we find that the number of prepositions is significantly lower in the simplified texts, while the number of nouns and pronouns is significantly higher.

The six best-performing features for classifying texts as regular vs. simplified, as identified in Vandeghinste and Bulté (2019), all score counter-intuitively here.[10]

We would like to emphasize that these characteristics were calculated over the entire dataset, without considering the quality (particularly adequacy) of the simplifications. Further research should determine whether these patterns persist when calculations are made only for instances where the simplification was judged accurate.

Experimentation with different prompts also remains future work. Placing less emphasis on explanation may lead to very different values for the characteristics of the synthetic text.

3.2 The Assessors

The crowd that created the evaluation dataset was recruited through social media posting. Apart from an email address for the participants in the competition, no information was collected about the assessors. 384 people have assessed questions, with a maximum of 3756 questions answered by a single user, with an average of 64 and a median of 19, indicating that more than half of the users only answered a single questionnaire of 19 cases.

3.3 The Application

3.3.1 Introduction Screen

The introduction screen invites the user to participate. To keep the application as accessible as possible, users are not required to log in initially, allowing them to start answering questions without any additional steps. However, users have the option to log in immediately, enabling their score to be retrieved and allowing questions to be tailored based on their profile, ensuring they receive questions they have not previously answered.

3.3.2 Evaluation Module

After clicking on PARTICIPATE, the user enters the evaluation module. Manual evaluation of automatically generated simplified sentences is performed across multiple dimensions, such as *complexity*, *accuracy*, and *fluency*. Each session presents twenty questions, five from each of the four tasks. For every evaluation, the user sees a slider on a scale with labelled extremes, as shown in Figures 1a and 2. The scale is an interval scale with 100 units and the slider starts in the *neutral* position with a value of 50. The users do not see the exact value.

For complexity evaluation, we have defined two tasks: assessing a sentence for its *complexity* and determining which of two sentences in a sentence pair is clearer.

Sentence Evaluation for *Complexity*

In this task, sentences are randomly selected from the dataset, and users are asked to evaluate their complexity. These sentences can come from either the original or the automatically simplified set, ensuring a condition-blind evaluation.

Figure 1a shows the slider that users can adjust along a scale with *simple* and *complex* as endpoints.

Pairwise Comparison for *Clarity*

In pairwise evaluation, a sentence pair from the database, consisting of an original sentence and its automatic simplification, is randomly presented to the user for assessment. This ensures that there is no order effect and that the user evaluates in a condition-blind manner.

As shown in Figure 1, the user sees a slider on a scale with *Text A* and *Text B* as endpoints.

[10] These are words/sentence, dependency tree depth, clause length, Brouwer, Flesch and Flesch Douma.

(a) Example of a simplicity evaluation question

(b) Example of a pairwise evaluation question

Figure 1: Screen shots for the two complexity-related dimensions

(a) Example of a fluency evaluation question

(b) Example of an accuracy evaluation question

Figure 2: Screen shots for Fluency and Accuracy

Sentence Evaluation for *Fluency*

For this task, both original and automatically simplified sentences are presented to the user, who is asked to determine how well the sentences conform to proper Dutch. This is again a condition-blind evaluation.

As shown in Figure 2a, the user sees a slider on a scale ranging from *incorrect Dutch* to *correct Dutch*.

Pairwise Evaluation for *Accuracy*

In this task, a sentence pair is presented to the user, who assesses whether the content and meaning of the target sentence match the source sentence.

As shown in Figure 2b, the user sees a slider on a scale ranging from *different meaning* to *same meaning*.

3.3.3 Gamification

To encourage participants to provide as many answers as possible in the evaluation module, various gamification elements have been added, as shown in Figure 3.

Figure 3: Gamification elements

Streak The *streak* is *active* if the user has completed a session within the past 30 hours. When the streak is active, session scores increase by 10%.

Score A scoring system has been added to the application to motivate users. The score is based on a combination of effort, speed, and accuracy, operationalized as follows:

- Effort is measured by the number of words read.
- Answering speed is measured by the time taken per word and adjusted based on a scale. Answering too quickly suggests a lack of thorough evaluation.
- Accuracy is measured for sentences with three or more judgements. If the new rating is outside of 0.5 standard deviations of prior ratings, the accuracy score is gradually reduced.

Scoreboard A scoreboard is maintained on the website, displaying the top three scores at all times. A separate page shows the full ranking.

Level A user's level is displayed at the top of the page and is determined as follows: *Pro*: Top 10%; *Advanced*: Top 66%; *Beginner*: Remaining users.

We do not think Gamification put pressure on the annotators to provide answers closer to previously given answers, as this information was not available to the annotators. We did not provide much detail on the scoring system to the annotators to avoid adapting their annotation behaviour to the operationalisation of the scoring system. The main result of the scoring mechanism was that more answers lead to higher scores, so the main effect was that, for at least some users, this seemed to motivate them to provide many answers.

4 Results

The answers are stored in a database and can be downloaded in CSV format. Both individual responses and averages are made available for download at https://hdl.handle.net/10032/tm-a2-y8.

In total, we received 24,745 responses to the six questions that could be asked per sentence pair. Table 2 shows the distribution of answers across the six questions and the 1,071 sentence pairs. The *Count* column indicates the number of different annotations that were received for a specific question. The columns with the specific question headers show how many of such questions received a certain number of answers. For instance there were 58 *Clarity* questions that received 0 answers, and 523 *Clarity* questions that received 6 answers. The *All* column shows the number of sentence pairs for which all questions were answered at least so many times as indicated in the *Count* column.

It is noteworthy that for questions concerning parallel text pairs (clarity and accuracy), questions that were answered by six different annotators have the highest frequency. The system was set up to first select questions from a limited pool of 500 questions. Once those were exhausted for a user, a new pool of questions was activated.

For the questions where only one side of the parallel pairs was assessed, we see the highest number of responses for texts answered by three people. For sentence pairs where all questions were answered, we observe the highest frequency when two individuals answered all the questions.

4.1 Inter-annotator Agreement

Table 3 presents baseline inter-annotator agreement (IAA) scores, measured by Krippendorff's α, (Krippendorff, 2019) for the six evaluation scores. Overall, agreement levels are low across all conditions, with α values ranging from 0.10 (Clarity) to 0.17 (Complexity B), indicating substantial inconsistency among annotators. Accuracy and Clarity, despite having the highest number of ratings and items (around 1000), show particularly low agreement ($\alpha = 0.12$ and $\alpha = 0.10$ respectively), suggesting that annotators may have interpreted these criteria in varied ways or applied them inconsistently. The other four scales (Complexity A and B, and Fluency A and B) exhibit slightly higher αs (between 0.13 and 0.17) but are based on fewer ratings and items, around 3200–3300 ratings across approximately 740–750 items. These results collectively highlight the subjectivity of judgements on these criteria.

Count	Clarity	Accuracy	Fluency A	Fluency B	Simplicity A	Simplicity B	All
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						

Table 2: Distribution of responses across the six different questions. The *Count* column represents the number of different answers given to the same question. The *All* column represents the number of sentence pairs for which all questions were answered the number of times indicated in the *Count* column.

	All ratings			Outliers removed					
Condition	α	Ratings	Items	θ	α	Ratings	%Ratings	Items	%Items
Clarity	0.10	5848	1013	0.7	0.72	2619	45	928	92
Accuracy	0.12	5880	1000	0.8	0.70	3344	57	999	100
Fluency A	0.13	3222	744	0.7	0.79	1358	42	614	83
Fluency B	0.13	3256	750	0.7	0.76	1399	43	645	86
Complexity A	0.15	3263	746	0.7	0.76	1482	45	638	86
Complexity B	0.17	3276	749	0.7	0.79	1417	43	647	86

Table 3: Inter-annotator agreement measured in Krippendorff's α. Behind the vertical line, you see the threshold (θ) for the Z-score, which was the best outlier removal method.

To improve the reliability of the human ratings by filtering out inconsistent judgements, we have tested three outlier removal methods: Z-score, Modified Z-score (Z-mod), and Interquartile Range (IQR). *Z-score* (Montgomery & Runger, 2010) identifies outliers by measuring how many standard deviations a value lies from the mean; values beyond a threshold θ are flagged as outliers. *Z-mod* (Iglewicz & Hoaglin, 1993) uses median and median absolute deviation, making it more robust to skewed or non-normal distributions. *IQR* (Tukey, 1977) classifies outliers as values that fall below the first quartile (Q1) minus θ times the IQR, or above the third quartile (Q3) plus θ times the IQR. While each technique balances sensitivity to outliers and robustness differently, experimenting showed that Z-score provided the best outlier filtering method, allowing us to keep the largest number of ratings while achieving an acceptable agreement ($\alpha > 0.667$).[11]

For each of the conditions we have gradually raised the θ value until we reach the acceptability limit of $\alpha > 0.667$. The resulting θ and α values are presented on the right side of Table 3. The table also includes the number of remaining answers after filtering out the outliers. The Ratings and Items columns indicate how many ratings and items remained after removing the outliers.

Table 4 shows the updated distribution of answers after outlier removal. We see that for the pairwise

[11]The attentive reader may notice a difference in the numbers and percentages in the last four columns of Table 3 compared to the numbers presented in Vandeghinste et al. (2025). The previously reported numbers came from a different practical operationalisation and implementation of the outlier-removal procedure. The current numbers are the ones obtained after actually generating the dataset without the outliers.

Count	Clarity	Accuracy	Fluency A	Fluency B	Simplicity A	Simplicity B	All
0	60	72	329	322	326	324	342
1	195	158	156	153	123	134	398
2	234	243	323	307	323	299	291
3	202	176	119	143	150	195	39
4	152	157	78	94	109	84	1
5	133	177	38	45	30	27	
6	39	35	21	5	9	3	
7	18	24	5			4	
8	16	8	2	1	1	1	
9	11	8		1			
10	6	3					
11	3	6					
12		3					
13	1	1					
15	1						

Table 4: Distribution of answers after outlier removal

conditions, having two remaining answers has the highest frequency, and this also holds for the single sentence assessments. Overall, we see that having at least a single answer per condition has the highest frequency.

4.2 Analysis of the Answers

A summary of the answers is shown in Table 5, both for the full dataset as well as the one with outliers removed. If we set the threshold for an accurate simplification at 50 (*Accuracy 50*), we find that 83% of the simplifications are rated as accurate. If we set the threshold at 70, this is true for only 49% of cases. After outlier removal, this does not change much, with nearly 82% of the remaining data points indicating an accurate simplification of at least 50%.

Condition	**% of Texts**	**% of Texts after outlier removal**
Accuracy 50	83.00	81.88
Accuracy 70	48.60	61.16
Complexity A $>$ B	69.44	68.86
Fluency B $>$ A	50.54	51.15
Clarity B $>$ A	79.57	80.42

Table 5: Pairwise comparisons. Text A is the original text. Text B is the synthetic text.

When we raise the bar and only assume accuracy when the average user assessment is higher than 70, we see the accuracy levels drop to 49% for the full dataset. After outlier removal, this increases to 61%, indicating that in the majority of cases (with outliers removed) there is a strong judgement that the accuracy is good. Together with the *Accuracy 50* numbers, we can conclude that the accuracy is considered better than neutral between *different meaning* and *same meaning* in more than 80% of the cases (interpretation: *more ok than not ok*, and scoring above 70/100 accurate (interpretation: *good*) in about 50-60% of the cases.

In terms of complexity (in which source and synthetic simplification were assessed independently), we see that the complexity of the source sentences is assessed to be higher in nearly 70% of the cases. This does not change much with outlier removal. So according to the assessors, simplification took place in about 70% of the cases.

In terms of fluency (which was also assessed independently), there is virtually no difference in assessment between the original sentences and the synthetic versions. This does not change much after outlier removal either.

In terms of clarity, the synthetic versions are judged to be clearer in about 80% of the cases. This does not change much with outlier removal.

Looking at raw averages (Table 6), we see that the accuracy of the simplification is rated at 67, which rises up to 71 after outlier removal. The complexity of the original sentence (A) receives an average score of 50, while the complexity of the synthetic sentence (B) is rated at 41. This difference is statistically significant ($p < .01$), and even increased after outlier removal.

The difference in fluency between sentence A and sentence B is negligible, both before and after outlier removal. In pairwise comparison of clarity, the average assessment was 64% (67% after outlier removal), indicating that on average, synthetic sentences were assessed to be clearer.

	Before		After outlier removal	
Condition	**Mean**	**Stdev**	**Mean**	**Stdev**
Accuracy	67.08	19.59	71.16	23.44
Complexity A	50.18	14.12	51.78	17.52
Complexity B	40.95	15.87	41.22	18.71
Fluency A	62.73	19.50	64.43	23.61
Fluency B	63.65	18.56	64.97	23.15
Clarity	64.19	19.47	66.73	23.21

Table 6: Raw averages of the evaluations, before and after outlier removal

This information can be leveraged to improve future ATS systems by focusing on aspects that most impact readability. Additionally, discrepancies between automatic readability scores and human judgements underscore the need for refined evaluation metrics that align more closely with human perception.

5 Conclusions

In the preceding sections, we described a crowd sourcing evaluation campaign to evaluate state-of-the-art automatic text simplifications for Dutch. Our findings show that there is little to no parallel data where we can reliably assume that the simplified version is indeed a proper simplification. To obtain human evaluations, we established a crowdsourcing infrastructure and gathered opinions from the crowd on a number of sentence pairs.

The responses from the crowd indicate that a distinction should be made between *clarity* and *simplicity*. Sentences in which the most difficult concepts are clarified are not necessarily syntactically or lexically simpler.

It should further be noted that there is a large amount of inter-annotator disagreement in our dataset, and that we also applied pairwise comparisons after extensive outlier removal. As Tables 5 and 6 show, this has only a minor effect on the results, and does not alter our interpretation of these results.

The dataset can assist in building or fine-tuning a rewriting system, as demonstrated in Seidl and Vandeghinste (2024). Additionally, we can sample a subset of this dataset, retaining only those parallel texts with the highest inter-annotator agreement, high accuracy, and substantial simplification or clarification. This subset could then be made available separately to the research community, for example, as an evaluation set.

Based on the available data, work can also be done on developing automatic evaluation metrics for text simplification in Dutch that optimally correlate with human judgement, following the approach of the COMET metric in machine translation (Rei et al., 2020). The data can also be used to build classification systems that assess a given text for fluency, simplicity, and clarity. We can think of integration into tools such as the Schrijfassistent.[12] Furthermore, research can be conducted on the impact of specific linguistic features on aspects such as language clarity.

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

References

Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., & Specia, L. (2020, July). ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4668–4679). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.424

Alva-Manchego, F., Scarton, C., & Specia, L. (2020). Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, *46*(1), 135–187.

Benedictus, H. (2025). Lees Simpel App. Personal Communication. https://www.leessimpel.nl

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., & Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 272–303. https://doi.org/10.18653/v1/W18-6401

Brouwer, R. (1963). Onderzoek naar de leesmoeilijkheden van Nederlands proza. *Pedagogische Studiën*, *40*, 454–464.

Iglewicz, B., & Hoaglin, D. (1993). *How to detect and handle outliers*. ASQC Quality Press.

ISO. (2023). *Plain language — part 1: Governing principles and guidelines* (tech. rep. No. ISO 24495-1:2023). International Organization for Standardization. Geneva, Switzerland. https://www.iso.org/standard/78907.html

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th edition). Sage Publications.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 74–81. https://aclanthology.org/W04-1013

Montgomery, D., & Runger, G. (2010). *Applied statistics and probability for engineers*. John Wiley & Sons.

Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In *Essential speech and language technology for Dutch* (pp. 219–247). Springer.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67. http://jmlr.org/papers/v21/20-074.html

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

Seidl, T., & Vandeghinste, V. (2024). Controllable Sentence Simplification in Dutch. *Computational Linguistics in the Netherlands Journal*, *13*, 31–61. https://www.clinjournal.org/clinj/article/view/171

Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Van de Velde, C., Vanroy, B., & Vandeghinste, V. (2023). Automatic sentence-level simplification for Dutch. *Computational Linguistics in the Netherlands. Abstracts*.

Vanackere, N., & Vandeghinste, V. (2022). *Building a comparable corpus between easy-to-read Dutch Wablieft and De Standaard* [Master's thesis, KU Leuven. Faculteit Ingenieurswetenschappen].

Vandeghinste, V., & Bulté, B. (2019). Linguistic proxies of readability: Comparing easy-to-read and regular newspaper dutch. *Computational Linguistics in the Netherlands Journal*, *9*, 81–100. https://www.clinjournal.org/clinj/article/view/97

Vandeghinste, V., Muller, A., François, T., & Declercq, O. (2021). Easy Language in Belgium. In C. Lindholm & U. Vanhatalo (Eds.), *Handbook of Easy Languages in Europe* (pp. 53–90). Frank & Timme.

Vandeghinste, V., Vanroy, B., & Van Doeslaar, J. (2025). Human evaluation of automated text simplification through crowdsourcing. *Proceedings of the CLARIN Annual Conference*, 143–147.

Vlantis, D., Gornishka, I., & Wang, S. (2024, May). Benchmarking the simplification of Dutch municipal text. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2217–2226). ELRA; ICCL. https://aclanthology.org/2024.lrec-main.199

Wubben, S. (2013). *Text-to-text generation by monolingual machine translation* [Doctoral dissertation, Tilburg University] [Series: TiCC Ph.D. Series Volume: 26].

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, *4*, 401–415. https://doi.org/10.1162/tacl_a_00107

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

Zhang, X., & Lapata, M. (2017). Sentence simplification with deep reinforcement learning. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 584–594. https://doi.org/10.18653/v1/D17-1062