LuxDiagRC: A Diagnostic Reading Comprehension Corpus for Luxembourgish with Linguistic and Cognitive Annotation Layers

Christophe Friezas Gonçalves and Salima Lamsiyah and Christoph Schommer
Department of Computer Science, Faculty of Science, Technology and Medicine
University of Luxembourg, Esch-sur-Alzette, Luxembourg
{christophe.friezas,salima.lamsiyah,christoph.schommer}@uni.lu

Abstract

Reading comprehension resources for low-resource languages remain limited, particularly datasets designed for educational assessment and diagnostic analysis in contrast to binary correctness. We present a diagnostically rich reading comprehension corpus for Luxembourgish, annotated using a two-layer framework that separates linguistic sources of textual difficulty from cognitive and diagnostic properties of comprehension questions. The linguistic layer captures span-level lexical, syntactic, morphological, and discourse-related features, while the cognitive layer annotates multiple-choice questions according to the PIRLS cognitive processes and diagnostically meaningful distractor types following the STARC framework. This design enables fine-grained analysis of reading comprehension errors by linking response patterns to underlying linguistic phenomena. The resulting corpus consists of 640 multiple-choice questions based on 16 annotated Luxembourgish texts. We describe the annotation methodology agreement measures, and will release the dataset as a publicly available resource for educational and low-resource NLP research.

1 Introduction

Reading comprehension datasets have played a central role in advancing natural language processing (NLP), particularly for high-resource languages, where large-scale benchmarks such as SQuAD and related resources have driven progress in language understanding and educational applications (Rajpurkar et al., 2016; Lai et al., 2017; Singh et al., 2024).

In contrast, comparable reading comprehension resources for low-resource languages remain scarce, both in terms of data volume and annotation depth. This scarcity limits the development, evaluation, and diagnostic analysis of NLP models and educational technologies in such languages and has been identified as a central challenge in low-resource NLP research (Magueresse et al., 2020).

Luxembourgish is a low-resource language spoken by approximately 400,000 people worldwide (Plum et al., 2025) and is classified as vulnerable by UNESCO.[1] In recent years, this status has motivated increased research attention, including work on language modeling, proficiency assessment, and educational applications for Luxembourgish (Plum et al., 2025; Nouzri et al., 2025; Lothritz et al., 2025). Despite these efforts, the availability of high-quality, annotated corpora remains a major bottleneck for systematic linguistic analysis and downstream NLP tasks.

This limitation is particularly present in the domain of reading comprehension and language acquisition. Most existing reading comprehension datasets, especially those widely used in NLP, conceptualize comprehension as a binary decision problem, where responses are evaluated solely as correct or incorrect (Rajpurkar et al., 2016). While such datasets are effective for benchmarking model performance, they provide limited insight into the underlying causes of comprehension errors. From an educational and diagnostic perspective, this lack of structured error information restricts their usefulness for analyzing learner behavior, identifying systematic sources of difficulty, and supporting fine-grained assessment (Jang, 2008; Berzak et al., 2020).

To address this gap, recent work has highlighted the importance of diagnostically informative question design, structured distractors, and explicit modeling of linguistic and cognitive sources of error (Mullis and Martin, 2019; Berzak et al., 2020). However, such approaches remain largely unexplored for low-resource languages due to annotation costs and limited expert availability. In this

[1] https://unesdoc.unesco.org/ark:/48223/pf0000192416

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 532–541
March 29, 2026

paper, we introduce a diagnostically rich reading comprehension corpus for Luxembourgish, constructed from texts used in secondary education and annotated using a two-layer framework. The first layer captures linguistic features of the source texts that are known to contribute to comprehension difficulty, while the second layer annotates multiple-choice questions and answer options according to cognitive processes defined by the Progress in International Reading Literacy Study (PIRLS) framework (Mullis and Martin, 2019) and diagnostically meaningful error types following the Structured Annotations for Reading Comprehension (STARC) paradigm (Berzak et al., 2020). By explicitly linking question design and response patterns to annotated linguistic features in the source texts, the resulting dataset supports fine-grained analysis of reading comprehension behavior beyond simple accuracy measures.

To summarize, the main contributions of this work are as follows:

- We present **LuxDiagRC**, a publicly available reading comprehension corpus of 640 items for Luxembourgish, addressing a gap in resources for low-resource languages.
- We propose a two-layer annotation framework that integrates linguistic feature annotation with cognitive and diagnostic question annotation.
- We describe the annotation workflow and inter-annotator agreement procedures used to ensure the reliability of the dataset.
- We will publicly release the dataset to support future research on Luxembourgish reading comprehension.

Together, these contributions provide a foundation for future research in educational NLP, reading comprehension assessment, and low-resource language processing.

2 Related Work

We structure the related work according to the main themes addressed in this paper. Specifically, we focus on three core topics: (i) reading comprehension datasets, (ii) diagnostic assessment and structured question design, and (iii) Luxembourgish and low-resource NLP.

Reading Comprehension Datasets Reading comprehension (RC) datasets have been central to the development and evaluation of NLP models, particularly for high-resource languages. Large-scale benchmarks such as SQuAD (Rajpurkar et al., 2016) and RACE (Lai et al., 2017) have driven substantial progress in both extractive and multiple-choice question answering. More recent domain-specific datasets, including MedMCQA (Pal et al., 2022) and SciDQA (Singh et al., 2024), have extended RC evaluation to specialized domains such as medicine and scientific literature. However, despite their scale and diversity, these datasets typically frame reading comprehension as a binary decision problem, evaluating responses as correct or incorrect. As a result, they provide limited insight into the underlying sources of comprehension errors and are less suitable for diagnostic analysis, particularly in educational settings where understanding learner difficulties is essential (Berzak et al., 2020).

Diagnostic Assessment and Structured Question Design To address the limitations of accuracy-based evaluation, prior work has emphasized the importance of diagnostically informative question design and structured distractors. The Structured Annotations for Reading Comprehension (STARC) framework explicitly models different types of incorrect answers, enabling fine-grained analysis of comprehension failures (Berzak et al., 2020). STARC has been applied across multiple domains, including reading comprehension, temporal reasoning, and music-related texts (Zhang et al., 2025, 2024; Weck et al., 2024; Shubi et al., 2024), and has been shown to support interpretable evaluation of both human and model behavior. In parallel, the Progress in International Reading Literacy Study (PIRLS) provides a widely adopted educational framework that categorizes reading comprehension questions according to cognitive processes such as retrieval, interpretation, inference, and evaluation (Mullis and Martin, 2019; Bulté et al., 2025). PIRLS has been extensively used across Europe to assess reading literacy and inform educational standards (García-Crespo et al., 2021; Papadopoulos et al., 2021; Kennedy and Strietholt, 2023).

The construction of multiple-choice reading comprehension questions also poses challenges for automatic question generation. Prior work has identified limitations related to distractor quality, cognitive alignment, and pedagogical validity (Du-

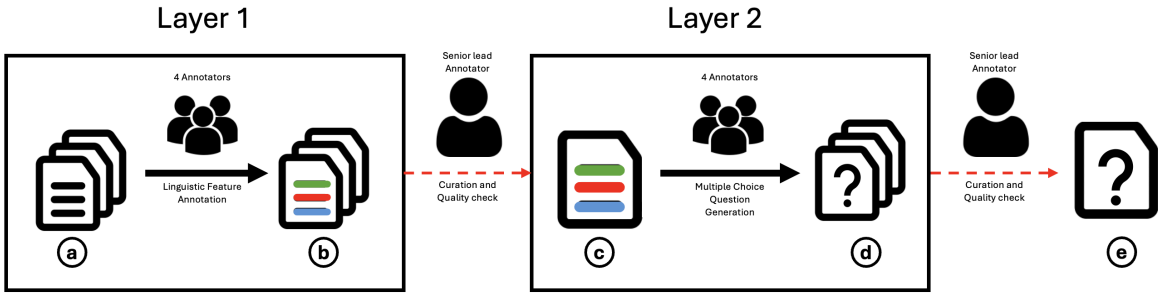

Figure 1: Overview of the Annotation pipeline. The raw Luxembourgish texts (**a**) are inputed into Layer 1. After the annotation process, 4 bundles of individually annotated texts (**b**) pass through a senior annotator for curation. Layer 2 uses the unified and curated emerging text (**c**) as input for the MCQ generation. This process culminates in 4 bundles of 200 questions (**d**). After curation of these bundles the final dataset (**e**) is set.

tulescu et al., 2025; Zeinalipour et al., 2025, 2024). Dutulescu et al. (2025) proposed a reasoning-enhanced approach to multiple-choice question (MCQ) generation in English, demonstrating the reliance of current methods on large, high-quality datasets. Similarly, Hwang et al. (2023) used GPT-3.5 to generate MCQs aligned with Bloom's Taxonomy (Bloom et al., 1956) and observed a degradation in question quality as the cognitive complexity of questions increased. These findings further motivate the need for expert-designed, diagnostically grounded datasets for educational applications.

Luxembourgish and Low-Resource NLP Research on Luxembourgish NLP has gained momentum in recent years, particularly following the emergence of large language models. Nouzri et al. (2025) explored a multi-agent tutoring approach for Luxembourgish language acquisition, highlighting both the scarcity of task-specific datasets and the multilingual context faced by learners. Their work emphasizes the need for human-in-the-loop approaches and specialized datasets for pedagogical tasks. Similarly, Lothritz et al. (2025) evaluated the proficiency of state-of-the-art LLMs in Luxembourgish across CEFR levels, reporting uneven performance and underscoring the challenges posed by code-switching and cross-lingual interference. Beyond Luxembourgish, low-resource NLP research has consistently shown that progress is constrained not only by data quantity but also by the lack of high-quality, task-specific annotations (Magueresse et al., 2020). In the context of language acquisition, linguistic complexity and difficulty play a crucial role. Bulté et al. (2025) distinguish between structural complexity and learner-oriented difficulty, a distinction that has been shown to impact second language acquisition. Subsequent studies applying these definitions and applying targeted methodology and taxonomies have demonstrated their effectiveness in second language proficiency (Yamazaki and Hiver, 2024; Sung et al., 2025; Lu, 2025).

Taken together, this prior work motivates the need for diagnostically rich reading comprehension resources for low-resource languages. Our work builds on established educational and NLP frameworks while addressing a clear gap in the availability of deeply annotated, diagnostic evaluation datasets for Luxembourgish.

3 Annotation Framework

This section describes the annotation framework used to construct the Luxembourgish reading comprehension dataset. The framework consists of two interdependent layers: a linguistic feature annotation layer (Layer 1) and a cognitive and diagnostic annotation layer (Layer 2). Layer 1 captures linguistic phenomena in the source texts that are known to contribute to comprehension difficulty, while Layer 2 annotates reading comprehension questions and answer options in terms of cognitive processes and diagnostically meaningful error types.

By explicitly linking question design and response behavior to underlying linguistic features, the framework enables fine-grained analysis of reading comprehension errors. Figure 1 provides an overview of the annotation pipeline.

3.1 Layer 1: Linguistic Feature Annotation

Layer 1 annotates linguistic phenomena in the source texts that are known to affect reading comprehension, particularly for learners of Luxembourgish as a second language. Annotations are applied at the span level and are intended to capture lexi-

cal, syntactic, morphological, and discourse-related sources of processing difficulty.

The selection of linguistic features is informed by prior work on linguistic complexity and learner difficulty, as well as by established annotation guidelines for educational corpora (Housen et al., 2012; Ide and Pustejovsky, 2017; Bulté et al., 2025). The annotation schema is organized into three categories: lexical–semantic, syntactic–morphological, and discourse–orthographic. Table 1 summarizes the full set of linguistic feature tags used in the dataset.

Luxembourgish developed from a West Germanic dialect and has been historically influenced by both German and French (SIP). This multilingual background motivates the inclusion of lexical tags capturing loanwords and cross-lingual interference, such as `LEX-LOAN-FR`, `LEX-LOAN-DE`, and `LEX-FALSE-FRIEND`.

Additional lexical-semantic tags target language-specific difficulties, including idiomatic expressions and rare or archaic vocabulary that may not be familiar to learners. Syntactic and morphological tags focus on structural phenomena known to increase processing load, including verb-final clauses, separable-prefix verbs, and embedded constructions. The `MORPH-N-RULE` tag captures instances of the *Eifeler Regel*, a Luxembourgish-specific morphophonological rule that poses difficulties for new learners (Moulin and Nübling, 2006).

Finally, discourse and orthographic tags capture ambiguity in pronoun reference and divergences between written and spoken forms, which represent additional challenges for non-native readers.

3.2 Layer 2: Cognitive and Diagnostic Annotation

Layer 2 annotates reading comprehension questions and answer options. Question design follows the PIRLS reading comprehension framework (Mullis and Martin, 2019), which distinguishes four cognitive processes: *Retrieve*, *Interpret*, *Infer*, and *Evaluate*.

Each text is associated with ten multiple-choice questions per annotator distributed across these four cognitive categories to ensure balanced coverage of both surface-level and higher-order comprehension skills. The distribution follows PIRLS recommendations [2] and is summarized in Table 2.

[2]https://www.iea.nl/studies/iea/pirls

Cognitive Process	Proportion (%)	Questions per Text
Retrieve	20	2
Interpret	30	3
Inferential	30	3
Evaluate	20	2

Table 2: Distribution of reading comprehension questions across cognitive processes, following the PIRLS framework.

Beyond cognitive categorization, answer options are designed according to the STARC framework (Berzak et al., 2020), which emphasizes diagnostically meaningful distractors. For each question, annotators first identify a *critical span* in the text on which they base their question on and from which stems the correct answer. The remaining answer options are constructed as: (i) a plausible misunderstanding of the critical span, (ii) a distractor derived from a different but irrelevant span in the text, and (iii) a plausible answer with no textual support.

To increase diagnostic value, annotators were instructed to select critical spans containing a minimum of three linguistic feature tags from Layer 1. Distractor spans were required to contain at least one linguistic feature tag. This design explicitly links observed comprehension errors to specific linguistic phenomena.

Table 3 gives an example of a created question and corresponding answers to clarify the annotation process and desired question complexity. The given critical span has fourteen linguistic tags and the distractor span six. The *misunderstanding* answer targets the LEX-FALSE-FRIEND, MORPH-N-RULE and ORTHO-PHONO-DIVERGE tagged tokens in the critical span as causes for misunderstanding. The *Distractor span* answer uses the cognitive load created by the LEX-FALSE-FRIEND and LEX-RARE-ZLS to increase the distractive factor if the answer. The *no support* answer is unrelated and checks a general assumption of the reader.

4 Annotation Procedure and Agreement

This section describes the annotation workflow, including the selection and training of annotators, the tools and resources used during annotation, and the procedures applied to ensure annotation reliability. In addition, we report inter-annotator agreement measures and describe the adjudication process used to produce the final gold-standard annotations.

Table 1: Linguistic feature annotation tags used in the dataset, grouped by category.

Tag ID	Description
Lexical and Semantic Features	
LEX-LOAN-FR	Word of clear French origin used in Luxembourgish
LEX-LOAN-DE	Word of clear German origin used in Luxembourgish
LEX-FALSE-FRIEND	Lexical item prone to cross-lingual confusion for multilingual speakers
LEX-IDIOM-ZLS	Idiomatic expression or typical Luxembourgish saying
LEX-RARE-ZLS	Rare, archaic, or low-frequency Luxembourgish word
Syntactic and Morphological Features	
SYN-VERB-FINAL	Subordinate clause with verb-final word order
SYN-VERB-SEP	Separable-prefix verb with split verbal construction
SYN-EMBEDDED	Embedded or nested relative clause
MORPH-N-RULE	Instance of the *Eifeler Regel* ("n-Reegel") morphophonological rule
Discourse and Orthographic Features	
DISC-COREF-AMBIG	Pronoun with distant or ambiguous referent
ORTHO-PHONO-DIVERGE	Word exhibiting divergence between written and spoken forms

Table 3: An example question of the type *Interpret* taken from the final dataset. For the text spans depicted, each number corresponds to the tagged linguistic feature presented in Table 1.

Question	
Wéi léisen si den Problem mam net nokommen? *How do they solve the Problem of not following?*	
Span Type	**Span**
Critical Span	No der Schoul souzen mer zesumme (9,11) beim Ralph doheem. Déi (3) ganz Clique (1) aus dem sechste (9,11) Schouljoer aus dem Brill (3). Mir hu (9,11) versicht, aus all deene (9,11) Puzzlestécker e (9,11) kompletten Text (2) ze bastelen, vun deem mer awer och guer näischt verstanen hunn (6).
Distractor Span	Mir hunn eis (3) alleguer wéi (11) Moundkaalwer (5) ugekuckt (6). Sou séier konnte (9) mir net schreiwen (6).
Answer Type	**Answer**
Correct	Si setzen sech zesummen an hëllefen sech géigensäiteg *They sit together and help each other*
Misunderstanding	Si setzen am brill ze puzzelen *They sit at brill to puzzle*
Distractor Span	Si kënnen net sou séier schréiwen *They cannot write that fast*
No Support	Si hun keng Loscht *They don't want to*

Table 4: A list of all texts used in the corpus

Title	**Author**	**Amount of Tokens**
Catherine, ech sinn esou glécklech	Cathy Clement	1139
De Mathematik-Proff	Lucien Blau	1169
De Pablo an d'Juliette	Josy Braun	830
Dräizéng	Tullio Forgiarini	540
Blues	Pol Greisch	375
Ech denken nach vill un de Mike	Cathy Clement	574
Meng éischt Zäit am Lycée	Lucien Blau	1305
Hausaufgaben – derfir, dergéint, oder wéi?	Nico Graf	508
Den Ersatzschoulmeeschter	Henri Losch	855
Schoulliewen am Krich	Roger Manderscheid	866
De Kunibert vun Hesper	Jemp Schuster	988
De Siegfried an d'Melusina	Nico Brettner	540
Poker	Nico Helminger	2120
D'Schnurreli	Josy Braun	1383
E Muerd am Gréngewald	Pol Pütz	1929
Vakanzën	Raymond Schaak	1082

4.1 Source Texts

The reading comprehension questions are grounded in a corpus of 16 Luxembourgish short stories selected from the textbooks *Lies de bal: lëtzebuergesch Texter* (2014 and 2021 editions). These textbooks are officially used in Luxembourgish secondary education and target students at both lower and upper secondary school levels.

The selected texts cover a diverse range of genres, including fables, historical narratives, and contemporary everyday-life stories. The complete corpus comprises 16,399 tokens. A detailed list of titles and authors is provided in Table 4. The texts have a mean length of 1025 tokens. The shortest text spanning 375 tokens and the longest 2120 tokens with a standard deviation of 487.53 between all texts.

The final question corpus results in 640 questions with 40 questions per text given 10 questions per annotator. Each text has 8 *Retrieve*, 12 *Interpret*, 12 *Infer*, and 8 *Evaluate* questions.

Although the corpus size may appear limited in the context of large-scale NLP, it is well suited to the goals of this work. Rather than serving as a training dataset for deep learning models, the corpus is designed as a gold-standard evaluation benchmark that prioritizes annotation depth and quality over data quantity. Its primary value lies in enabling fine-grained diagnostic analysis of reading comprehension behavior.

4.2 Annotators and Training

Four annotators participated in the dataset construction. All annotators are native speakers of Luxembourgish, hold C2-level proficiency, and completed their primary and secondary education within the Luxembourgish school system. This background

ensures extensive exposure to language norms, idiomatic usage, and educational conventions.

Prior to annotation, all annotators were trained using a detailed annotation manual. The manual specifies annotation guidelines, provides worked examples, and references authoritative, government-supported linguistic resources for Luxembourgish.

4.3 Annotation Workflow and Tools

The annotation process was organized as a multi-phase workflow to ensure annotation quality and to prevent bias between linguistic feature annotation and question generation.

Phase 1: Linguistic Feature Annotation (Layer 1). In the first phase, annotators worked independently to annotate all source texts with linguistic features using the Layer 1 schema (Table 1). Annotations were applied to the full corpus of approximately 40 pages. This phase was conducted in isolation from question generation to avoid any influence of downstream task design on linguistic annotation decisions.

Phase 2: Adjudication and Inter-Annotator Agreement. Following the independent annotation phase, inter-annotator agreement was calculated for the Layer 1 annotations (see Section 4.4). All disagreements were subsequently reviewed in adjudication sessions led by a senior annotator. These discussions resulted in a single, consolidated gold-standard version of the linguistically annotated texts. Annotation guidelines were refined based on recurring sources of disagreement identified during this process.

Phase 3: Cognitive and Diagnostic Annotation (Layer 2). In the final phase, annotators were provided with the adjudicated gold-standard Layer 1 annotations and independently generated multiple-choice reading comprehension questions following the Layer 2 schema. Question generation was explicitly guided by the linguistic feature annotations; annotators were encouraged to design questions targeting specific tagged phenomena (e.g., lexical false friends or syntactic complexity) and to construct diagnostically meaningful distractors accordingly.

All annotation phases were carried out using the INCEpTION platform (Klie et al., 2018), which supports multi-layer, span-based annotation, collaborative workflows, and built-in agreement analysis. The platform enables parallel annotation of multiple layers and is well suited for workflows that combine detailed linguistic annotation with cognitive and diagnostic labeling, as required by the framework adopted in this work.

4.4 Inter-Annotator Agreement and Adjudication

To ensure that the corpus constitutes a reliable scientific resource, all annotations were statistically validated using inter-annotator agreement (IAA) measures. IAA quantifies the extent to which multiple annotators, working independently, assign consistent annotations to the same data.

Two types of annotation tasks are involved, each requiring an appropriate agreement metric. For the linguistic feature annotation layer (Layer 1), the task consists of span-based labeling, where agreement depends on both span boundaries and assigned labels (e.g., `LEX-LOAN-FR`). Agreement for this layer was therefore measured using Krippendorff's α, which is robust to multiple annotators, missing data, and different label types (Krippendorff, 2011).

For the cognitive and diagnostic annotation layer (Layer 2), the task is of a generative nature, where all questions were checked by the senior annotator for redundancy and refined if two questions were interchangeable.

Annotation was carried out using an iterative, batch-based protocol. Texts were annotated in small batches (approximately 3 texts at a time), after which agreement scores were computed. If agreement for a given batch fell below an acceptable threshold (e.g., $\alpha < 0.7$), annotation was temporarily paused and the annotators met to discuss sources of disagreement. These disagreements typically reflected ambiguities or edge cases not fully specified in the annotation manual. The guidelines were subsequently refined, and annotators were aligned on the updated rules before proceeding to the next batch.

Following completion of the independent annotation phases, all remaining disagreements were reviewed in joint adjudication sessions led by a senior annotator. Conflicts were resolved by consensus with reference to official Luxembourgish linguistic resources. This process resulted in a single, consolidated gold-standard version of the dataset.

5 Use Cases

The proposed dataset is designed to support both educational and NLP research by enabling fine-grained analysis of reading comprehension behavior. Unlike standard reading comprehension benchmarks that focus primarily on overall accuracy, the diagnostic structure of this corpus allows errors to be interpreted in terms of their underlying linguistic and cognitive causes. In this section, we outline two primary use cases that illustrate the practical value of the dataset.

5.1 Educational Use Case: Diagnostic Profiling of Student Weaknesses

For the educational use case, we illustrate the educational support and pedagogical effectiveness by using one example question of the dataset and the different underlying conclusions that can be made based on the answers. Table 3 shows an example question with the associated critical and distractor spans in addition to the corresponding answers. The linguistic feature tags are given by reference. In this case, the question asks the student to interpret and summarize the underlying message of a given paragraph in the text.

A student could choose the *misunderstanding* distractor. A deeper look at the critical span could then be taken to analyze this choice. Said span has four close cases of the MORPH-N-RULE and ORTHO-PHONO-DIVERGE, hinting towards a general misunderstanding of the sentence if the student has issues grasping the N-rule itself. A second caveat that can be concluded and analyzed by the teacher is the combination of the N-rule and the LEX-FALSE-FRIEND word "Brill". The word can be misunderstood in German to mean glasses, whereas in this case, it represents the name of a specific place in Luxembourg. This phenomenon in addition to the N-rule could lead to a higher cognitive load in German-speaking individuals and throw them of the true answer, hinting towards language bias while reading. A bias that has to be tackled by the teacher.

The *distractor span* answer uses a span that is set some lines before the critical span. This span has a mix of different linguistic tags, a false friend, a rare word in addition to three verb related tags, increasing the cognitive load for the overall understanding of the span. For the student in our use case, if they were to choose this answer. The analyses could lead to the conclusion that the combination of a rare word and a false friend could increase the importance a student would put on the sentence without taking into account its relevance to the question at hand. This combined with the fact, that the second part of the span looks near identical to the distractor answer would bring bias to the student to choose said answer.

The *no support* answer leads to two different conclusions, either the student has not read the given text or completely misunderstood the paragraph containing the critical and distractor span altogether.

An important note to stress as seen in this example. The corpus brings higher insight the more questions a student answers. One question alone hints towards gaps in the student's capabilities yet does not solidify suspicions. The more answers a student gives the clearer the gaps become.

5.2 NLP Use Case: Diagnostic Evaluation of Language Models

In terms of NLP, the corpus allows for a deeper insight and analysis into Luxembourgish comprehension of LLMs. The corpus can be used to enhance and build onto the research done by Lothritz et al. (2025), enhancing the proficiency level accuracy by highlighting the specific gaps in comprehension in which the different models fall into. Thus, treating a model similar to a student, as seen in our previous use case, and thus by not only evaluating on numerical values but also analyzing the critical and distractor spans of committed mistakes on similarities and patterns in associated linguistic features.

An additional insight can be gained on the type of questions. The corpus is made up of 60% interpretation and inference-based questions. In other words, questions that rely on the ability of using text information to form a specific conclusion and not basic information retrieval. This fact combined with the linguistic link, creates direct pointers towards internal mechanisms of LLMs in language usage and processing.

The diversity of the dataset components leads to the additional benefit of creating specialized test subsets for language models. The linguistic features and question types can be mixed and matched to filter out questions to create specific test cases such as false friend-based questions and or orthographic to phonetically divergent questions to pinpoint said linguistic phenomena in tested models.

6 Conclusion

This work introduced a diagnostically rich reading comprehension corpus for Luxembourgish, addressing a critical gap in resources for low-resource languages. The dataset is constructed from educationally grounded texts and annotated using a two-layer framework that separates linguistic features of the source texts from cognitive and diagnostic properties of reading comprehension questions. By combining linguistic feature annotation with structured multiple-choice questions grounded in the PIRLS and STARC frameworks, the corpus enables analysis of reading comprehension behavior beyond simple accuracy-based evaluation.

The proposed resource supports both educational and NLP research. For education, it enables diagnostic assessment of learner difficulties and targeted analysis of comprehension failures. For NLP, it provides a fine-grained evaluation benchmark for probing model behavior under specific linguistic conditions in a low-resource setting. More broadly, this work demonstrates the value of deep, expert-driven annotation as an alternative to large-scale but shallow datasets, particularly for languages with limited digital resources.

Limitations and Future Work

This work introduces a novel diagnostically rich reading comprehension corpus for Luxembourgish, however, several limitations should be acknowledged. The dataset is relatively small compared to large-scale benchmarks, reflecting a deliberate focus on deep, gold-standard annotation rather than model training at scale. Accordingly, the corpus is primarily intended for diagnostic evaluation and analysis.

The annotation process relies on expert annotators with high proficiency in Luxembourgish and educational experience. While this ensures annotation quality, it limits scalability and increases the cost of dataset expansion. Moreover, although the selected linguistic features and diagnostic error types are grounded in established frameworks, they may not capture all sources of comprehension difficulty.

These limitations suggest several directions for future work. The dataset can be expanded with additional texts, genres, and proficiency levels, and the annotation framework may be adapted to other low-resource and multilingual settings. Future studies may also involve empirical evaluation with human learners to validate the diagnostic value of the annotated error types. From an NLP perspective, the dataset provides a basis for fine-grained evaluation of reading comprehension models with respect to linguistic features and distractor behavior.

Acknowledgments

We thank the four Luxembourgish native-speaker annotators for their careful work on the linguistic and question annotations, and the senior annotator for leading adjudication and curation. Their expertise was essential to producing a reliable gold-standard resource.

References

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735, Online. Association for Computational Linguistics.

B Bloom, M.D Engelhart, E.J Furst, W.H Hill, and D.R Krathwohl. 1956. *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*.

Bram Bulté, Alex Housen, and Gabriele Pallotti. 2025. Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*, 75(2):533–574.

Andreea Dutulescu, Stefan Ruseti, Denis Iorga, Mihai Dascalu, and Danielle S. McNamara. 2025. Ymcq: Reasoning-enhanced mcq generation. In *Artificial Intelligence in Education*, pages 308–315, Cham. Springer Nature Switzerland.

Francisco J García-Crespo, Rubén Fernández-Alonso, and José Muñiz. 2021. Academic resilience in european countries: The role of teachers, families, and student profiles. *Plos one*, 16(7):e0253409.

Alex Housen, Folkert Kuiken, and Ineke Vedder. 2012. Dimensions of l2 performance and proficiency: Complexity, accuracy and fluency in sla. language learning & language teaching. volume 32. *Language Learning & Language Teaching (MS)*.

Kevin Hwang, Sai Challagundla, Maryam Alomair, Lujie Karen Chen, and Fow-Sen Choa. 2023. Towards ai-assisted multiple choice question generation and quality evaluation at scale: Aligning with bloom's taxonomy. In *Workshop on Generative AI for Education*.

Nancy Ide and James Pustejovsky. 2017. *Handbook of Linguistic Annotation*, 1st edition edition. Springer Netherlands, Dordrecht.

Eunice E Jang. 2008. A review of:"cognitive diagnostic assessment for education: Theory and application" jacqueline p. leighton and mark j. gierl, editors, cambridge university press, may 2007, 384 pages, us 80.00hardcover,us 29.99 paperback, isbn 978-0-521-68421-7.

Alec I Kennedy and Rolf Strietholt. 2023. School closure policies and student reading achievement: Evidence across countries. *Educational Assessment, Evaluation and Accountability*, 35(4):475–501.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018).

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Cedric Lothritz, Jordi Cabot, and Laura Bernardy. 2025. Testing low-resource language support in llms using language proficiency exams: the case of luxembourgish. *Preprint*, arXiv:2504.01667.

Xiaofei Lu. 2025. Meaning and function dimensions of linguistic complexity in second language writing. *Research Methods in Applied Linguistics*, 4(1):100191.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Claudine Moulin and Damaris Nübling. 2006. *Perspektiven einer linguistischen Luxemburgistik*. Universitätsverlag Winter Heidelberg.

Ina VS Mullis and Michael O Martin. 2019. *PIRLS 2021 Assessment Frameworks.* ERIC.

Sana Nouzri, Meryem EL Fatimi, Titouan Guerin, Mahfoud Othmane, and Amro Najjar. 2025. Beyond chatbots: Enhancing luxembourgish language learning through multi-agent systems and large language model. In *PRIMA 2024: Principles and Practice of Multi-Agent Systems*, pages 385–401, Cham. Springer Nature Switzerland.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Timothy C. Papadopoulos, Valéria Csépe, Mikko Aro, Marketa Caravolas, Irene-Anna Diakidoy, and Thierry Olive. 2021. Methodological issues in literacy research across languages: Evidence from alphabetic orthographies. *Reading Research Quarterly*, 56(S1):S351–S370.

Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. Text generation models for Luxembourgish with limited data: A balanced multilingual strategy. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 93–104, Abu Dhabi, UAE. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Omer Shubi, Yoav Meiri, Cfir Avraham Hadar, and Yevgeni Berzak. 2024. Fine-grained prediction of reading comprehension from eye movements. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 3372–3391. Association for Computational Linguistics.

Shruti Singh, Nandan Sarkar, and Arman Cohan. 2024. SciDQA: A deep reading comprehension dataset over scientific papers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20908–20923, Miami, Florida, USA. Association for Computational Linguistics.

Service information et presse du gouvernement SIP. Introduction to luxembourgish.

Hakyung Sung, Mikyung Kim Wolf, Michael Suhan, and Kristopher Kyle. 2025. Lexical richness in young english learners' writing: A focus on opinion and listen-write task types. *Assessing Writing*, 66:100975.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. *Preprint*, arXiv:2408.01337.

Joseph S Yamazaki and Phil Hiver. 2024. Learners' behavioral engagement and performance on linguistically difficult l2 reading tasks: The effects of effort feedback, self-efficacy, and attributions. *Language Teaching Research*, page 13621688241304871.

Kamyar Zeinalipour, Neda Jamshidi, Fahimeh Akbari, Marco Maggini, Monica Bianchini, Marco Gori, and 1 others. 2025. Persianmcq-instruct: A comprehensive resource for generating multiple-choice questions in persian. In *Proceedings of the First Workshop*

on Language Models for Low-Resource Languages, pages 344–372. Association for Computational Linguistics.

Kamyar Zeinalipour, Yusuf Gökberk Keptiğ, Marco Maggini, and Marco Gori. 2024. Automating turkish educational quiz generation using large language models. In *International Conference on Intelligent Systems and Pattern Recognition*, pages 246–260. Springer.

Zhihan Zhang, Yixin Cao, and Lizi Liao. 2025. XFinBench: Benchmarking LLMs in complex financial problem solving and reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8715–8758, Vienna, Austria. Association for Computational Linguistics.

Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. 2024. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. *Preprint*, arXiv:2406.02472.