Getting Close to Cloze: Investigating Language Model and Human Cloze-test Performance in Afrikaans

Susan Lotz[♢♠], **Rik van Noord**[♢], **Gertjan van Noord**[♢]
[♢]CLCG, University of Groningen; [♠]SU Language Centre, Stellenbosch University
{s.lotz, r.i.k.van.noord, g.j.m.van.noord}@rug.nl

Abstract

Models that can estimate the readability of a given text automatically are a valuable resource for any language. There are however many languages for which such models do not work well or simply do not exist yet. In this paper, we lay the groundwork for developing a high-quality application for Afrikaans by having encoder-only language models (LMs) complete a set of cloze tests already completed by humans. Strong correlation between the cloze-test performance of humans and an LM is an indication that the LM could possibly serve as a proxy for human participants. We show that the output of models trained on (some) Afrikaans correlates reasonably well with human answers, underscoring the potential of LMs to be used in automatic readability assessment. A more fine-grained analysis confirms that the correlation is not driven by only a few strongly correlating word classes, but spread relatively evenly over all word classes. We further establish by means of a manual evaluation that, in cases where the cloze-test performance of humans and an LM correlate strongly because both were wrong, LM answers tend to be further off than human answers for the same cloze items. It is noteworthy that the model with the best correlation, afRoBERTa (r=0.62; Spearman's ρ=0.62), is neither the most accurate nor the largest model, but a model trained on Afrikaans only, showing the benefit of small, monolingual LMs compared to large, multilingual models for specific purposes.

Keywords: cloze tests, readability, surprisal, language models, Afrikaans

1. Introduction

Writers want readers to read their texts. Readers prefer to read texts that they can understand relatively easily. Readability assessment indicates how readable a text would be for a certain audience, thereby helping to connect readers with texts at an appropriate level. In traditional readability research, cloze tests have been used as a measure to determine the readability of a given text (Taylor, 1953), with several studies showing strong correlations between cloze scores and traditional comprehension test scores (Bormuth, 1967, 1968; Kamalski, 2007; Gellert and Elbro, 2013). Cloze tests require participants to fill in words that have been left out from a text at certain intervals. The extent to which participants succeed in this task gives an indication of how readable the text is, as it shows to what extent the participant was able to preempt the next (left out) word (see Section 2.1 for more).

Automatic readability assessment employs automatic measures, nowadays mostly deep learning models (Wilkens et al., 2024), to assess readability. If cloze tests aimed at determining the readability of a text for a certain human audience can be completed by a language model (LM) with strong correlation with the human answers, that LM could possibly serve as a proxy for human participants. A usually costly and time-consuming step in the development of readability assessment measures could this way be performed more easily, and large automated cloze-test sets reflecting reliable proxied answers for under-resourced languages could become more attainable.

Some prior work already used LMs to complete cloze tests. Benzahra and Yvon (2019) may have been the first to investigate cloze difficulty by means of neural language models (LMs), using GPT-2 (Radford et al., 2019), but without success. Olney (2022) argues that GPT-2, being autoregressive and only allowing leftward context to be used, was not the best neural LM to use for this task. Subsequently, Olney (2022) employed T5 (Raffel et al., 2020) to complete cloze tests and found that, across all three sets of corpora he investigated, T5 predictions correlated significantly with human cloze scores. In addition, Hofmann et al. (2022) and Shain et al. (2024) found compelling evidence for using LMs as proxies for human behavior in cloze tests. In more recent research, Lopes Rego et al. (2024) use LMs to augment a cognitive model of reading, and Sadlier-Brown et al. (2024) found that, among the 5 LMs they investigated, RoBERTa appeared to give the most human-like output in the cloze tests used in their experiments.

Previous research focused mostly on English, however, and there are many other languages that will benefit from having automatic readability assessment models. In this paper, we take the first step towards building such a system for Afrikaans, a non-agglutinative Indo-European Germanic language, used by over 6.5 million speakers in southern Africa. We explore whether multilingual or language-specific encoder-only LMs can successfully complete an Afrikaans cloze-test set aimed at determining readability previously also completed

Proceedings of Resources for African Indigenous Languages (RAIL) 2026 @ LREC 2026, pages 107–120
12 May 2026.

by human participants (Lotz et al., forthcoming). Then, we ascertain which LM's output correlates best with the existing human participants' cloze answers, and whether this correlation may indicate the possibility of developing an adequate model for predicting readability. Our aim is not to find the *best-performing* LM in the cloze-completion task, but the one that correlates best with human answers, two aspects that can differ quite substantially (Oh and Linzen, 2025).

Contributions To the best of our knowledge, we are the first to run LM experiments on a cloze-test set and investigate correlation with human answers for the same set for a language other than English. In this paper, we do the following:

1. We show that there is a modest, but clear correlation between the cloze output of LMs and humans, highlighting the potential of using such models for automatic readability prediction for Afrikaans.
2. We do a manual annotation of a subset of our data showing that, in cases where the cloze-test performance of humans and an LM correlate strongly because both were wrong, LM answers tend to be further off than human answers for the same cloze items.
3. By means of a part-of-speech analysis, we establish that the reported correlation is also not driven by a few strongly correlating word classes only.
4. We find that the best correlation is obtained for a relatively small model trained on Afrikaans only (afRoBERTa), a clear reminder that there is still a need for such smaller monolingual LMs, despite the recent success of large, multilingual LMs.

2. Previous work

2.1. Cloze tests and readability

The cloze test, introduced by Taylor (1953) as a criterion for readability, has been an important instrument in the development of readability measures for over 70 years. As a gold-standard data set, human cloze test results represent a benchmark for readability prediction that could be used in developing rule-based applications such as the classic formulas (DuBay, 2004; François, 2015; Jansen et al., 2017), and, more recently, for automatic readability assessment, where systems use machine learning (Collins-Thompson, 2014; Kleijn, 2018; Vajjala and Lučić, 2018; Crossley et al., 2019; Vajjala, 2022).

Cloze tests require participants to fill in words that have been left out from a text at certain intervals (Bormuth, 1967). Blanks may be selected according to different criteria for different tests: every nth word may be removed or specific kinds of words be blanked out (Kobayashi, 2002). Cloze tests are scored using the Exact scoring method, where only the original word is considered a correct answer, or using the Semantic or Acceptable scoring method, where synonyms and other plausible words may be accepted as correct (Bachman, 1985; O'Toole and King, 2011).

The extent to which participants succeed in filling the blanks gives an indication of how readable the text is, as it shows to what extent the participant was able to preempt the next (left out) word in the context of the text. When reading, humans also preempt the words that follow those they are actually reading, an observation that leads Goodman (1967) to describe reading as a "psycholinguistic guessing game". Both readers and cloze test participants thus rely on language knowledge, memory cues and context to make sense of what they read or need to fill in. The similarity of the action of filling in a cloze test and reading confers strong construct validity on the cloze test (Horton, 1974; Jansen et al., 2017).

Several studies also show strong correlations between cloze scores and traditional comprehension test scores (Bormuth, 1967, 1968; Kamalski, 2007; Gellert and Elbro, 2013). We have recently done some work on developing a new, empirically tested traditional readability measure for specifically Afrikaans (Lotz et al., forthcoming). We have not been able to put forward a reliable formula using the traditional method, and will therefore embrace more modern approaches, of which incorporating LMs is the next step, reported in this paper. We will use the existing cloze test data set completed by humans in that study for the current research.

2.2. Cloze tasks in NLP

In natural language processing (NLP), a cloze task builds on the original cloze procedure. It entails a fill-in-the-blank prediction, where an LM has to infer missing word(s) from the words that precede and follow the blanks. Cloze tasks have been used in NLP in several ways over the past 10 years: among others as tests of how well a system uses the surrounding context to choose an appropriate missing word (Paperno et al., 2016); as reading-comprehension benchmarks where a model has to fill in a missing word in a question using information from a given passage (Hermann et al., 2015); as a way to train encoder-only LMs (Devlin et al., 2019); and as simple fill-in-the-blank prompts to check what factual or linguistic knowledge such pretrained models appear to have stored (Petroni et al., 2019).

Although the cloze task has been applied in nu-

merous studies to train or evaluate LMs, it has been used less often in the way cloze tests are usually presented to humans, particularly in languages other than English. An example of such a study is that of Puccinelli et al. (2021), who use encoder-only LMs to take an Italian assessment cloze test usually taken by newcomer university students to ascertain their starting level. They found that LMs could successfully pass such tests, but they did not correlate human performance with that of LMs. Their setup also differed from the cloze tests in our study in that participants had to choose from a preselected list of words. Another example is a study by Nikiforova et al. (2020), who use a Russian cloze-test set (Laurinavichyute et al., 2017) originally completed by humans to investigate how selected LMs perform on the task of predicting the next word, given the corpus. The original cloze task for human respondents was to successively predict the next words for each context in the cloze test. No correlation between the performance of humans and LMs was calculated, although the authors used the actual human expectations about the next word for a given sequence as reflected in the cloze results as ground truth against which to evaluate the LMs' output.

Our work In our research, we apply the cloze task as originally used for human participants, with LMs having to complete the exact same cloze-test set. Instead of having only one cloze item per sentence, the cloze-test set we use contains multiple cloze items per sentence. We believe we are the first to use LMs to complete a cloze-test set designed for human participants in Afrikaans, and that we are the first to correlate human performance to that of LMs for a language other than English.

2.3. NLP resources for Afrikaans

Afrikaans is an indigenous African language (Kotzé, 2018; PanSALB, 2021), having developed from 17th-century Dutch dialects in contact with several indigenous and foreign languages spoken at the settlement at the Cape on the southernmost tip of Africa from the middle of the 17th century onwards (Davids, 1994; Conradie and Coetzee, 2014). It is one of the 12 official languages of South Africa and is used by over 6.5 million speakers in southern Africa. From an NLP perspective, Afrikaans is considered a low-resource language (Dirix, 2023; Eiselen and Gaustad, 2023), with Joshi et al. (2020b) placing it in the 'rising star' category, the third level of their six-point language classification, in which 0 represents exceptionally low-resource languages and 6 represents languages that benefit from every NLP breakthrough.

Some foundational work for NLP in Afrikaans has indeed been done: The South African Centre for Digital Language Resources (SADiLaR Language Resource Repository) houses, among other resources, lemmatized, POS-tagged, morphologically analyzed text corpora and other resources, the Autshumato English–Afrikaans parallel corpus (McKellar, 2022), and a RoBERTa-based LM that has been trained exclusively on Afrikaans (Eiselen, 2023). Afrikaans has also been included in several multilingual LMs (Conneau et al., 2020; Alabi et al., 2022; Dossou et al., 2022; Adebara et al., 2023). A dependency treebank for Afrikaans, AfriBooms (Augustinus et al., 2016), has been developed, enabling the development of automatic tokenization, POS tagging, lemmatization and dependency parsing (Qi et al., 2020), and the VivA Corpus Portal makes several Afrikaans corpora available for online searches (Virtuele Instituut vir Afrikaans (VivA)). Some initial work has been done on the impact of data scarcity on a generative question-answering (QA) model for Afrikaans (Moape et al., 2025), and Afrikaans has been included in multilingual QA evaluation in the BELEBELE Benchmark (Bandarkar et al., 2024) and AfroBench (Ojo et al., 2025).

3. Method

3.1. Data set

We use our existing cloze test data set as collected in Lotz et al. (forthcoming). This study included 595 Afrikaans-speaking participants, all over the age of 18. A set of 40 articles of approximately 300 words each from 7 genres was used (6 articles from a popular weekly magazine, 6 texts produced by the South African government, 6 newspaper articles, 6 insurance brochures, 6 health brochures, 5 electronic newsletters and 5 informed consent texts). The texts were chosen to represent different readability levels, for example, the popular weekly magazine articles would be expected to be in a lower register and therefore easier to read than the selected government texts. Five cloze tests were created for each of the 40 texts, which yielded 200 hardcopy cloze tests[1], with approximately 50 cloze items in each test. There were approximately 6 participants per cloze test, varying between 2 and 10. Each participant completed around 100 cloze items, leading to a data set of 59,111 annotations. The Exact scoring approach (O'Toole and King, 2011) was followed to ensure consistent scoring.

3.2. Language models

We ran the cloze completion experiments using encoder-only transformer-based masked LMs, as their pretraining objective directly matches the

[1]One Excel sheet unfortunately became corrupted, so we work with 199 files.

cloze task of predicting masked tokens. We used three types of LMs: models trained specifically for Afrikaans, multilingual models with Afrikaans in their training data, as well as monolingual models trained on a single language related to Afrikaans in different ways: Dutch as a fellow Low Franconian language, and English as a more distant West Germanic Anglo-Frisian relative.[2]

Afrikaans models The main model we use in this study is afRoBERTa (Eiselen, 2023), a model based on RoBERTa-base (Liu et al., 2019), but trained solely on Afrikaans texts – approximately 350 million words. afRoBERTa was developed by the Centre for Text Technology (CTexT) at North-West University in South Africa. In addition, we use Afro-XLM-R (Alabi et al., 2022), which is an adaptation of XLM-R-large (Conneau et al., 2020) with multilingual adaptive fine-tuning for 17 African languages, including Afrikaans (trained on the mC4 corpus: 752.2 MB; 3,697,430 sentences). Afro-XLM-R is in fact a multilingual model, but differs from other multilingual models in that it still has a specific focus on Afrikaans.

Multilingual models We employ three different multilingual LMs. The first two are XLM-R base and large (Conneau et al., 2020), which are RoBERTa-based models trained on 100 languages across 2.5 TB of CommonCrawl data. Finally, we use the smaller mmBERT (Marone et al., 2025), which was trained on 3 TB of data across 1,800 languages, with an extra focus on low-resource languages.

Non-Afrikaans monolingual models For completeness, we also ran two models that were not trained on Afrikaans itself at all, only on related languages, namely the Dutch BERTje model (De Vries et al., 2019) and the English DeBERTaV3 model (He et al., 2021). However, we found that these models were, in essence, unable to perform the task, as their vocabularies do not contain enough words in Afrikaans. For the sake of brevity, we do not include results on those models in the next section.

3.3. LM settings

Context To complete cloze tests, the LMs in this study have to predict multiple masks per sentence. However, if all of these were to be predicted simultaneously, the model would be disadvantaged compared to humans, as humans can keep track of their previous predictions and use them as context, which helps with subsequent predictions. Therefore, to ensure a process as close as possible to the one humans would follow, we have the model go through the cloze test from left to right, and insert its final prediction as additional context for subsequent predictions.

Tokenization Since transformer LMs use subword tokenisation, it may happen that a masked word consists of more than one subword. We opted to have the model either (1) predict only the first subword per mask ('first' setting) or (2) generate each subword step-by-step until the full word is reconstructed (iterative reconstruction) and compare those results. Some researchers opt for the simplicity of Option 1 (Kalo and Fichtel, 2022; Jacobs et al., 2024), however if only one token per mask is allowed, one cannot fully evaluate multi-piece word answers. The iterative reconstruction strategy (Kalinsky et al., 2023) in Option 2 is related to SpanBERT's span prediction (Joshi et al., 2020a), and ensures that models are evaluated fairly when reconstructing multi-token words. In our experiments, there was little difference between the two settings, and therefore we opt to show only the results for the more realistic iterative reconstruction setting.

Normalization Allowances for capitalization, Afrikaans diacritics and the use of hyphens in Afrikaans are made through normalization. Evaluation of the use of the indefinite article *'n*, a contraction of its historical Dutch form *een*, is also relaxed, allowing several variations ('n, ' n, and n, to name a few, even if they are not strictly speaking correct due to the wrong apostrophe being used or left out) to be considered correct (English equivalent: *a*).

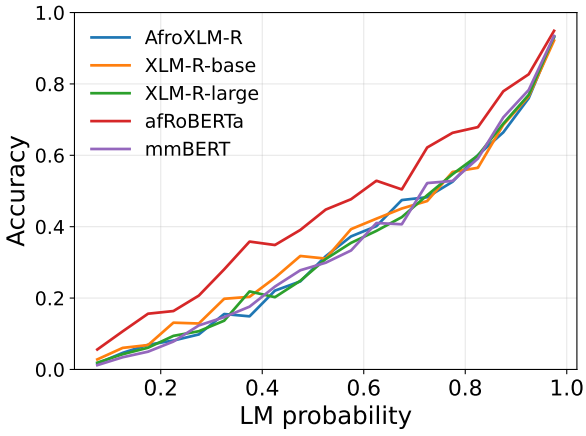


Figure 1: Accuracy vs. LM probability for the models used in this paper. Predictions are binned per 0.05 confidence and need at least 100 instances to be included.

[2]All code is available at: https://github.com/lotzdata/LM_human_cloze_performance

Cloze sentence: Van Vuuren is verras deur die _____ vrae wat in die _____ voorgekom het.		
Gold sentence: Van Vuuren is verras deur die (1) **tipe** vrae wat in die (2) **vraestel** voorgekom het.		
English translation: Van Vuuren was surprised by the (1) **type** of questions occurring in the (2) **paper**.		
Annotation option	**Response example**	**English**
1. Match with spelling/other difference	(1) tiepe, (2) vrastel	(1) tipe, (2) paperr
2. Synonym	(1) soort, (2) toets	(1) kind, (2) test
3. Plausible fit in context	(1) verskeidenheid	(1) range
(semantically close, but not exact; syntactically acceptable)	(2) eksamen	(2) exam
4. Alternative fit (different meaning)	(1) moeilike	(1) level
(syntactically acceptable, but may deviate in meaning)	(2) roman	(2) novel
5. Incorrect (nonsensical sentence)	(1) lamppaal, (2) straat	(1) lamp, (2) street

Table 1: An example of applying our annotation scheme.

3.4. Evaluation

We are interested in how well LM outputs correlate with human answers. We correlate each LM's output with the averaged score of humans on a given cloze item. For example, if 4 out of 6 humans gave a correct answer, this would yield an accuracy score of 0.67. We calculate Pearson's r and r^2, as well as Spearman's ρ over the 10,996 unique cloze items.[3] It is clear how to score human output, but there are multiple ways to calculate LM performance. The three methods we employed are outlined below.

Probabilities The most obvious metric would be the use of the softmax probabilities of the correct answer.[4] This is irrespective of the rank of the answer: if a model gave a 21% probability to the correct answer, we simply correlate 0.21 with the human score, whether the model had this as its best prediction or not. As a sanity check, we plot the confidence (=probability) of each model's final prediction versus its accuracy, and find a clear correlation between the two in Figure 1.

Reciprocal rank One could argue that the ranking of answers is a better signal than probabilities only. For this metric, we assign a score based on the reciprocal rank of the correct answer, defined as $\mathrm{RR} = 1/i$, where i is the index of the correct answer. This gives more weight to small differences at the top of the ranking, while differences at the bottom of the ranking are negligible (Olney, 2022).

Top-K A different method of including ranking information is to look at the Top-K predictions, and simply assign a score of 1 if the correct prediction occurs in the Top-K. For example, if a model predicts the correct word at the 10th position, it would get a score of 1 for K=10 and higher, but a score of 0 for K=9 and lower.

[3]Given the large sample size, all our correlations are significant, with very small confidence intervals, which we omit for brevity.

[4]This is often referred to as **surprisal** as well.

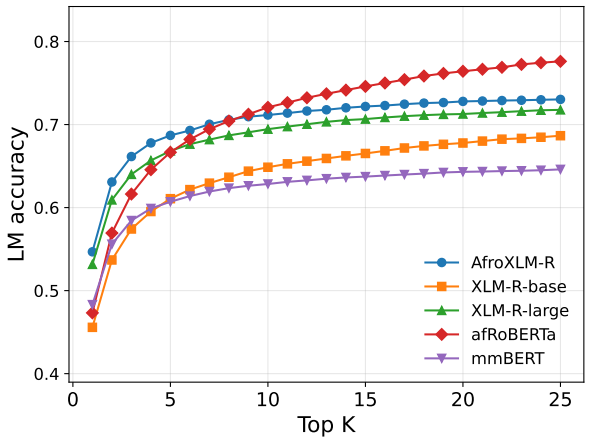
3.5. Manual annotation

For a better understanding of how humans and LMs differ when providing answers for the cloze-test set, we perform a manual annotation on a subset of the data. Cloze answers for seven texts, one of each of our seven text genres, were annotated by four annotators. The annotators speak Afrikaans as their native language, all have language-related undergraduate degrees, and three have postgraduate qualifications. Two are experienced language practitioners with over 25 years of experience each, one a linguistics master's student, and one a retired librarian. The annotators received clear instructions for the task, including detailed annotation examples. No distinction was made between the human and LM cloze answer subsets when presented to the annotators; the answers were scrambled before annotation and unscrambled afterwards.

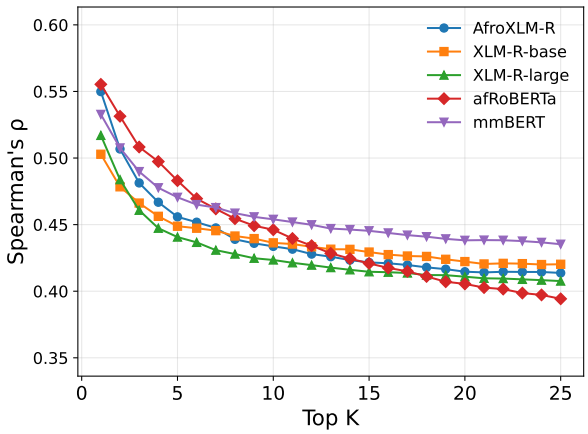
Following Carrell et al. (1993)'s Acceptable cloze scoring procedure for their study on first- and second-language reading strategies, we developed an annotation scheme consisting of five categories:

1. Match, with spelling/other difference
2. Synonym
3. Plausible fit in context
4. Alternative fit (different meaning)
5. Incorrect

These categories were used to annotate answers that were already considered wrong according to the Exact cloze scoring method. If the human or LM answer was an exact string match with the gold answer, no annotation was needed. For each applicable cloze item, the annotators judged the output of afRoBERTa, as well as the most frequent human answer. We aggregated the four annotators' annotations, by selecting the most frequent annotation for each cloze item, and in cases where there was a tie, reflecting the verdict by Annotator 4, the most experienced annotator, and also an author of the



(a) Accuracy at top-K.



(b) Spearman's ρ correlation at top-K.

Figure 2: Performance at top-K for the LMs under investigation.

current study. The annotations can be regarded as points on a continuum with reasonable acceptability at the one end and complete incorrectness at the other end. An example of the application of the annotation scheme can be seen in Table 1.[5]

Agreement We assessed inter-rater agreement by means of Krippendorff's alpha (α) and mean linear weighted pairwise Cohen's kappa (Cohen's κ_w) due to the ordinal nature of the annotation scheme. Across the four annotators, we obtain a Krippendorff's α of 0.63. Pairwise Cohen's κ_w values for the 6 annotator pairs ranged from 0.44 to 0.64, indicating moderate agreement between annotators.[6]

	Acc	**Probability**			**Reciprocal Rank**			
	$K=1$	r	r^2	ρ	Sc	r	r^2	ρ
afRoBERTa	0.47	0.62	0.38	0.62	0.56	0.58	0.34	0.57
Afro-XLM-R	0.55	0.61	0.37	0.61	0.61	0.55	0.31	0.54
XLM-R-base	0.46	0.56	0.31	0.57	0.53	0.52	0.27	0.52
XLM-R-large	0.53	0.58	0.33	0.59	0.59	0.52	0.27	0.52
mmBERT	0.48	0.59	0.35	0.58	0.54	0.54	0.29	0.53

Table 2: Results of applying several LMs on our data set of cloze tests in Afrikaans. Pearson's r, r^2 and Spearman's ρ are measures of correlation with human answers. "Acc" denotes general accuracy of the LMs (K=1). "Sc" denotes the average RR score of a model. Human accuracy was 0.51.

4. Results and discussion

Table 2 shows the main results of our study. All models have a clear, though modest, correlation with human answers. We obtain the best correlation when looking at LM probabilities directly, instead of the more sophisticated reciprocal rank method. It is noteworthy that the model with the best correlation is in fact afRoBERTa (r=0.62; r^2=0.38; Spearman's ρ=0.62), which was trained on Afrikaans only. This observation shows that there is still a clear need for developing such smaller, encoder-only models for a single language, even if large LMs can take care of many tasks for a given language.

Bigger is not better Figure 2a and 2b show accuracy and correlation (ρ) when using Top-K values of 1 to 25, where a K-value of 1 is usually used to calculate the accuracy of the LM on the task. These figures help bring a crucial observation to light: the models performing better on the task overall (Afro-XLM-R and XLM-R-large) in fact **do not** correlate better with human answers. Simply training *better* LMs is evidently not the way to put forward better models for estimating readability. This observation corroborates the findings by Oh and Linzen (2025), who suggest more research needs to be done on training more cognitively plausible LMs.

Results per genre In Table 3 we show the accuracy (K=1) and correlations of the best correlating model, afRoBERTa, per text genre. It turns out that the metrics do not differ much from each other per text type. On the one hand this is good news: using LMs in this way is a robust method that is not easily thrown off by texts in a different style. On the other hand, one would assume that these texts would have different readability levels – something that LMs do not automatically capture yet. Determining how to implement this effectively remains a key challenge for future research.

[5] Detailed annotation instructions are shown in Table 7 in Appendix C.

[6] See Figure 5 in Appendix B for a confusion matrix between two annotators.

Genre	Inst.	LM acc.	Hum acc.	r	r^2	ρ
Consent	1,340	0.51	0.56	0.61	0.37	0.59
eNewsletters	1,321	0.46	0.51	0.60	0.36	0.59
Government	1,628	0.51	0.45	0.62	0.39	0.61
Health	1,606	0.42	0.51	0.59	0.34	0.57
Insurance	1,676	0.46	0.50	0.55	0.30	0.55
Magazine	1,643	0.45	0.53	0.57	0.33	0.57
Newspaper	1,583	0.50	0.52	0.56	0.31	0.55

Table 3: Scores per text genre for afRoBERTa, using the probability method to calculate the correlations. "Inst" denotes the number of cloze items that were filled, and K=1 for LM accuracy.

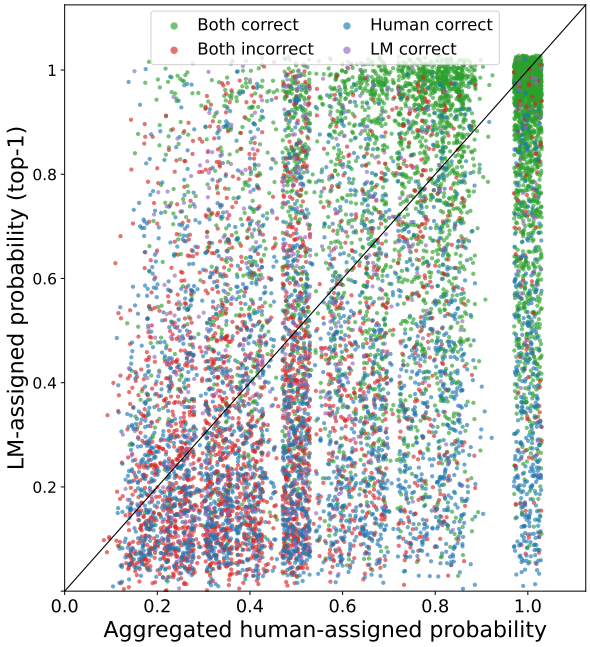


Figure 3: Comparison of LM (afRoBERTa) and aggregated human probabilities for the plotted instances. The black line marks the points where the aggregated human collective and LM assign the same probability to their respective top-1 predictions.

Human and LM probabilities To explore how human and LM probabilities relate, we follow Goldstein et al. (2022) and aggregate human and LM (afRoBERTa) expectations for each cloze item in Figure 3. The x-axis shows human-assigned probability for the most frequent human response: for each cloze item, we aggregate the available human responses (varying from 2 to 10) and define the probability as the proportion of participants who produced the most frequent answer. The y-axis shows the LM-assigned probability of its top-1 prediction (K=1). Note that the plotted probabilities reflect confidence, not accuracy. Accuracy is assessed separately by comparing the most frequent human answer and the top-1 prediction of the LM to the gold cloze answer.[7] Of the 10,797 plotted instances, both LM and humans are correct in 4,643 cases (43%, green); both are incorrect in 2,714 cases (25%, red), the human answers only are correct in 2,974 cases (28%, blue) and the LM only is correct in 466 cases (4%, purple). This shows that, while individual humans have a very similar accuracy to LMs (see Table 3), aggregated groups of humans are clearly more accurate.

	#	Iterative 1	Iterative 5	Iterative 10	First 1	First 5	First 10
afRoBERTa	1,039	0.1	0.3	0.6	1.7	4.1	5.4
Afro-XLM-R	2,644	2.0	3.4	3.9	2.6	8.8	11.8
XLM-R-base	2,644	1.4	2.5	3.3	2.1	6.6	9.3
XLM-R-large	2,644	2.2	3.3	4.6	3.2	8.6	12.8
mmBERT	3,559	0.7	1.0	1.6	1.3	3.7	6.3

Table 4: Analysis of LM accuracy (%) on multi-piece gold words for different values of Top-K evaluation, evaluated across two settings.

Multi-piece accuracy The number of gold words that are split into multiple pieces depends on the tokenizer, which can differ across models. It is an extra challenge for the models to generate such multi-token words correctly. We report the accuracy of the LMs on such words in Table 4. A number of important insights emerge from the data: First, afRoBERTa has to predict considerably fewer multi-piece words, since this is the only model with a tokenizer specifically attuned to Afrikaans. Second, although the Afrikaans-specific tokenizer did not improve the model's overall performance, Figure 2a suggests that it may have contributed to the stronger correlation between the model's output and human responses. Third, the accuracy for multi-piece words is still quite low, even when the forgiving 'first' setting is applied, where models need to predict the first piece only. We interpret this as showing that the technicality of having to predict multi-piece tokens does not influence the analysis negatively. In fact, it is expected that it would be hard for an LM to get multi-piece tokens right, independent of the technical implementation. These words are by definition relatively obscure, since they did not get merged to a single token during training, and are therefore harder to predict anyway.

[7] Since aggregated human probabilities are based on small and varying numbers of responses, the x-axis values are discrete, which produces visible stripes containing many points on top of each other. We add a random variation (jitter) of up to 0.03 to the data points to improve visibility. The original plot is available as Figure 4 in Appendix A.

Category	All instances		Both wrong	
	Human	LM	Human	LM
Match, with spelling diff	0.5	0.2	2.0	1.0
Synonym	4.5	2.7	9.1	6.1
Plausible fit in context	13.3	11.7	41.8	21.4
Alternative fit	6.1	11.4	22.5	24.5
Incorrect	8.0	28.7	24.5	46.9
Exact match	67.6	45.2	0.0	0.0

Table 5: Percentage of annotations per category, split by human or LM answers. The two rightmost columns show the percentages on a subset (98 instances) where both the human as well as the LM initially answered the cloze item wrong.

Tag	#	LM acc.	Hum acc.	r	r^2	ρ
ADJ	735	0.20	0.27	0.48	0.23	0.44
ADP	1,453	0.69	0.63	0.49	0.24	0.47
ADV	721	0.37	0.43	0.62	0.38	0.60
AUX	903	0.73	0.69	0.42	0.17	0.39
CCONJ	470	0.61	0.62	0.49	0.24	0.46
DET	1,192	0.63	0.68	0.33	0.11	0.33
NOUN	2,289	0.21	0.31	0.48	0.23	0.48
PART	366	0.93	0.87	0.28	0.08	0.25
PRON	1,082	0.55	0.60	0.42	0.18	0.41
SCONJ	269	0.58	0.54	0.54	0.29	0.54
VERB	1,255	0.31	0.41	0.51	0.26	0.50

Table 6: UPOS analysis with LM (afRoBERTa) and human accuracies.

4.1. Human Evaluation

Our analysis shows strong correlation when both an LM and all humans are wrong. But humans and LMs can also be wrong in different ways. Of the 2,714 cases where afRoBERTa as well as the aggregated humans answered incorrectly, they answered with the same exact word in 614 instances (23%). To gain a deeper understanding of the differences in mistakes humans and afRoBERTa made, we had a subset of human and LM cloze answers annotated based on the annotation scheme outlined in Section 3.5.

Since we only annotate instances where there is no exact string match, there were 122 human cloze answers and 206 LM cloze answers to annotate. Table 5 shows the combined annotation results. There is a clear trend here: the aggregated humans were much more often correct than the LM (AfRoBERTa), and also much less often completely incorrect. When the humans and LM were both wrong according to the initial Exact scoring method (the two rightmost columns in Table 5), we observe that the humans are more often at least close to the correct answer, in particular answering more frequently with an option that would be a plausible fit. This analysis shows that, in cases where the cloze-test performance of humans and LM correlate strongly because both were wrong, LM answers tend to be further away from the correct answer than human answers for the same cloze items.

4.2. Fine-grained analysis

To make sure that the correlation we found is not only driven by a few high-correlating word classes, we use Stanza (Qi et al., 2020) to tag the word categories of the gold data set and correlate those with the human and LM cloze answers.[8] Table 6 shows the results per UPOS tag for the afRoBERTa model.[9] It is clear that humans and the LM find the same words easy or difficult: the largest difference in LM and human accuracy per UPOS tag is only 10%. The overall correlation between the accuracy of human answers and that of the LM is not caused by a small number of word categories correlating very strongly – there is at least a modest correlation for all part-of-speech groups.

The part-of-speech results also correspond to the intuition that an LM would be better at getting function words right, whereas humans would be better at providing the correct content words as cloze answers (Bachman, 1985; Xie et al., 2018; Goldstein et al., 2022). Function words, such as determiners, prepositions, pronouns, conjunctions, and particles, follow distinct patterns and are constrained by grammar, making it more likely for an LM to guess them correctly. In contrast, content words, such as nouns, main verbs, adjectives and adverbs, are less constrained and carry more meaning – something that humans would pick up on better.

The LM indeed performed the worst on adjectives (20%), nouns (21%) and verbs (31%). While humans clearly do better, they were still at most 10% better than the LM. The LM outperformed humans on the functional words on average by 5% (UPOS classes adpositions, auxiliaries and particles), with one exception: determiners (DET). For language-specific XPOS classes (see Table 8 in Appendix D), LM accuracy is in fact 17% higher for the definite article (LB) category. However, the LM could not get any of the 262 indefinite articles (LO) instances right, even after the relaxation of the evaluation of those articles. This category is entirely made up of the indefinite article *'n* in Afrikaans, a contraction of its historical Dutch form *een*. As it turns out, the afRoBERTa model curiously does not have this word as a single token in its vocabulary. Because predicting multiple tokens correctly is something all LMs struggle with in our setup, this technical issue is likely the reason for the bad performance on this XPOS tag.[10]

[8] Tagging accuracy is reported as 98.6% for UPOS and 95.8% for XPOS (Stanford NLP Group, 2024).

[9] The full forms for the abbreviations are available here: https://universaldependencies.org/u/pos/

[10] The developer confirmed *'n* was in the training data, but could not clarify why it was not a single token.

5. Conclusion

In this paper, we lay the groundwork for developing a high-quality readability assessment system for Afrikaans. We are the first to use a data set of cloze tests filled in by humans to investigate correlation with LMs on said data for a language other than English. We find that the LMs in this study indeed correlate well with human answers, showing their potential for automatically assessing readability. The correlation is also not driven by only a few word classes correlating strongly; rather it is spread relatively evenly over all word classes. We further show that, in cases where the cloze-test performance of humans and an LM correlate strongly because both were wrong, LM answers tend to be further off than human answers for the same cloze items. It is noteworthy that the model with the best correlation is neither the largest nor the most accurate model, but a smaller monolingual model – an observation that highlights the need for smaller, monolingual LMs.

In future work, we aim to establish a large corpus of Afrikaans texts across various levels of readability, to train and evaluate high-quality systems that can automatically determine the readability of any given text in Afrikaans.

6. Limitations

This research is only a starting point for developing a high-quality readability assessment system for Afrikaans. The study is conducted on a data set in Afrikaans, a choice that is apt for our current purpose, but does limit the study, as other languages are excluded. The data set is also not representative of all texts available in Afrikaans. We further chose to experiment with encoder-only language models, which is sufficient for our current purpose, but there are indeed many more models that could be experimented with. The annotation of the subset of cloze answers could also have been more extensive, but served its purpose for our exploration.

7. Acknowledgments

We would like to thank Carel Jansen for initiating the collection of the human cloze test data set, and express appreciation for his input on this paper. We would further like to acknowledge our anonymous reviewers for their valuable feedback on previous versions of the paper. We are also very grateful to our annotators. Funding from Stellenbosch University, the University of Groningen and the Van Ewijck Foundation made this study possible.

8. Bibliographical References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. SERENGETI: Massively multilingual language models for Africa. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537, Toronto, Canada. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Liesbeth Augustinus, Peter Dirix, Daniel van Niekerk, Ineke Schuurman, Vincent Vandeghinste, Frank Van Eynde, and Gerhard van Huyssteen. 2016. AfriBooms: An online treebank for Afrikaans. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 677–682, Portorož, Slovenia. European Language Resources Association (ELRA).

Lyle F Bachman. 1985. Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, 19(3):535–556.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele Benchmark: A parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Marc Benzahra and François Yvon. 2019. Measuring text readability with machine comprehension: a pilot study. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 412–422, Florence, Italy. Association for Computational Linguistics.

John R. Bormuth. 1967. The cloze readability procedure: A review of research on its use for evaluating instructional materials. Research Report ED010983, University of California, Los Angeles, CRESST / Centre for the Study of Evaluation.

John R. Bormuth. 1968. Cloze as a measure of readability: Criterion-reference scores. *Yearbook of the International Reading Association*, 17:303–317.

Patricia L Carrell, Joan J. Carson, and Dong Zhe. 1993. First and second language reading strategies: Evidence from cloze. *Reading in a Foreign Language*, 10(1):953–65.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jac Conradie and Anna Coetzee. 2014. *47. Afrikaans*, pages 897–918. De Gruyter Mouton, Berlin, Boston.

Scott A Crossley, Stephen Skalicky, and Mihai Dascalu. 2019. Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, 42(3-4):541–561.

Achmat Davids. 1994. Afrikaans—die produk van akkulturasie. In G. Olivier and A. Coetzee, editors, *Nuwe perspektiewe op die geskiedenis van Afrikaans*, pages 110–119. Southern Boekuitgewers, Pretoria.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Dirix. 2023. The need for a large(r) Afrikaans treebank. *Stellenbosch Papers in Linguistics Plus*, 67.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of the Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

William H. DuBay. 2004. The principles of readability. Report, Impact Information, Costa Mesa, CA. Available at ERIC. Accessed 20 September 2022.

Roald Eiselen. 2023. NCHLT Afrikaans RoBERTa language model. SADiLaR Language Resource Repository, License: Creative Commons Attribution 4.0 International (CC-BY 4.0). Accessed 5 Oct 2023.

Roald Eiselen and Tanja Gaustad. 2023. Deep learning and low-resource languages: How much data is enough? A case study of three linguistically distinct South African languages. In *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 42–53. Association for Computational Linguistics.

Thomas François. 2015. When readability meets computational linguistics: A new paradigm in readability. *Revue française de linguistique appliquée*, (2):79–97.

Anna S. Gellert and Carsten Elbro. 2013. Activation of background knowledge for inference making: Effects on reading comprehension. *Scientific Studies of Reading*, 17(5):435–452.

A. Goldstein, Z. Zada, E. Buchnik, M. Schain, A. Price, B. Aubrey, S. A. Nastase, A. Feder, D. Emanuel, A. Cohen, A. Jansen, H. Gazula, G. Choe, A. Rao, C. Kim, C. Casto, L. Fanda, W. Doyle, D. Friedman, P. Dugan, L. Melloni, R. Reichart, S. Devore, A. Flinker, L. Hasenfratz, O. Levy, A. Hassidim, M. P. Brenner, Y. Matias, K. A. Norman, O. Devinsky, and U. Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25:369–380.

K. S. Goodman. 1967. Reading: A psycholinguistic guessing game. *Journal of the Reading Specialist*, 6:126–135.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28*, pages 1693–1701.

Markus J Hofmann, Steffen Remus, Chris Biemann, Ralph Radach, and Lars Kuchinke. 2022. Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, 4:730570.

Raymond Joseph Horton. 1974. The construct validity of cloze procedure: An exploratory factor analysis of cloze, paragraph reading, and structure-of-intellect tests. *Reading Research Quarterly*, 10(2):248–251.

Cassandra L Jacobs, Loïc Grobol, and Alvin Tsang. 2024. Large-scale cloze evaluation reveals that token prediction tasks are neither lexically nor semantically aligned. *arXiv preprint arXiv:2410.12057*.

Carel Jansen, Rose Richards, and Liezl Van Zyl. 2017. Evaluating four readability formulas for Afrikaans. *Stellenbosch Papers in Linguistics Plus*, 53:149–166.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020a. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020b. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Tom Kalinsky, Assaf Kasirer, and Yoav Goldberg. 2023. Simple and effective multi-token completion from masked language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2345–2359. Association for Computational Linguistics.

Jan-Christoph Kalo and Leandra Fichtel. 2022. Kamel: Knowledge analysis with multitoken entities in language models. In *Automated Knowledge Base Construction (AKBC)*.

Judith M. H. Kamalski. 2007. *Coherence Marking, Comprehension and Persuasion: On the Processing and Representation of Discourse*. Ph.D. thesis, Utrecht University.

Suzanne Kleijn. 2018. *Clozing in on readability: How linguistic features affect and predict text comprehension and on-line processing*. Ph.D. thesis, Utrecht University.

Miyoko Kobayashi. 2002. Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *The Modern Language Journal*, 86(4):571–586.

Ernst Kotzé. 2018. Die klassifikasie van Afrikaans. LitNet (Menings). Published 18 April 2018. Accessed: 2026-02-16.

Anna Laurinavichyute, Irina A Sekerina, Kristina Bagdasaryan, Svetlana Alexeeva, and Nikita Zmanovksy. 2017. Russian sentence corpus: Benchmark measures of eye movements in reading in cyrillic. *International Journal of Corpus Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adrielli Tina Lopes Rego, Joshua Snell, and Martijn Meeter. 2024. Language models outperform cloze predictability in a cognitive model of reading. *PLOS Computational Biology*, 20(9):e1012117.

Susan Lotz, Bo Blankers, Rik van Noord, and Carel Jansen. forthcoming. Readability assessment in Afrikaans: Cloze scores, linguistic features and cross-validated linear regression. *Southern African Linguistics and Applied Language Studies*, forthcoming (preprint).

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmBERT: A modern multilingual encoder with annealed language learning. *arXiv preprint arXiv:2509.06888*.

Cindy McKellar. 2022. Autshumato English-Afrikaans parallel corpora. SADiLaR Language Resource Repository.

Tebatso Gorgina Moape, Fulufhelo Mthombeni, and Annemie Stoman. 2025. Evaluating the impact of data scarcity on model performance in a low-resource Afrikaans question answering model. In *2025 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–6.

Anastasia Nikiforova, Sergey Pletenev, Daria Sinitsyna, Semen Sorokin, Anastasia Lopukhina, and Nick Howell. 2020. Language models for cloze

task answer generation in Russian. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 28–37, Marseille, France. European Language Resources Association.

Byung-Doh Oh and Tal Linzen. 2025. To model human linguistic prediction, make LLMs less superhuman. *arXiv preprint arXiv:2510.05141*.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

Andrew M Olney. 2022. Assessing readability by filling cloze items with transformers. In *International Conference on Artificial Intelligence in Education*, pages 307–318. Springer.

JM O'Toole and RAR King. 2011. The deceptive mean: Conceptual scoring of cloze entries differentially advantages more able readers. *Language Testing*, 28(1):127–144.

PanSALB. 2021. The status of Afrikaans as an indigenous South African language. Accessed: 2026-02-16.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Daniele Puccinelli, Silvia Demartini, and Pier Luigi Ferrari. 2021. Tackling Italian University assessment tests with transformer-based language models. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, pages 404–409, Milan, Italy. CEUR Workshop Proceedings.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

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(1).

SADiLaR Language Resource Repository. Search results for Afrikaans. Accessed: 2026-02-17.

Emily Sadlier-Brown, Millie Lou, Miikka Silfverberg, and Carla Kam. 2024. How useful is context, actually? Comparing LLMs and humans on discourse marker prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 231–241, Bangkok, Thailand. Association for Computational Linguistics.

C. Shain, C. Meister, T. Pimentel, R. Cotterell, and R. Lévy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121.

Stanford NLP Group. 2024. Model performance – Stanza. Accessed: 19 February 2026.

W. L. Taylor. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Virtuele Instituut vir Afrikaans (VivA). Korpusportaal: Oop (explore corpus). Accessed: 2026-02-17.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331, St. Julian's, Malta. Association for Computational Linguistics.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale cloze test dataset created by teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2344–2356, Brussels, Belgium. Association for Computational Linguistics.

A. Original scatterplot

Figure 4 shows the scatter plot in Figure 3 without any jitter to improve visualization. If too many human respondents skipped an item, that item could not be plotted, which resulted in fewer plotted instances than the total number of cloze items.

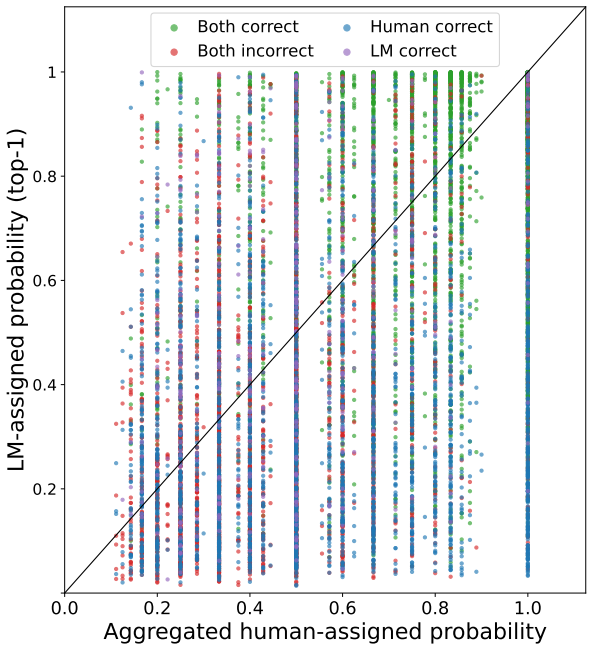


Figure 4: Original scatter plot: Comparison of LM (afRoBERTa) and aggregated human probabilities for the plotted instances. The black line marks the points where the aggregated human collective and LM assign the same probability to their respective top-1 predictions.

B. Confusion Matrix

Figure 5 shows a confusion matrix for Annotators 1 and 2. It shows the general trend, also observed across other annotators, that annotators often agreed on the general severity or closeness of fit, but not always on the exact cut-off point between neighboring categories.

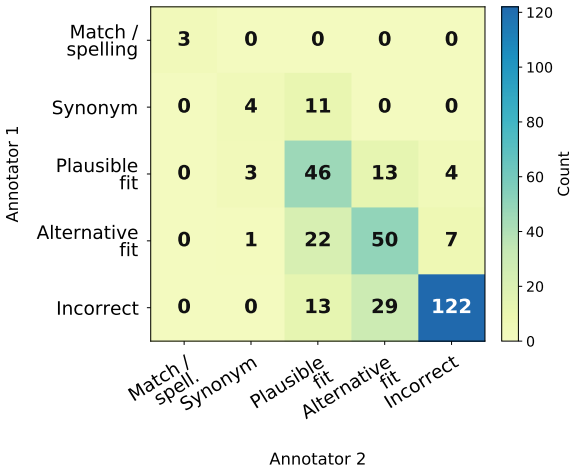


Figure 5: Confusion matrix for annotations by Annotators 1 and 2 (Pairwise Cohen's κ_w = 0.64), showing some confusion with adjacent categories.

Annotation option	Extra instructions
1. Match with spelling/other difference	The response differs in spelling or form, but it is still clearly the gold answer.
2. Synonym	Can the gold answer be replaced with this response without changing the meaning of the sentence or context?
3. Plausible fit in context *(semantically close, but not exact; syntactically acceptable*)	Can the gold answer be replaced with this response so that the sentence remains well-formed and keeps the same broad meaning, even though the word is not a synonym?
4. Alternative fit (different meaning) *(syntactically acceptable, but may deviate in meaning)*	Can the gold answer be replaced with this response so that the sentence remains well-formed and meaningful, even if the intended meaning or contextual fit changes?
5. Incorrect (nonsensical sentence)	The response makes the sentence ungrammatical and/or nonsensical.

Table 7: Extra instructions for the annotators.

C. Detailed annotation instructions

Table 7 shows the more detailed annotation instructions per category that were given to the four annotators.

D. Fine-grained XPOS results

Table 8 shows the comparison between LM and human performance per XPOS tag.

Tag	#	LM acc.	Hum acc.	r	r^2	ρ
ASA	526	0.17	0.24	0.43	0.18	0.37
ASP	163	0.31	0.36	0.55	0.30	0.57
BS	562	0.34	0.38	0.61	0.38	0.58
KN	470	0.61	0.62	0.49	0.24	0.46
KO	241	0.62	0.55	0.53	0.29	0.53
LB	755	0.93	0.76	0.29	0.08	0.26
LO	262	0.00	0.60	-	-	-
NA	351	0.18	0.26	0.41	0.17	0.42
NM	152	0.24	0.38	0.41	0.17	0.44
NSE	1,105	0.25	0.35	0.50	0.25	0.49
NSM	672	0.17	0.27	0.48	0.23	0.47
PB	256	0.59	0.59	0.46	0.21	0.45
PDOENP	124	0.73	0.76	0.19	0.04	0.22
PTENP	116	0.83	0.73	0.40	0.16	0.32
SVS	1,431	0.68	0.63	0.49	0.24	0.47
UPI	211	0.96	0.88	0.20	0.04	0.17
VTHOG	693	0.28	0.37	0.46	0.21	0.45
VTHOK	251	0.76	0.70	0.50	0.25	0.46
VTHOO	190	0.43	0.56	0.58	0.33	0.57
VTUOM	298	0.57	0.59	0.36	0.13	0.33
VTUOP	131	0.88	0.80	0.37	0.14	0.27
VUOT	163	0.87	0.84	0.26	0.07	0.30
VVHOG	200	0.30	0.40	0.51	0.26	0.52

Table 8: XPOS analysis with LM and human accuracies. The full forms for the abbreviations are available here: https://www.sketchengine.eu/afrikaans-part-of-speech-tagset/