# ITEC at MLSP 2024: Transferring Predictions of Lexical Difficulty from Non-Native Readers

**Anaïs Tack**
KU Leuven
Faculty of Arts, Research Unit Linguistics
imec research group itec
anais.tack@kuleuven.be

## Abstract

This paper presents the results of our team's participation in the BEA 2024 shared task on the multilingual lexical simplification pipeline (MLSP; Shardlow et al., 2024a). During the task, organizers supplied data that combined two components of the simplification pipeline: lexical complexity prediction and lexical substitution. This dataset encompassed ten languages, including French. Given the absence of dedicated training data, teams were challenged with employing systems trained on pre-existing resources and evaluating their performance on unexplored test data.

Our team contributed to the task using previously developed models for predicting lexical difficulty in French (Tack, 2021). These models were built on deep learning architectures, adding to our participation in the CWI 2018 shared task (De Hertog and Tack, 2018). The training dataset comprised 262,054 binary decision annotations, capturing perceived lexical difficulty, collected from a sample of 56 non-native French readers. Two pre-trained neural logistic models were used: (1) a model for predicting difficulty for words within their sentence context, and (2) a model for predicting difficulty for isolated words.

The findings revealed that despite being trained for a distinct prediction task (as indicated by a negative $R^2$ fit), transferring the logistic predictions of lexical difficulty to continuous scores of lexical complexity exhibited a positive correlation. Specifically, the results indicated that isolated predictions exhibited a higher correlation ($r = .36$) compared to contextualized predictions ($r = .33$). Moreover, isolated predictions demonstrated a remarkably higher Spearman rank correlation ($\rho = .50$) than contextualized predictions ($\rho = .35$). These results align with earlier observations by Tack (2021), suggesting that the ground truth primarily captures more lexical access difficulties than word-to-context integration problems.

## 1 Introduction

The aim of predicting and simplifying lexical difficulty is to enhance text readability by focusing on vocabulary. Drawing from a simplified perspective on reading (Hoover and Gough, 1990), we can divide these difficulties into two main categories: decoding and comprehension. Decoding issues relate to difficulties in accessing words (also known as "lexical access"), where readers struggle to recognize and recall the form and meaning of words from memory. Conversely, comprehension difficulties involve struggles in integrating words into the broader textual context (sometimes termed "word-to-context integration"). Therefore, simplifying lexical difficulty entails employing various strategies to boost clarity and comprehension. This may involve substituting complex terms with simpler alternatives or providing contextual clues or definitions. Ultimately, the goal is to enhance accessibility while maintaining the integrity of the conveyed message.

Over the last decade, several tasks have been organized to advance the development of automated models, including the complex word identification shared task (Paetzold and Specia, 2016), the second complex word identification shared task (Yimam et al., 2018), the shared task on lexical complexity prediction (Shardlow et al., 2021), and the shared task on multilingual lexical simplification (Saggion et al., 2022). Lastly, Shardlow et al. (2024a) organized the shared task on multilingual lexical simplification pipeline (MLSP).[1]

This system description paper outlines our team's involvement in the MLSP shared task, focusing on our approach. Specifically, we leveraged predictions of lexical difficulty for French from previous research (Tack, 2021) in the initial phase of the lexical simplification pipeline, known

[1] https://sites.google.com/view/mlsp-sharedtask-2024/



*Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 635–639

as lexical complexity prediction. Our approach also entailed comparing predictions for individual words (approximating lexical access difficulties) with predictions for words within context (approximating word-to-context integration difficulties). Subsequent sections detail our methodology and findings.

## 2 Method

The shared task progressed through two distinct phases. In the development phase, which took place from February 15 to March 14, 2024, teams were tasked with developing systems using existing resources. Due to the absence of dedicated training data and the small size of only 30 trial items per language, our emphasis was on employing pre-trained models for making zero-shot predictions of lexical difficulty (see Section 2.1).

During the evaluation phase, from March 15 to March 26, 2024, teams were provided with test data for ten languages within the MultiLS framework (Shardlow et al., 2024b; North et al., 2024). During this phase, we used our pre-trained models to predict scores of lexical complexity for the French test set (see Section 2.2) and made two submissions. In the subsequent sections, we will provide a more detailed description of the pre-trained models and test data.

### 2.1 Pre-Trained Models for French

We employed two neural models for predicting lexical difficulty in French, previously developed by Tack (2021) in her Ph.D. thesis. These models represented an improved version of the deep learning architecture developed by De Hertog and Tack (2018) for the second shared task on complex word identification (Yimam et al., 2018) and the earlier models developed by Tack et al. (2016b).

The first model featured a bidirectional long short-term memory neural network architecture, depicted in Figure 1. This model, constructed using TensorFlow, incorporated two word representations as input: character embeddings (generated through a convolutional neural network) and pre-trained FastText word embeddings. Furthermore, the model integrated learner-specific encodings to tailor predictions accordingly. However, in the transfer approach, personalization was not possible, resulting in these encodings being set to zero for the shared task.

The second model comprised a feedforward neu-

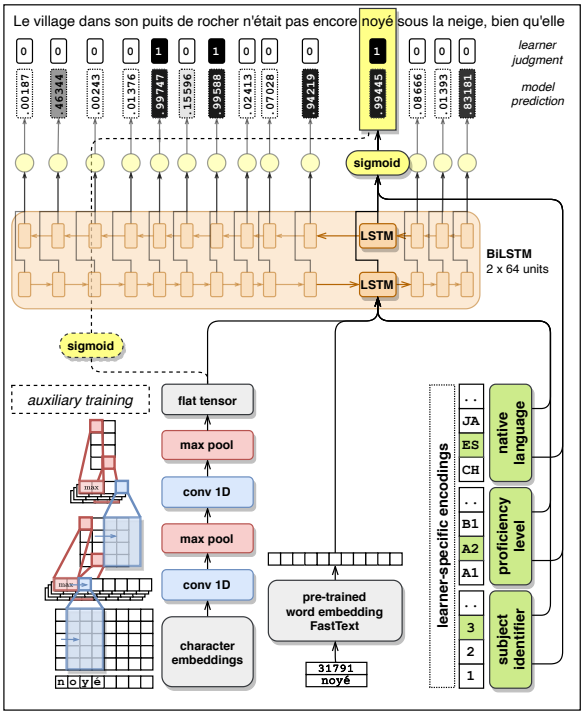


Figure 1: Bidirectional Long-Short Term Memory Neural Network Architecture in Tack (2021), Making Contextualized Predictions of Lexical Difficulty for French

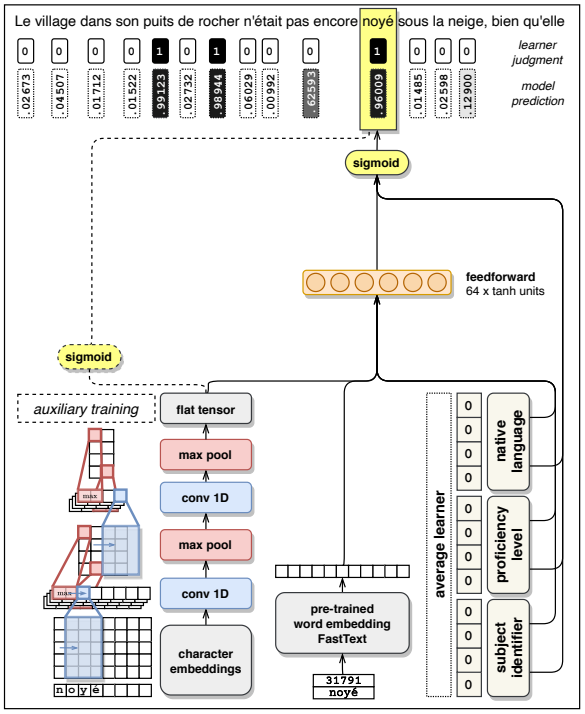


Figure 2: Feedforward Neural Network Architecture in Tack (2021), Making Isolated Predictions of Lexical Difficulty for French

| ID | Language | Sentence Context | Target Word |
|---|---|---|---|
| fr_549 | french | Bien sûr, on peut me rétorquer que je n'ai qu'à acquérir la nationalité française. | rétorquer |
| fr_550 | french | Bien sûr, on peut me rétorquer que je n'ai qu'à acquérir la nationalité française. | acquérir |
| fr_551 | french | Bien sûr, on peut me rétorquer que je n'ai qu'à acquérir la nationalité française. | nationalité |

Figure 3: Examples of Items in the French Test Data

ral network architecture, as depicted in Figure 2. Built using TensorFlow, this model utilized two word representations as input: character embeddings (generated through a convolutional neural network) and pre-trained FastText word embeddings. Additionally, learner-specific encodings were incorporated into the model to customize predictions. However, in the transfer approach where personalization wasn't possible, these encodings were also set to zero, as depicted in the figure.

It's worth mentioning that Tack (2021) conducted fine-tuning on contextualized BERT models. However, these models were not employed due to their underperformance compared to the previous two models, as indicated by the results presented in Tack (2021).

The two models presented in Figures 1 and 2 were trained using the dataset detailed in Chapter 5 of Tack's thesis, which expanded upon the initial data collected by Tack et al. (2016a). This training dataset comprised 262,054 binary decision annotations gathered from a sample of 56 non-native[2] French readers. These annotations captured *perceived* lexical difficulty, as participants were instructed to read texts and highlight words they did not understand. This method differed from measuring *actual* lexical difficulty. Since participants were prompted to highlight words, they could potentially overlook genuinely challenging words that they didn't recognize while reading the text.

### 2.2 Test Data for French

The French test data, as supplied by Shardlow et al. (2024a), contained 570 items. Each item included an identifier, the language, contextual word usage, and the target word requiring difficulty prediction, as depicted in Figure 3. Among the total 570 target words, the dataset comprised 560 unique word types and covered 191 distinct sentence contexts.

[2]Most readers were native Dutch speakers, with a minority being speakers of Chinese, Japanese, and Spanish.

For the lexical complexity prediction track, the French test data was annotated by 10 raters, all of whom were non-native French speakers. Their native languages included Arabic (2), Mandarin (2), German (1), Hindi (1), Italian (1), Japanese (1), Spanish (1), and Turkish (1).

## 3 Results

Figure 4 illustrates the model predictions for the French test dataset. As shown, both models generally predicted a high difficulty level (> 0.5) for most test items, with the isolated model (run 2) indicating a higher difficulty level compared to the contextualized model (run 1).

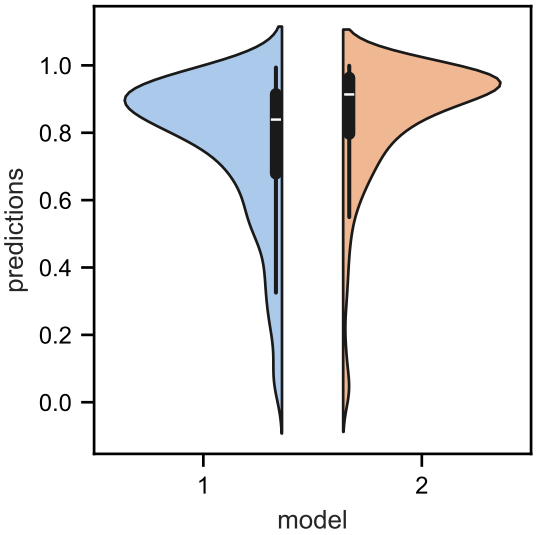


Figure 4: Predictions of Lexical Complexity for the French Test Data

Table 1 showcases the leaderboard results for the French test dataset. Notably, the $R^2$ metric suggests that both models exhibited a negative fit with the true complexity scores, as supported by the high (worse) scores for MAE and MSE. One likely explanation is that both models were trained

| # | Team | Run | $r$ | $\rho$ | MAE | MSE | $R^2$ |
|---|---|---|---|---|---|---|---|
| 1 | TMU-HIT | A | 0.6253 | 0.6302 | 0.1669 | 0.0452 | 0.2704 |
| 2 | Archaeology | 1 | 0.5335 | 0.5310 | 0.1898 | 0.0487 | 0.2136 |
| 3 | TMU-HIT | A | 0.5278 | 0.5343 | 0.1744 | 0.0471 | 0.2391 |
| 4 | RETUYT-INCO | A | 0.4868 | 0.4651 | 0.2063 | 0.0602 | 0.0279 |
| 5 | Archaeology | 2 | 0.4411 | 0.4188 | 0.1851 | 0.0504 | 0.1862 |
| 6 | Archaeology | A | 0.4411 | 0.4188 | 0.1851 | 0.0504 | 0.1862 |
| → 7 | ITEC | 2 | 0.3607 | 0.4972 | 0.5302 | 0.3373 | –4.4459 |
| → 8 | ITEC | 1 | 0.3253 | 0.3533 | 0.4545 | 0.2694 | –3.3488 |
| 9 | GMU | 1 | 0.3193 | 0.3207 | 0.2089 | 0.0589 | 0.0484 |
| 10 | GMU | A | 0.1557 | 0.1756 | 0.2136 | 0.0617 | 0.0039 |
| 11 | SCaLAR | A | 0.1035 | 0.0674 | 0.2093 | 0.0616 | 0.0061 |

Table 1: Leaderboard of Lexical Complexity Prediction for French Including the Predictions by the Two Models

for a notably distinct prediction task, namely logistic regression instead of linear regression. Another conceivable factor contributing to the negative fit is the variation in native languages among the non-native readers who annotated the data in Tack (2021) compared to those who annotated the French test dataset (see Section 2.2). Since annotators' native languages influence their perception of word difficulty, this variation is likely to impact the accuracy of the predictions.

However, the findings presented in Table 1 also demonstrate that, despite the weak fit, the models still exhibited a modest positive correlation with the true complexity scores. Specifically, the findings indicated that isolated predictions showed a slightly stronger correlation ($r = .36$) compared to contextualized predictions ($r = .33$).

These results align with earlier observations by Tack (2021), indicating that the ground truth predominantly reflects greater challenges in lexical access (i.e., difficulty recognizing the form and meaning of the word, regardless of its context) rather than issues in word-to-context integration (i.e., difficulty in interpreting the word within its context). Specifically, Tack (2021) noted that words identified as challenging by non-native readers exhibited more lexical access difficulties, as indicated by the higher predictive power of features associated with isolated word surprisal compared to contextualized word surprisal. This finding is unsurprising, given that the annotators had elementary to intermediate proficiency levels and, therefore, had a significantly smaller vocabulary size compared to native speakers. Consequently, it is reasonable to assume that the non-native annotators of the French test dataset also had a lower vocabulary size and were thus more susceptible to encountering words not yet ingrained in their mental lexicon, resulting in greater challenges in recognizing the form and meaning of words.

Furthermore, the results depicted in Table 1 reveal that isolated predictions demonstrated a notably higher (and fourth-best) Spearman rank correlation ($\rho = .50$) compared to contextualized predictions ($\rho = .35$). This suggests that although the logistic scores predicted by the model might not closely match the continuous complexity scores, they still preserve the same ranking of difficulty as the continuous complexity scores would. Therefore, even though transferring the difficulty scores may pose uncertainty, there is an interesting potential in transferring the ranking of lexical difficulty from this model to new data.

## 4 Conclusion

This study delved into predicting lexical complexity in French test data employing two models: an isolated model and a contextualized model. The findings underscore that while the transfer of difficulty scores remains uncertain, the ranking of lexical difficulty from this model can still be applied to new data. This emphasizes the potential usefulness of the models in comprehending lexical complexity in French texts, while also spotlighting the limitations in transferring the raw predicted scores. Moving forward, we also intend to explore the implications of transferring zero-shot predictions made with pre-trained French models to other languages.

## References


Dirk De Hertog and Anaïs Tack. 2018. Deep Learning Architecture for Complex Word Identification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, volume 13, pages 328–334, New Orleans, Louisiana. Association for Computational Linguistics.

Wesley A. Hoover and Philip B. Gough. 1990. The Simple View of Reading. *Reading and Writing*, 2(2):127–160.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2024. Multils: A multi-task lexical simplification framework. *arXiv preprint arXiv:2402.14972*.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 Shared Task on Multilingual Lexical Simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024a. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Marcos Zampieri, and Horacio Saggion. 2024b. An Extensible Massively Multilingual Lexical Simplification Pipeline Dataset using the MultiLS Framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Anaïs Tack. 2021. *Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers*. Ph.D. thesis, UCLouvain & KU Leuven, Louvain-la-Neuve, Belgium.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016a. Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, volume 10, pages 230–236, Portorož, Slovenia. European Language Resources Association.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016b. Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Articles longs)*, volume 23, pages 221–234, Paris, France. AFCP - ATALA.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, volume 13, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.