UNIVERSITY OF GOTHENBURG

DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

UNTANGLING LANGUAGE MODELS

Analysis of XLM-RoBERTa Predictions in the Framework of Grammatical Error Detection

Céline Leuzinger

Master’s Thesis:	15 credits
Programme:	Master’s Programme in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2023
Supervisor:	Elena Volodina, David Alfter
Examiner:	Staffan Larsson
Keywords:	GED, LLMs, XLM-RoBERTa, m-DeBERTa-V3

Table of Contents

Acknowledgements

First and foremost, I would like to express my warmest thanks to my supervisor, **Elena Volodina**, for her expertise, unwavering support, and great kindness throughout this journey. I am also extremely thankful to my co-supervisor, **David Alfter**, for the helpful and detailed guidance he provided. I have truly enjoyed collaborating with both of you, and be assured that the knowledge and insights I have gained during our work will continue to influence my academic journey.

I would also like to thank my friend and fellow linguist **Elisa**, for being a patient and insightful language consultant, and answering all of my questions on the Italian language. I also extend my gratitude to **Emil**, who served as an exceptional Swedish consultant and, beyond that, a loving and caring partner.

I also express my profound gratitude to all of **the professors, teaching assistants and staff members** from the Department of Philosophy, Linguistics and Theory of Science, for their understanding and flexibility throughout the year, and for all the invaluable knowledge that I have gained.

Finally, I would like to thank all of the wonderful **friends and classmates** that I met in Sweden during this eventful yet beautiful year, for all the support they provided and all the fun we had. Tack så mycket!

Abstract

The field of Grammatical Error Detection (GED) holds significant importance in NLP, as both L2 learners and native speakers make use of GED systems to enhance their writing abilities (Madi & Al-Khalifa, 2018: 142). In the last few years, the field of Grammatical Error Detection has been the object of numerous studies, which contributed to the remarkable improvement of GED systems. Large Language Models (LLMs), and more particularly XLM-RoBERTa, achieved state-of-the-art results for GED in several languages (Colla et al. 2023), in the framework of the MultiGED-2023 shared task (Volodina et al., 2019), presented at the 12th NLP4CALL workshop. To further improve GED systems, the present work carries out a linguistic analysis of XLM-RoBERTa's predictions for Italian and Swedish, in order to investigate the strenghts and weaknesses of LLMs in GED. Results show that syntax, lexicon and morphology are more problematic than punctuation and orthography. The data suggests that the model struggles to take into account world knowledge, sentence context and logic while making predictions. It also has difficulties signalling missing tokens, or tokens that should be deleted.

Additionally, the present work fine-tunes the m-DeBERTa-V3 language model for the purpose of GED: this model is pre-trained on Replaced Token Detection (He et al., 2023: 1), a task that is closely related to GED (Yuan et al., 2021: 8730), which hints that m-DeBERTa-V3 should perform well in error detection. Results show that m-DeBERTa-V3 performs poorly in GED, opposite of what was initially hypothesised. These poor results could be due to numerous factors, related to hyperparameter tuning, the choice of the

loss function or the lack of optimization processes. Future studies could keep on exploring the use of m-DeBERTa-V3 in GED and show whether the results can be improved.

1 Introduction

1.1 Grammatical Error Detection at the NLP4CALL workshop

Every year since 2012, researchers have gathered to present and share their findings in the area of Computer Assisted Language Learning, within the framework of the annual NLP4CALL workshop (Natural Language Processing for Computer-Assisted Language Learning). In 2023, a shared task on Multilingual Grammatical Error Detection (often abbreviated "MultiGED") was organised as part of the workshop (Volodina et al. 2023). Participants were encouraged to build and submit GED systems capable of identifying grammatical errors in a set of underrepresented languages in NLP: Czech, Italian, Swedish and German (Volodina et al. 2023: 1). English was also included in the workshop so as to allow comparison with previous GED systems and research (Volodina et al. 2023: 3). Six teams of researchers participated in the task and submitted systems for GED in one or more languages, using a wide range of techniques: LSTMs, Machine Translation (MT) techniques, as well as Large Language Models (LLMs) (see Volodina et al. 2023: 8). The winning team fine-tuned the XLM-RoBERTa language model, topped with a linear classifier, and outperformed the other teams for the majority of datasets provided, thereby presenting a new State of the Art (SOTA) for Czech, Italian, Swedish and German GED (Volodina et al. 2023: 12, Colla et al. 2023). On the other hand, no team from the shared task managed to improve the results of existing English GED systems, and the SOTA for English GED remains Yuan et al. (2021), who fine-tuned the ELECTRA (Clark et al. 2020) language model for grammatical error detection.

1.2 Aim of the present work

A natural follow-up question is how to further improve the field of GED. This can be achieved in at least two ways. First, one could propose and implement a new system, for example, using another kind of language model. Another possibility is to carry out an in-depth linguistic analysis of the predictions of current SOTA language models, so as to understand their strengths and weaknesses, and hence suggest targeted improvements. A possibility is to analyse the model's incorrect predictions, and see whether a particular area of the grammar (such as punctuation, orthography, lexicon, morphology, syntax) is problematic for the model. The present work shall explore both of these possibilities, with a focus on the latter one. The need for such linguistic analyses in the field of GED is indeed considerable; pinpointing the areas in which models performs poorly would allow further research to find targeted solutions to these issues. To the best of our knowledge, no such analysis has been performed in the context of the MultiGED-2023 shared task so far. Further, such investigation could deepen our general understanding of language models. Language models are still regarded as complex and opaque, and there has often been more efforts towards creating new LMs rather than understanding the ones that already exist (Bender et al., 2021: 619). Casting light on how such models understand language would allow future research to improve them further.

The present work will therefore analyse the output of the winning model, XLM-RoBERTa. The choice of this model stems from multiple reasons: first of all, if one has in mind the general improvement of GED, it is natural to focus on the model that is already SOTA for most languages. Second, XLM-RoBERTa is a development of BERT (Liu et al., 2019), a highly popular model which has been used in a very wide range of NLP tasks since its creation in 2018 (Devlin et al., 2018). Finally, further research keeps on improving BERT models: for example, He et al. (2021) recently developed another improvement of BERT, DeBERTa, which is known to outperform other BERT-models in every NLP task (He et al. 2021: 9). Gaining insights into the functioning of XLM-RoBERTa can prove advantageous for other BERT-based Language Models.

Another choice that the present thesis makes is to analyse the model's prediction for Swedish and Italian. The reason are once more numerous: first of all, annotating the predictions of the model is a time-consuming task, and the set of languages should hence be restricted for a work of this scope. Second, it has already been underlined that finding an annotation scheme that is relevant cross-linguistically is an extremely difficult task (Volodina et al. 2023: 3), and reducing the set of languages allows to preserve both the coherence and quality of the annotations. The choice of Italian and Swedish in particular is motivated by the following reasons: linguistic abilities (which ruled out the analysis of Czech), availability of native speaker consultants, and will to choose languages that are underrepresented in the field of GED. English has hence been left for further studies to analyse, as well as German, since these are more widely spoken than Swedish and Italian. Further, Swedish and Italian come from two different linguistic families, respectively Germanic and Romance, and hence offer a wide linguistic representation.

The second goal of this thesis is to make use of another language model for GED, to see whether the results of XLM-RoBERTa could be further improved. The language model m-DeBERTa-V3 (He et al. 2023) is of particular interest in the present case, since it shares structural similarities with two languages models that are already SOTA in GED: ELECTRA in English GED (Yuan et al., 2021), and XLM-RoBERTa in multilingual GED (Colla et al., 2023). We shall therefore fine-tune m-DeBERTa-V3 for Grammatical Error Detection using the MultiGED Swedish and Italian datasets, and compare the results with the ones of XLM-RoBERTa.

1.3 Structure

The present work consists of three main sections: section 2 presents XLM-RoBERTa as fine-tuned by the winning team of the MultiGED shared task, EliCoDe (Colla et al. 2023). Section 3 introduces our analysis of its predictions for Swedish and Italian. Section 4 presents m-DeBERTa-V3 and its fine-tuning for the purpose of GED.

Section 2.1 introduces the MultiGED-2023 shared task and data, while section 2.2 presents the different teams that participated in the task, along with the systems they used and the results obtained. Section 2.3 shortly summarises the use of Large Language Models in GED. Section 2.4 delves into the specificities of the winning model, XLM-RoBERTa: the story of its development, and the fine-tuning carried out by the EliCoDe team. Section 3 presents a linguistic analysis of XLM-RoBERTa's predictions. First, the annotation framework used to annotate the predictions is introduced in section 3.1, while section 3.2 presents the linguistic analysis for each grammatical domain: punctuation, orthography, lexicon, morphology and syntax. Section 3.3 summarises the findings and concludes the section. Section 4 introduces m-DeBERTa-V3, starting with a summary of its architecture and development in section 4.1. This is followed by a short description of the data used for the training in section 4.2. Section 4.3 details the fine-tuning steps, section 4.4 presents the results, and section 4.5 discusses and critically assess these results. Section 4.6 summarises the findings and concludes this section, while section 5 provides a general conclusion.

2 Introducing XLM-RoBERTa in GED

2.1 MultiGED-2023 at the NLP4CALL: shared task and data

The MultiGED-2023 shared task at the NLP4CALL focuses on Grammatical Error Detection. Grammatical Error Detection consists in the detection of tokens that require a correction, as described by Volodina et al. (2023: 1). This task is of great importance in NLP, as GED systems are commonly used by both native speakers and L2 language learners as a tool to improve their language skills (Madi & Al-Khalifa, 2018: 142).

Grammatical Error Detection is to be distinguished from Grammatical Error Correction (often abbreviated "GEC"): while the former detect an ungrammatical input, the latter generates a correction for an ungrammatical input. However, GEC systems are not always able to accurately detect all types of errors (Rei & Yannakoudakis, 2016: 1181). For example, lexical errors are not easily detectable by correction systems (Kochmar and Briscoe, 2014: 1740). It is therefore necessary to develop powerful error detection systems in parallel to error correction systems.

GED can be either binary or multi-class, and can be performed at the sentence-level or at the token-level. The multiGED shared task focuses on binary token classification: namely, the model has to tell whether a certain token is correct ("c") or incorrect ("i") within a sentence, and make a prediction for each token, using data that contains tokens manually annotated as "c" or "i". The incorrect predictions made by the model can be either False Negatives or False Positives. False Negatives are incorrect tokens that were marked as correct by the model. False Positives are correct tokens that were marked as incorrect by the model. The following sentence is taken from the English FCE dataset, used during the shared task, and illustrates the predictions of the model used by EliCoDe compared to the gold standards, along with annotations of False Negatives:

Sent:	I	am	sad	to	read	about	Richard	not	being	at	his	best	.
Gold:	c	c	c	c	c	c	c	c	c	i	i	i	c
Pred:	c	c	c	c	c	c	c	c	c	c	c	c	c
FNs:										FN	FN	FN	

Example 1: gold standard and predicted sentence for an English example, along with annotations of False Negative (FN)

In the example above, the model predicted that every token is correct "c", while the gold standard contains in fact three incorrect tokens "i". Hence, the model missed three erroneous token: they should therefore be labelled as "False Negative". On the other hand, the model correctly identified ten correct tokens, which are therefore not annotated, since they are neither False Negatives nor False Positives. Such sequence of predictions is made for every sentence within the datasets, and is then compared to the gold standard. The model's predictions may also contain False Positives. An example of this is illustrated below:

Sent:	The	college	is	just	stairs	away	from	the	bottom	end	.
Gold:	c	c	c	c	i	c	c	c	c	c	c
Pred:	c	i	c	c	c	c	c	c	c	c	c
FNs/FPs:		FP			FN						

Eaxample 2: gold standard and predicted sentence for an English example, along with annotations of False Positive (FP) and False Negative (FN)

In the example above, the sentence contains one False Positive and one False Negative. XLM-RoBERTa therefore made two incorrect predictions in this sentence.

The shared task covers five languages: Czech, Swedish, Italian, German, and for the purpose of comparison with previous studies, English. Train, development and test datasets were provided to the participants for each language (Volodina et al. 2023: 4). The datasets contain individual sentences, which are taken from corpora of annotated L2 learners essays (Volodina et al. 2023: 4). It is worth noting that the sentences come in random order; this means that the order of the sentences within the essay is not preserved. The data for each language was taken from the following corpora:

Language	Corpus
Czech	GECCC (Náplava et al., 2022)
Swedish	SweLL-gold (Volodina et al., 2019)
Italian	MERLIN (Boyd et al., 2014)
German	Falko-MERLIN (Boyd 2018)
English	FCE Corpus (Yannakoudakis et al., 2011) REALEC (Vinogradova and Lyashevskaya, 2022)

Table 1: corpora from which the datasets were taken at the MultiGED-2023 shared task. See Volodina et al. (2023) for a detailed description of each dataset. The datasets are available on Github: https://github.com/spraakbanken/multiged-2023.

The present thesis focuses on Swedish and Italian. The data statistics for these two languages is presented in the table below:

Language	Source Corpus	Nr. Sentences	Nr. Tokens	Nr. Errors	Error Rate
Swedish	SweLL-gold	8,553	145,507	27,274	0.187
Italian	MERLIN	7,949	99,698	14,893	0.149

Table 2: data statistics for Swedish and Italian, taken from Volodina et al. (2023: 4)

These statistics give us many insights about the datasets. Taking the Italian dataset as reference, the Swedish dataset has about 7.6% more sentences. However, it has 45.9% more tokens, thereby suggesting that the Swedish datasets contains longer sentences. Further, the Swedish dataset contains almost twice as many errors as the Italian dataset, i.e. 83.1% more errors in Swedish than in Italian, with an error rate that is also slightly higher than the Italian error rate. We therefore expect more False Negatives in Swedish than in Italian, a hypothesis that is confirmed in the linguistic analysis (see section 3).

2.2 MultiGED-2023 at the NLP4CALL: teams, systems and results

Six teams participated in the MultiGED shared task, and used a wide range of techniques to achieve GED. The table below, taken from Volodina et al. (2023: 8) summarises the techniques used by each team:

System	Description
EliCoDe (Colla et al., 2023)	XLM-RoBERTa language model pretrained on ≈100 languages with a stacked linear classifier on top, with a dropout layer in-between. Fine-tuned 5 different models for 5 languages on train (or train+dev) data.
DSL-MIM-HUS (Ngo et al., 2023)	XLM-RoBERTa language model from the HuggingFace repository pretrained on ≈100 languages, fine-tuned jointly on all MultiGED datasets. There is only one trained model for prediction of all the test datasets.
Brainstorm Thinkers	mBERT, for all six datasets
VLP-char (no eng-realec) (Ngo et al., 2023)	Character-based LSTM model with two recurrent layers, unidirectional supervised approach, separate model for each dataset, REALEC excluded. No external datasets.
NTNU-TRH (Bungum et al., 2023)	Multilingual system based on LSTMs, GRUs, and standard RNNs with multilingual Flair embeddings for a sequence-to-sequence labeling multitask learning.
Su-dali (only Swedish) (Kurfalı and Östling, 2023)	Distantly-supervised transformer-based machine translation (MT) system trained solely on an artificial dataset of 200 million sentences, only Swedish. No supervision, training, or fine-tuning on any labeled data.

Table 3: Summary of the systems used by each team, taken from Volodina et al. (2023: 8)

All teams opted for deep learning techniques. Half of the teams made use of language models. Two teams used LSTMs, while one team resorted to machine translation techniques.

The predictions made by each system were then evaluated using Precision, Recall, and Fβ. Precision calculates the number of correct positive predictions made by the model by dividing the True Positives by the total number of positive predictions made:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall divides the number of correct positive predictions by the total number of positives in the dataset:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Finally, Fβ considers both precision and recall to make a balance estimation of the model's performances. The formula for Fβ is shown below:

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

The value assigned to β determines whether more weight is assiged to precision or recall. A small value puts more weight on precision, whereas a high value puts more weight on recall. Since β=0.5 in the present case, the formula puts more weight on precision than recall.

It is worth noting that the number of True Negatives (in other words, correct tokens that are accurately predicted by the systems) is not taken into account in these metrics. The task indeed consists in identifying incorrect tokens, and the ability of the model to recognise correct tokens is hence not taken into account.

The tables below rank the teams for the Italian and Swedish data, as well as their scores, starting with the best performing team:

Team	**Precision**	**Recall**	**F0.5**
EliCoDe	81.80	66.34	78.16
DSL-MIM-HUS	74.85	44.92	66.05
Brainstorm Thinkers	73.81	39.94	63.11
su-dali	82.41	27.18	58.60
VLP-char	26.40	55.00	29.46
NTNU-TRH	80.12	5.09	20.31
Majority	89.90	45.37	75.15

Table 4: Ranking of the teams for Swedish, starting with the best performing team. The "Majority" row is calculated based on the majority token prediction of all systems taken together. This table was taken from Volodina et al. (2023: 9).

Team	Precision	Recall	F0.5
EliCoDe	86.67	67.96	82.15
DSL-MIM-HUS	75.72	38.67	63.55
Brainstorm Thinkers	70.65	36.46	59.49
NTNU-TRH	93.38	19.84	53.62
VLP-char	25.79	44.24	28.14
Majority	90.25	40.95	72.74

Table 5: Ranking of the teams for Italian, starting with the best performing team. Note that the su-dali team only worked with Swedish, and therefore does not appear in the Italian ranking. The "Majority" row is calculated based on the majority token prediction of all systems taken together. This table was taken from Volodina et al. (2023: 9).

EliCoDe (Colla et al., 2023) ranks best for both languages, and represents the new SOTA for Swedish and Italian (Volodina et al., 2023: 12). Interestingly, the teams that made use of language models are also the three best ranking teams, therefore showcasing the efficiency of LMs in GED.

2.3 The use of Language Models in GED

Historically, a wide range of systems and techniques have been used for GED and GEC. Madi and Al-Khalifa (2018: 143) distinguish three types of approaches: rule-based, syntax-based and machine learning and transformer-based approaches. Rule-based approaches make use of manually written rules to check for errors in text, but are extremely time-consuming and hence not efficient (Madi and Al-Khalifa, 2018: 143). Syntax-based approaches use parsers to identify mistakes; if the parser cannot produce a syntactic tree, the sentence is erroneous (Madi and Al-Khalifa, 2018: 143). However, this technique cannot locate the mistake within the sentence, and is also time-consuming to design (Madi and Al-Khalifa, 2018: 142). Recent approaches have therefore focused on machine learning for GED. The use of neural models, and in particular transformer-based language models, contributed to significantly improve both GED and GEC (Bryant et al. 2022: 18). In fact, Language Models have become the new State of the Art in both GED and GEC (Clark et al., 2020, Bryant et al., 2022: 18, Volodina et al., 2023). In the case of Swedish and Italian, XLM-RoBERTa is the best performing model to this day (Volodina et al. 2023, Colla et al. 2023).

2.4 XLM-RoBERTa: specifications and fine-tuning

2.4.1 From BERT to XLM-RoBERTa

XLM-RoBERTa is a transformer-based multilingual model developped in 2020 by Conneau et al. XLM-RoBERTa is built upon its monolingual counterpart, RoBERTa, developped in 2019 by Liu et al., itself an improvement of BERT, introduced by Google in

2018 (Devlin et al., 2018).

The introduction of transformers by Google in 2017 represented a major breakthrough in NLP research, as they were shown to outperform both recurrent and convolutional neural networks (Vasawani et al., 2017: 10). One of the key advantages of transformers is their self-attention mechanism: transformers are able to look for relevant context within a whole sequence, and can hence process long-range dependencies (Vasawani et al., 2017: 2). In 2018, Google introduced BERT (Bidirectional Encoder Representations from Transformers), a monolingual[1] transformer-based language model relying entirely on self-attention, which delivered state-of-the-art results in numerous NLP tasks (Devlin et al., 2018: 6). Two tasks were used for pre-training: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) (Devlin et al., 2018: 5). In the former task, a certain percentage of the input token is masked, and the model must predict the masked token; in the latter, the model is asked to tell whether two sentences follow another or not (Devlin et al. 2018: 5). Further studies have aimed at improving the already promising results given by BERT.

Liu et al (2019) significantly increased BERT's performance on a wide range of NLP tasks by implementing a certain number of improvements, thereby developing an analoguous monolingual model: RoBERTa (Robustly Optimized BERT Approach). A certain number of parameters were changed in RoBERTa: the number of layers increased from 12 to 24, another four attention heads were added, the batch size and vocabulary size increased to respectively 2k and 50k (Liu et al., 2019: 13). The greater number of layers is beneficial for the model to understand more complex patterns in the data. Attention heads, on the other hand, are used to capture dependencies between words; increasing the number of attention heads allows the model to better understand such dependencies. A larger batch size helps reduce training time while also improving the model generalisation capabilities, while the increase in vocabulary size allows RoBERTa to handle a much broader span of words. RoBERTa also benefits from a larger dataset, which includes the OPENWEBTEXT, CC-NEWS and STORIES datasets, as well as longer training sequences (Liu et al. 2019: 1). Changes were also made in pre-training. RoBERTa uses dynamic masking instead of the previous static masking: while BERT performed masking only once during the data processing, RoBERTa performs a new masking for every sequence fed into the model, thereby improving the results (Liu et al. 2019: 4). Next Sentence Prediction (NSP) was also dropped in RoBERTa, which resulted in equal or slightly improved overall performances (Liu et al. 2019: 5). A year later, Conneau et al. (2020) developped XLM-RoBERTa, a multilingual RoBERTa, trained on Common Crawl data for 100 languages, which performed equally well as the monolingual model, thereby proving that multilingual data does not necessarily hurt the model's performance (Conneau et al., 2020: 1). It is with this model that the winning team EliCoDe (Colla et al., 2023) yielded the best results overall at the MultiGED-2023 shared task. The following section explores the fine-tuning procedures conducted by the EliCoDe team to perform Grammatical Error Detection (GED) using XLM-RoBERTa.

[1] There also exists a multilingual version of BERT, called mBERT (see Pires et al., 2019).

2.4.2 Fine-tuning of XLM-RoBERTa

The EliCoDe team considered the GED task a token classification task, where each token should be assigned a label (Colla et al., 2023: 28). The team converted the labelled data into BIO format, which stands for "Beginning-Inside-Outside": "B" is used to mark the beginning of the error span, "I" is used to mark tokens inside the error span, and "O" is used to mark tokens outside of the error span (Colla et al., 2023: 28). In the absence of information on the error span, the team decided to only use "B" to mark incorrect tokens, and "O" to mark correct tokens (Colla et al., 2023: 28). EliCoDe's model is based on ClinicalTransformerNER (Yang et al., 2020), which they adapted so as to fit XLM-RoBERTa (Colla et al. 2023: 28). ClinicalTransformerNER is a transformer package that was developed in 2020 for the purpose of clinical concept extraction: in other words, extracting patient information (such as family or drug history) from clinical texts (Yang et al., 2020: 1935). Such package could possibly be used to extract erroneous tokens from the data - however, Colla et al. (2023) do not provide further details on the use and adaptation of ClinicalTransformerNER to the present task.

EliCoDe used XLM-RoBERTa, topped up with a dropout layer, and a linear classifier (Colla et al., 2023: 28). The dropout layer is a regularization technique that helps prevent overfitting: a certain percentage of the neurons (10% in the present case) are randomly dropped out during training. Hence, the model cannot rely exclusively on a specific subset of neurons, which prevents overfitting. The linear classifier provides a probability for each class, for each token in the sequence, and outputs the most probable label ("O" or "B", ie. correct or incorrect) for a given token.

It is worth noting that EliCoDe trained a different model[2] for each language (Colla et al., 2023: 28). The models were trained using the training and development set for 10 epochs (Colla et al., 2023: 28). EliCoDe also trained the models based solely on the training data (Colla et al., 2023: 28), but the best performance was achieved by the models trained on both the training and development sets - it is hence these models that we shall focus on in the following sections.

2.4.3 Model performance: False Positives and False Negatives

The next section of the present work performs a linguistic analysis of the model's output, and more particularly of the model incorrect predictions. But before proceeding, it is of importance to discuss what should be considered incorrect predictions.

As mentioned previously, the scores achieved by EliCoDe were the highest in the competition for both Italian and Swedish (Volodina et al., 2023: 9). However, the model also made a certain amount of incorrect predictions. Such incorrect predictions can be either False Negatives or False Positives: False Negatives are incorrect tokens that are

[2] The present work should therefore always speak of "models" when refering to the models presented by EliCoDe for Italian and Swedish. However, since each of the models present a similar architecture, the present work shall keep on using the singular "model" to make generalisation that can be applied to both models.

predicted as correct, while False Positives are correct tokens that are predicted as incorrect. The amount of False Negatives and False Positives can be illustrated in a confusion matrix:

		True labels	
		Correct	Incorrect
Predicted Labels	Correct	11,533	905
	Incorrect	397	1784

Table 6: confusion matrix for the Swedish data

		True labels	
		Correct	Incorrect
Predicted Labels	Correct	8,570	478
	Incorrect	156	1014

Table 7: confusion matrix for the Italian data

The present thesis focuses on False Negatives, and leaves aside False Positives. First, False Positives are less numerous, and their influence on overall results is hence minimal. Second, False Positives are difficult to analyse: one can hardly pinpoint why a correct token is considered incorrect by the model, unless one is willing to take the risk of making speculations. While the analysis of False Positives can certainly be undertaken despite this risk, the present thesis leaves it to future studies, and focuses on the analysis of False Negatives.

In a limited number of cases, the model was in fact correct about a False Negative: the token was labeled as "correct" by the model and manually annotated as "incorrect", yet there is no visible error. Let us illustrate this with an example taken from the Italian test data:

Sent:	Car-i	Mario	e	Francesco	,	[...]
Gloss:	dear-PL	Mario	and	Francesco	,	[...]
Gold:	i	i	i	i	i	
Pred:	c	c	c	c	c	
FNs/FPs:	*FN	*FN	*FN	*FN	*FN	

Example 3: "Dear Mario and Francesco". Gold standard and predicted sentence for an Italian example, along with annotations of False Negatives (FNs). The star in front of the annotation signifies that the prediction is incorrect.

If one follows the gold standard, then the entire sequence should be labeled as False Negative, since the model predicted that these tokens were correct, yet they are manually annotated as incorrect. However, taken as they are in the dataset (i.e. without any external context), these tokens are correct: the first word "dear"/"cari" is correctly spelled and has the correct plural marking, the proper names also seem to be correctly spelled, and there is no error in the conjunction "and"/"e". A possibility is that the human annotator marked the sentence as incorrect given a certain context. For example, it might be unnatural to write this sentence in an essay, especially if the task does not concern the writing of a letter. This underlines that corrections are relevant within a given context, which sometimes spawn beyond the sentence boundaries and is hence impossible to capture for a model that is trained on individual sentences. Let us also introduce an example of this discrepancy between a context-free sentence and the original sentence from the Swedish data. In the sentence below, the pronoun "Vi"/"we" was annotated as incorrect manually, certainly because the wider context of the essay required another pronoun to be coherent. However, outside of the scope of the essay, this is a grammatical use of "Vi":

Sent:	Vi	och	mina	familj	tycker	om	Rosaborg	.
Gloss:	we	and	my-PL	family	like	about	Rosaborg	.
Gold:	i	c	i	c	c	c	i	c
Pred:	c	c	i	c	c	c	i	c
FNs:	*FN							

Example 4: "We and my family". Gold standard and predicted sentence for a Swedish example, along with annotations of False Negative (FN). The star in front of the annotation signifies that the annotation is incorrect.

This underlines the importance of the context in annotating GED data: certain errors can only be considered as such within a wider context. Since the order of sentences is randomized in our datasets, the wider context is not preserved, and certain annotations have to be adapted in consequence.

The cases described above will be marked as "MC/FN", which stands for "Model Correct/False Negative". These cases are not taken into account in our linguistic analysis: since we argue that the token is not erroneous, no annotation can be assigned to it.

Inversely, the model might be correct about a False Positive: the annotator considered the token correct, while the model predicted it as incorrect, and we argue that the model is right. Let us illustrate this with an Italian example:

Sent:	Città	X	,	il	8	Gennaio	[...]
Gold:	c	c	c	i	c	c	
Pred:	c	c	c	i	c	i	
FPs:						*FP	
Gloss:	city	X	,	the	8	January	[...]

Example 5: "City X, the 8th of January". The formulation "City X" is used as an anonymisation process in the dataset.

In the example above, the human annotator considered that "Gennaio"/"January" was correct. However, there is an error in this token: in Italian, months do not take a capital letter. The correct token should hence be "gennaio". The model correctly predicted this token as incorrect. Such token is therefore labeled "MC/FP", which stands for "Model Correct/False Positive" in the present analysis.

Contrary to cases where the model is correct about a False Negative, MC/FP tokens are interesting for the present study, since there is a rationale to consider them erroneous. Further, they represent an error that was identified by the model, but not by the human annotator. Annotating and analysing them could therefore bring further insights into our understanding of LMs.

Additionally, certain "c"/"i" labels were assigned to tokens that are neither words nor punctuation marks and had to be removed from the present analysis. These are mostly cases of formatting tokens; for example, a line jump was wrongfully encoded and added to the dataset, and labeled "c" manually and "i" by the model. Such tokens were removed from the data.

Finally, there were cases in which both the human annotator and the model were correct, but suggested two different correction patterns. Let us take a segment of sentence taken from the Swedish data, which could be translated by "[...] for we who are adults [...]". Observe the discrepancy of the gold standard and predictions:

Sent:	för	vi	som	är	vuxna
Gloss:	for	we	who	are	adults
Gold:	i	c	c	c	c
Preds:	c	i	c	c	c
FPs/FNs:	FN?	FP?	-	-	-

Example 6: "[...] for we who are adults [...]"

In this example, there are two valid possibilities of corrections. The first one implies suppressing the word "för"/"for", hence changing the sentence to "[...] we who are adults [...]". The second possibility is to change the token "vi" into "oss"/"us", hence changing the sentence to "[...] for us who are adults [...]". In this case, we shall argue that the model's predictions should be considered correct, despite the difference with the truth labels. The sentence above therefore contains one "MC/FN" and one "MC/FP". This underlines another important aspect of data annotation for GED: there might be more than one way of correcting sentences, which might skew the results of GED systems.

The table below summarises the amount of FPs, FNs, MC/FPs and MC/FNs in the data after it was reviewed, for Swedish and Italian:

	Swedish	Italian
FPs	354	133
FNs	760	361
MC/FPs	42	21
MC/FNs	142	67
Removed	4	52

Table 8: Count of FPs, FNs, MC/FPs and MC/FNs in the datasets. The "Removed" row contains the number of tokens that were removed from the datasets.

The annotated and curated data will be shared with the organisers of the MultiGED-2023 shared task at the end of the revision period.

The present thesis exclusively concentrates on instances where the model's predictions diverge from those made by human annotators: the False Negatives (FNs), and cases where the model is correct about a False Positive (MC/FPs). It goes without saying that analysing all of the model's predictions (including the correct ones) would have been of great interest, since it would have allowed for better and more precise comparisons. For example, in the upcoming section, we will discuss the challenges XLM-RoBERTa faces in detecting missing tokens, as a significant portion of False Negatives are associated with such omissions. However, it's possible that in numerous instances, the model accurately predicted missing tokens errors. Unfortunately, we cannot validate this hypothesis: annotating correct predictions would involve analysing tens of thousands of tokens, a task beyond the scope of this thesis. However, it is still possible to formulate tentative hypotheses on the model's strengths and weaknesses by looking at the FNs and MC/FPs. Further studies could use the hypotheses formulated in the present thesis to carry out targeted analyses of correct predictions, and hence provide additional insights on the strengths and weaknesses of LMs. It is hence important to keep in mind that the conclusions that will be drawn when it comes to the model's difficulties with an area of the grammar are tentative, and should be reinforced or disputed by future analysis of the True Positives.

3 Analysing XLM-RoBERTa predictions

3.1 Annotation framework

The present work proposes to classify each FN and MC/FP into one of five grammatical categories: Punctuation, Orthography, Lexicon, Morphology, Syntax (POLMS), an annotation system already used by Volodina et al. (2019: 90) to annotated the SweLL corpus, from which the Swedish data used in MultiGED-2023 was taken. The Italian data was already annotated following the MERLIN taxonomy (see Boyd et al. 2014), which differs from POLMS. The Italian data was hence annotated for POLMS from scratch, while we re-used the POLMS annotations of the SweLL taxonomy to annotate the Swedish data. Minor changes were made to the POLMS classification used in SweLL, for example when the rationale behind a given annotation for Swedish could not be applied to the Italian data. Appendix B details the cross-referencing of the SweLL and the present taxonomy.

In addition to POLMS, the present thesis uses an other annotation system: the ADR error classification, initiated by Bryant et al. (2017), where "A" stands for Addition, "D" for Deletion, and "R" for Replacement. This annotation system describes the operation necessary to correct the erroneous token: suppress it, add another token, or replace it by another token. Once more, the Swedish data was already annotated for ADR. Such annotations were preserved in the present taxonomy. The Italian data, on the other hand, was annotated for ADR from scratch.

Let us illustrate the POLMS and ADR annotations with a custom English sentence:

Sent:	And	he	should	leave	but	he	do	not	want	.
Gold:	i	c	c	c	c	c	i	c	c	i
POLMS:	S						M			S
ADR:	D						R			A

Example 7: sentence, gold standard and POLMS-ADR annotation for a custom English example.

In the example above, there is an extra word: "And", which should be Deleted. The token is hence marked as "i", which stands for "incorrect". Additionally, it is annotated as "S" which stands for "Syntax" and "D" for "Deletion". The token "do" is also erroneous: it should be replaced by "does", it is hence a morphology "M" Replacement "R" mistake. Finally, there is a missing token: "to", as in "he does not want to", which should be added after "want" and is hence signalled on the following token ".". It is a syntactic error "S" which should be corrected by adding "A" a token.

In addition to POLMS and ADR, the present work also annotated each error type according to sub-categories, thereby giving further depth to the analysis. Both datasets were already annotated: the Swedish data was already divided into subcategories by Volodina et al. (2019), and the Italian data by Boyd et al. (2014). In order to keep a cross-linguistic coherence between the annotations, we re-annotated the data with our own annotation framework, utilising the already annotated data as a support for the annotation. See appendix B for a cross-reference between the current taxonomy and the SweLL taxonomy (Volodina et al., 2019).

The ADR annotation scheme was incorporated in the subcategories: each subcategory systematically represents either an Addition, Deletion or Replacement mistake. For example, the "missing comma" subcategory is always an "A" error type, since it requires the addition of a comma. Let us take the example presented above, and add the annotations for the subcategories:

Sent:	And	he	should	leave	but	he	do	not	want	.
Gold:	i	c	c	c	c	c	i	c	c	i
POLMS:	S						M			S
Subc.:	extra_word	-	-	-	-	-	agreement	-	-	missing_word
ADR:	D	-	-	-	-	-	R	-	-	A

Example 8: sentence, gold standard, POLMS annotations, subcategories of each error type and ADR annotations for a custom English example.

The Table below summarises the annotation framework used in the present research (for a detailed explanation of the annotation framework, along with illustrative examples, see Appendix A):

Punctuation	**Orthography**	**Lexicon**	**Morphology**	**Syntax**
Missing comma (A)	Capitalization (R)	Expression (R)	Wrong choice of definite/indefinite (R)	Formulation (R)
Missing punctuation (A)	Spelling (R)	Adjective (R)	Tense choice (R)	Missing word (A)
Wrong punctuation (R)	Word concatenation (R)	Adverb (R)	Plural morphology (R)	Extra word (D)
Conjunction-punctuation (R)		Conjunction (R)	Singular morphology (R)	
Punctuation-conjunction (R)		Noun (R)	Case (R)	Word order (R)
Extra comma (D)		Preposition (R)	Gender morphology (R)	
		Pronoun (R)	Mistake in the definite morpheme (R)	
		Verb (R)	Definite adjectival morphology (R)	
		Determiner (R)	Verbal morphology (R)	
		Auxiliary (R)		

Table 9: Summary of the present taxonomy. Each POLMS category is divided into subcategories, each of which corresponds to a type of ADR operation.

3.2 Results and discussion

After the annotation process, the total number of FNs for each POLMS category was computed, as well as the total number of FNs in each subcategory.

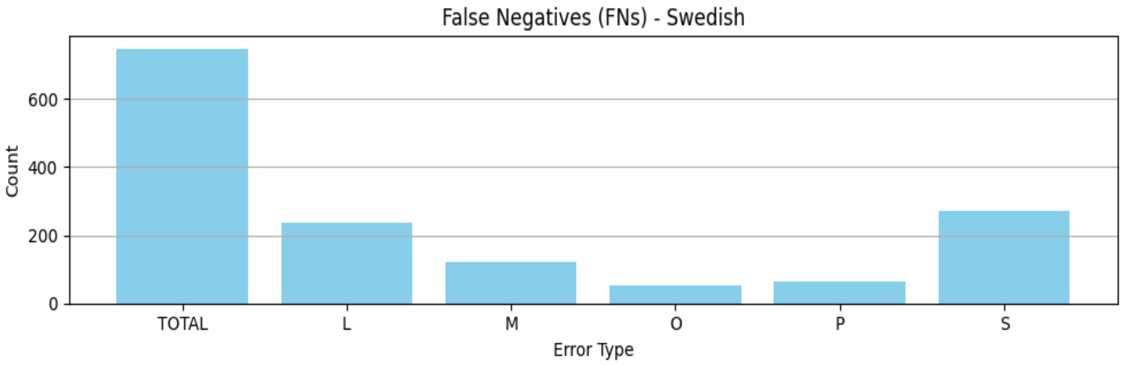


Figure 1: False Negatives classified into POLMS categories in Swedish.
Count: 'TOTAL': 747, 'L': 236, 'M': 120, 'O': 53, 'P': 65, 'S': 273
Converted into percentages: 'L': 31%, 'M': 16%, 'O': 6%, 'P': 8%, 'S': 36%

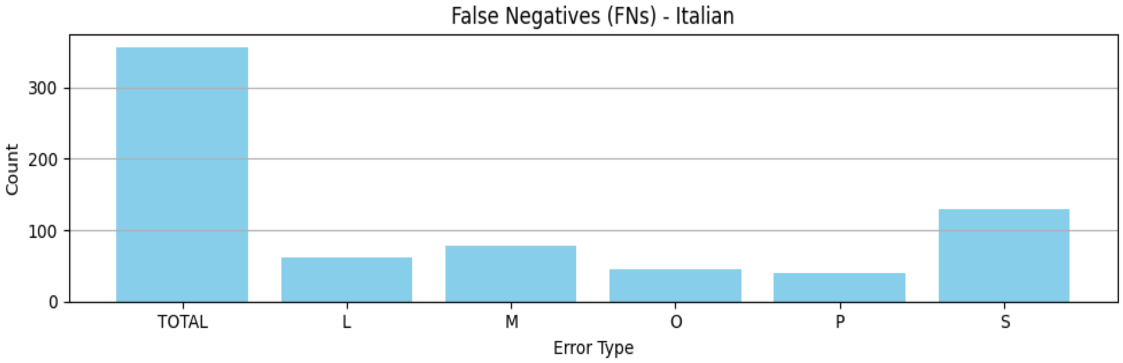


Figure 2: False Negatives classified into POLMS categories in Italian.
Count: 'TOTAL': 356, 'L': 62, 'M': 78, 'O': 46, 'P': 40, 'S': 130
Converted into percentages: 'L': 17%, 'M': 21%, 'O': 12%, 'P': 11%, 'S': 36%

A first striking observation is that syntax represents around one third of the FNs for both languages, while orthography and punctuation are much less represented. Lexical mistakes are also numerous in Swedish (31% of all FNs), whereas morphology is slightly more problematic in Italian (21% of all FNs).

The prevalence of syntactic errors is not unexpected. It is commonly known that certain syntactic processes, such as long-range dependencies, pose difficulties for language models, even though the invention of transformers largely improved this issue (Vasawani et al., 2017: 6). A closer look at syntactic subcategories could cast light on which syntactic processes are more represented among FNs.

The fact that orthography and punctuation only represent a minority of FNs can be explained by several factors. XLM-RoBERTa is based on a large vocabulary of sub-words units (Liu et al., 2019: 6), which should allow the model to correctly spot a majority of orthographic mistakes. Several factors can explain the small amount of punctuation mistakes: first of all, punctuation is the least-represented error type among the whole Swedish dataset (including training data), with 2'888 Punctuation mistakes out of a total of 27'489 errors (as annotated by Moner and Volodina, 2023) - less punctuation errors in the data naturally means less potential FNs for the model. Further, one could also expect limited punctuation complexity in learner's essays, thereby facilitating the task for the model.

Lexical FNs are numerous in Swedish. Lexical errors are often related to semantics: given a certain context, a word is incorrectly used, even though it fits the sentence syntactically. For a human being, these errors are often easily spotted. However, such mistakes require world knowledge to be correctly identified - and the amount of world knowledge that a model can have is limited by its size (Guu et al., 2020: 1), which might explain the high amount of Lexicon-related FNs in Swedish. On the other hand, morphology caused more problems than lexicon in Italian. This could be due to the fact that Italian has a rich morphology: for example, it has verbal agreement for each person, in each tense, as well as gender, singular and plural morphology, among others. This could explain why Italian has a relatively high amount of morphological FNs. The following section will hence carry an in-depth analysis of each subcategory of error for each POLMS category to test the hypotheses formulated above.

In addition to the FNs, the present work also analyses the cases where the model is correct about a False Positive. In other words, a certain token is labelled as "correct" manually, but rightfully predicted as "incorrect" by the model. These are hence the errors that the model identified, but not the human annotator. It is the case of 42 tokens in Swedish and 21 in Italian. The POLMS distribution of these is shown on the graph below:

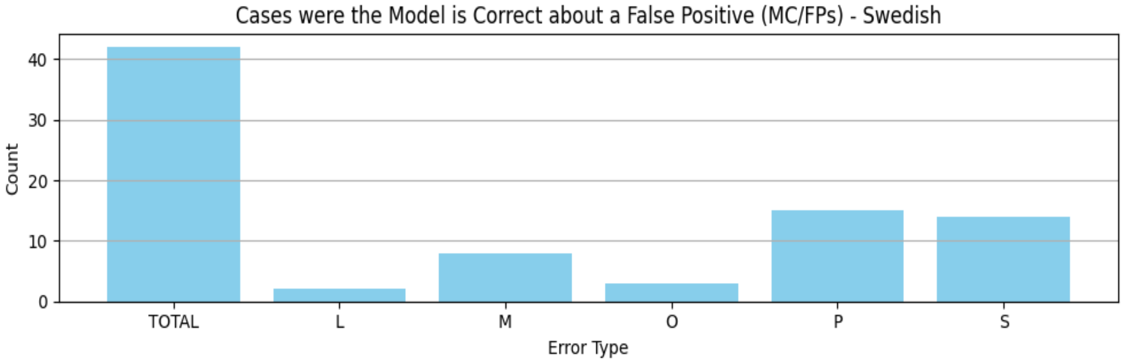


Figure 3: MC/FPs classified into POLMS categories in Swedish.
Count: 'TOTAL': 42, 'L': 2, 'M': 8, 'O': 3, 'P': 15, 'S': 14
Converted into percentages: 'L': 4%, 'M': 19%, 'O': 7%, 'P': 35%, 'S': 33%

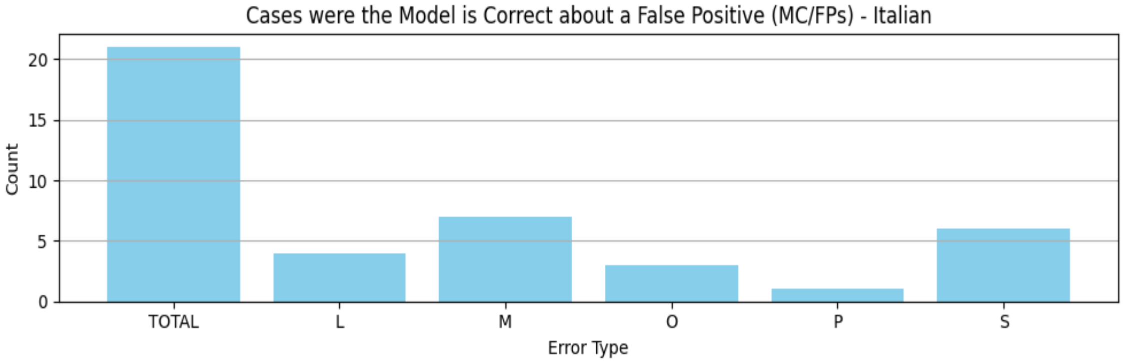


Figure 4: MC/FPs classified into POLMS categories in Italian.
Count: 'TOTAL': 21, 'L': 4, 'M': 7, 'O': 3, 'P': 1, 'S': 6
Converted into percentages: 'L': 19%, 'M': 33%, 'O': 14%, 'P': 4%, 'S': 28%

The model seems to perform well in punctuation in Swedish, morphology in Italian, and interestingly, syntax cross-linguistically. At this stage, at least two hypotheses can be formulated: the syntax category being the broadest category, the model might simply have more chances of being correct or incorrect about a syntactic mistake. Alternatively, the model might be performant in a certain type of syntactic error, and underperform for another type of syntactic mistake. The analysis of each subcategory of error, which we will delve into in the following subsections, should cast light on these hypotheses.

3.2.1 Punctuation

Punctuation only represents about 8% of the total number of False negatives in Swedish, and 11% in Italian. On the other hand, greater disparities exist between Swedish and Italian when it comes to errors that the model identified, but not the human corrector: 35% in Swedish, but only 4% in Italian (which represents only one case).

The distribution of punctuation FNs across subcategories for Swedish and Italian is illustrated in the graph below:

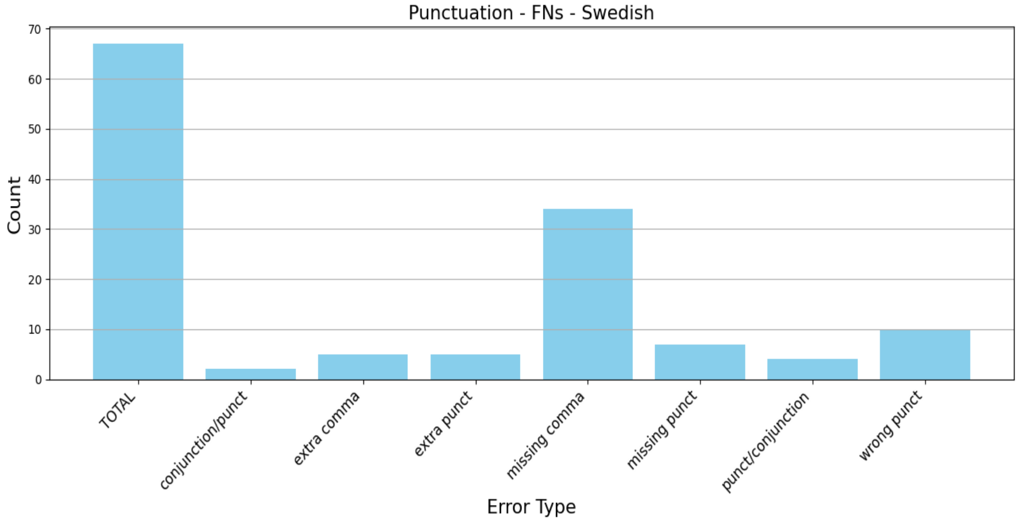


Figure 5: punctuation FNs classified into subcategories in Swedish.
Count: ’TOTAL’: 67, ’conjunction/punctuation’ (R): 2, ’extra comma’ (D): 5, ’extra punctuation’ (D): 5, ’missing comma’ (A): 34, ’missing punctuation’ (A): 7, ’punctuation/conjunction’ (R): 4, ’wrong punctuation’ (R): 10

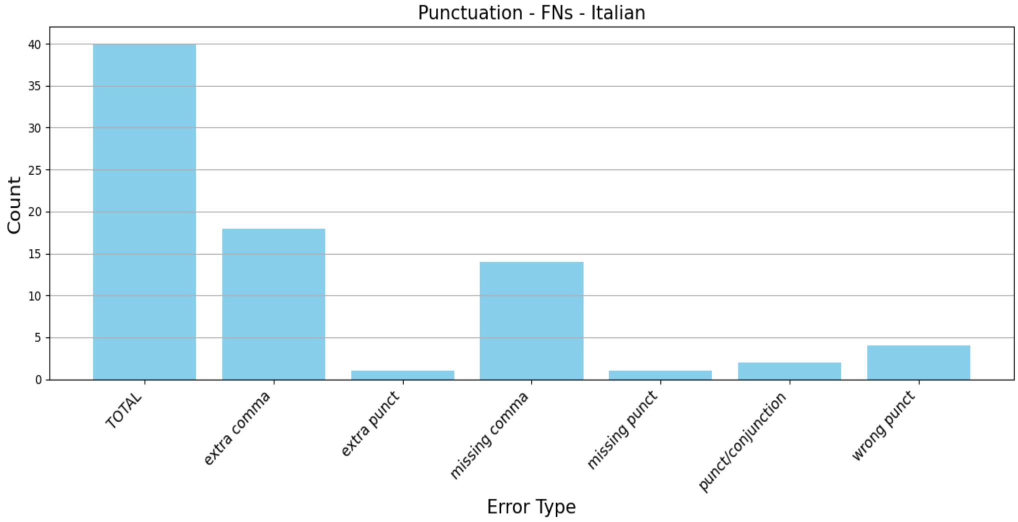


Figure 6: punctuation FNs classified into subcategories in Italian.
Count: ’TOTAL’: 40, ’extra comma’ (D): 18, ’extra punctuation’ (D): 1, ’missing comma’ (A): 14, ’missing punctuation’ (A): 1, ’punctuation/conjunction’ (R): 2, ’wrong punctuation’ (R): 4. The subcategory "conjunction/punctuation" does not have any occurrences in the Italian data, and is hence not represented on this graph.

The first observation is that the graphs do not mirror each other. In Swedish, the majority of punctuation FNs concern missing punctuation marks: 41 out of 67 cases. On the other hand, extra punctuation marks were most represented in Italian, with 19 out of 40 cases. However, in both cases, the most problematic type of punctuation mark are commas: 34 cases of missing commas in Swedish, and 18 cases of extra commas in Italian. In both languages, cases where a conjunction should be replaced by a comma (and inversely) are relatively rare: only two cases in Italian, and six in Swedish. Finally, the model did not identify an erroneous punctuation mark that should be replaced by another in 14 cases cross-linguistically.

Cross-linguistically, Addition and Deletion mistakes are the most represented among punctuation FNs. Addition errors are particularly hard to identify, since there must be a guideline as to how to signal them. In the MultiGED shared task, a missing token is signalled by labelling the following token as incorrect. Such an annotation guideline could pose difficulties for the model, since it forces it to label a token as incorrect, even if it technically does not contain an error. Making sure that this annotation guideline is implemented when fine-tuning the model could help reduce such cases of FNs. Deletion errors, on the other hand, are more straightforward to signal, since the token is present within the sentence. Yet, the fact that XLM-RoBERTa missed a certain number of those errors hints that the model struggles to identify whether a certain punctuation mark is needed or not. The analyses of the other POLMS category will reveal whether such difficulties are also encountered with missing and extra words.

Certain linguistic factors could also explain why Addition and Deletion errors are numerous among FNs. In Swedish, the subcategory "Missing comma" represents more than 50% of punctuation FNs. In 15 cases out of 34, the missing comma should have been used to separate a long sentence (>20 words) into shorter phrases. This suggests that the model has more chances of missing an error in long sentences. This could stem from the properties of the training data: the average sentence length of the training data is 17 tokens, which could explain why the model struggles with sentences that are longer than 20 tokens - some of which even contain more than 80 words. Further, in 7 cases out of 34, a comma should have been used before a relative clause which introduces non essential information. Such cases where the information is "non-essential" might be hard to decipher for a language model. For example, a sentence such as "The exams, which are corrected by professors, are in June" is as grammatically correct as "The exams which are corrected by professors are in June", and only knowledge of the world hints that the former sentence is the correct one. Further, punctuation has a relatively subjective nature: one can often debate whether a comma should or should not be added in a sentence, and factors such as writing style also play a role in the placement of punctuation. In Swedish in particular, commas are not rigidly used (p.c. Elena Volodina), and the data concerning commas should therefore be taken with a grain of salt.

Given the small number of punctuation FNs, we can hypothesise that the model performs well in this particular area of the grammar. This, however, could also be due to the proportion of punctuation errors in the test set: if there are few punctuation errors in total, we will also find few FNs related to punctuation.

Punctuation represents a large chunk of the mistakes that were identified by the model, but not by the human annotator for the Swedish data: 15 out of 42 cases, hence around 35% of the total number of False FPs - a number which is far lower for Italian, with only one case. The fact that these errors are not marked by the human annotator could stem from different causes. For example, the formation of the datastes required certain sentences to be re-formatted, which might have created punctuation inconsistencies. But the model's ability to find these errors remains remarkable. The graph below illustrates the distribution of such cases across subcategories in Swedish:

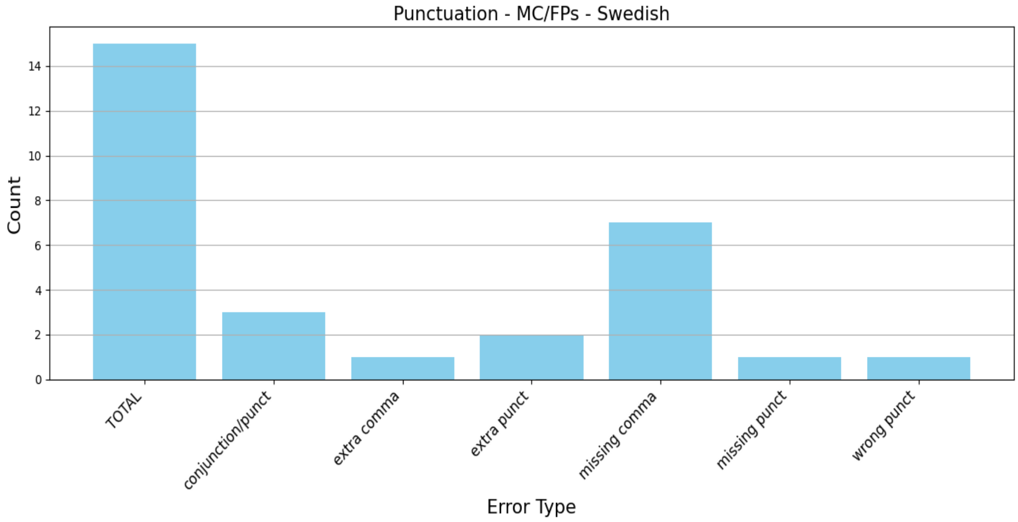


Figure 7: Punctuation MC/FPs classified into subcategories in Swedish.
Count: 'TOTAL': 15, 'conjunction/punctuation' (R): 3, 'extra comma' (D): 1, 'extra punctuation' (D): 2, 'missing comma' (A): 7, 'missing punctuation' (A): 1, 'wrong punctuation' (R): 1.

Once again, missing commas represent a vast majority of the errors: 7 out of a total of 15. Among these 7 MC/FPs, three are cases where the missing comma should have introduced a relative clause with non-essential information. Interestingly, this particular subtype of error was also common among FNs. This could suggest that missing commas is a common error within the test dataset, and the model therefore has more chances of both missing or identifying this particular error type. This, however, can only be verified by analysing the True Positives. For now, we shall therefore make the tentative conclusion that the model performs well in punctuation, given the fact that it is one of the least represented POLMS category among FNs.

3.2.2 Orthography

Orthographic errors do not seem particularly problematic for the language model: they only represent 6% of the False Negatives in Swedish and 12% in Italian. It is the lowest percentage among POLMS categories in Swedish and the second lowest in Italian, after punctuation. This could be due to the fact that language models make use of large vocabularies, and should hence be able to identify the most straightforward types of orthographic errors, such as spelling mistakes. The graph below shows the breakdown of orthographic mistakes in Swedish and Italian across subcategories:

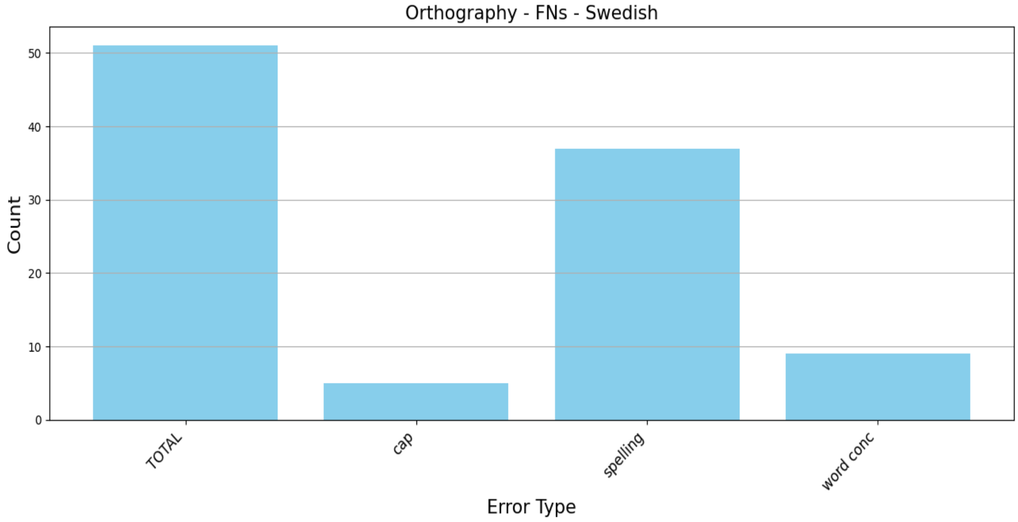


Figure 8: orthography FNs classified into subcategories in Swedish.
Count: ’TOTAL’: 51, ’capitalisation’ (R): 5, ’spelling’ (R): 37, ’word concatenation’ (R): 9.

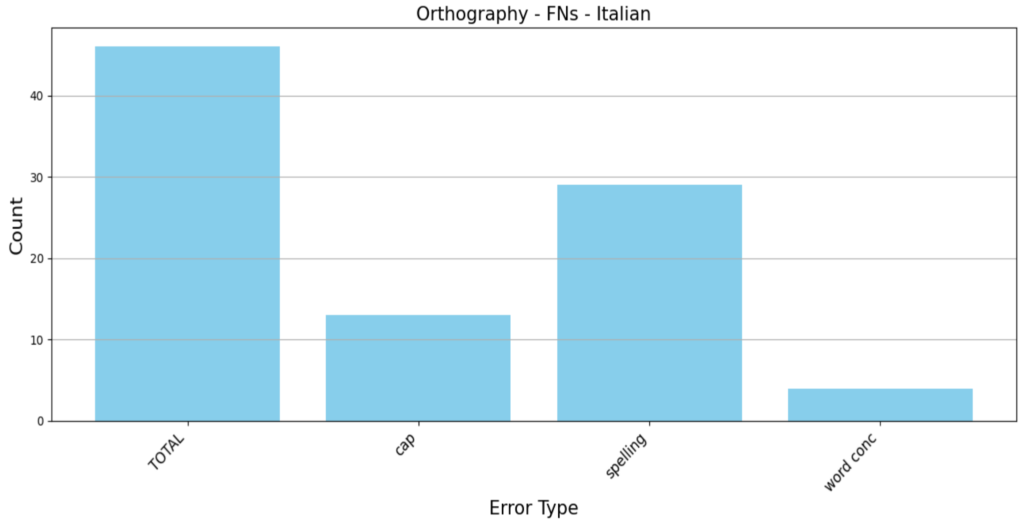


Figure 9: orthography FNs classified into subcategories in Italian.
Count: 'TOTAL': 46, 'capitalisation' (R): 13, 'spelling' (R): 29, 'word concatenation' (R): 4.

In both cases, spelling mistakes are by far the most represented subcategory among orthographic FNs. There are also a few cases of capitalisation FNs, i.e. tokens that are (or are not) supposed to start with a capital letter: 5 in Swedish and 13 in Italian. Additionally, respectively 9 and 4 FNs are cases of word concatenation: a certain token is supposed to be written in one word, but is written in two, or inversely.

Most cases of spelling FNs concern a missing character, or an inversion of characters. For example, the model counted "bätre" as correct instead of "bättre" ("better") and "rimboso" instead of "rimborso" ("reimbursement"). However, in Swedish, in 13 cases out of 37, the misspelled word did mean something - although not the meaning intended by the writer. For example, "rösta" ("vote") is misspelled into "rosta" ("grill"). A possible interpretation of these False Negatives could be that the model wrongly assumes that they are correct, since they do exist in the Swedish language and fit in the sentence grammatically, while making little sense semantically. However, we only count one similar case in Italian, which has yet a similarly high amount of spelling FNs. There is at least one more hypothesis as to why the model did not correctly identify some spelling mistakes: learners' essays often contain a high number of spelling mistakes, and it is hence not illogical that the model is more likely to miss some of them.

The model only identified a total of six errors that were not identified by the human annotator cross-linguistically: four concern capitalisation, two spelling mistakes. Cases of MC/FPs in orthography are therefore sparse. Future research could analyse the number of orthographic True Positives, and compare the results to the number of False Negatives. This would allow us to see the proportions of orthographic mistake the model actually identified and compare it to the number of errors that were missed. For now, the data suggests that the model performs well in orthography.

3.2.3 Lexicon

Interestingly, there exist a large difference between the ratios of Lexicon FNs in Swedish and Italian. The proportion of lexical mistakes is of 31% in Swedish, and only around 17% in Italian. In numbers, this represent 236 cases in Swedish and only 62 in Italian. This important difference could be explained by several factors; for example, the respective properties of each language's lexicon, or disparity in the training and test data. Further analyses of the lexical subcategories could cast light on this issue. There are only few cases where the model found an error that the human annotator did not identify: two cases in Swedish (one adjective, one noun), and four in Italian (two prepositions, one verb, one noun). Among all cases of errors that were identified by the model and not the human annotator, 4% and 19% are lexical errors in Swedish and Italian respectively.

Lexicon errors are all Replacement errors, and always describe an incorrect choice of word: the learner used a certain word, but should have used another one given the context. The Lexicon errors are divided into subcategories corresponding to the part of speech (POS) the token belongs to, such as adjective, adverb, conjunction, noun, preposition, pronoun, verb or determiner. If the erroneous token should be replaced by a token belonging to another part of speech, the error is categorised under the corrected token's part of speech. For example, the adjective "rapid" should be replaced by the adverb "rapidly" in the sentence "I went rapid". The error is hence adverbial. Additionally, a subcategory "expression" has been added for cases where a group of words is wrongly used, for example in an idiomatic expression. The figure below illustrates the distribution of these subcategories:

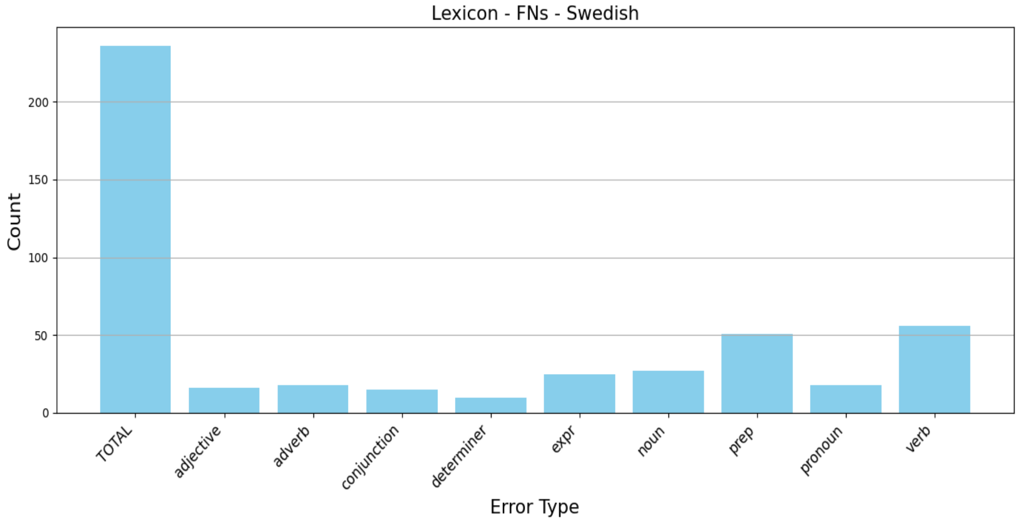


Figure 10: lexicon FNs classified into subcategories in Swedish.
Count: 'TOTAL': 236, 'adjective' (R): 16, 'adverb' (R): 18, 'conjunction' (R): 15, 'determiner' (R): 10, 'expr' (R): 25, 'noun' (R): 27, 'prep' (R): 51, 'pronoun' (R): 18, 'verb' (R): 56.

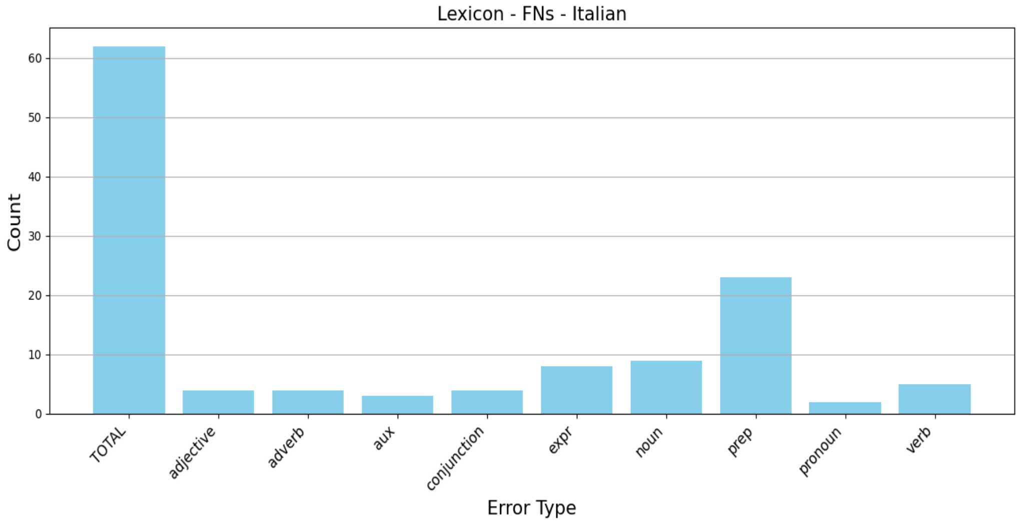


Figure 11: lexicon FNs classified into subcategories in Italian.
Count: 'TOTAL': 62, 'adjective' (R): 4, 'adverb' (R): 4, 'aux' (R): 3, 'conjunction' (R): 4, 'expr' (R): 8, 'noun' (R): 9, 'prep' (R): 23, 'pronoun' (R): 2, 'verb' (R): 5.
The graph contains one more subcategory compared to the Swedish data: "auxiliary". This error type only exists in Italian.

Cross-linguistically, preposition is the most represented POS, with 51 cases in Swedish and 23 in Italian, followed by nouns, verbs and expressions. On the other hand, adjectives, adverbs, conjunctions, determiners, pronouns and auxiliaries are much less problematic for the model in both languages.

When it comes to nouns, an interesting example taken from the Swedish dataset is the word "andning", which describes the biological process of breathing, compared to the word "andetag", which simply means "breath". In the equivalent of the expression "take a deep breath", the latter should be used. However, "andning" was used instead by one of the learners in the sentence "Ahhh jag orkar inte mer (och tar ett djupt andning)", which can be translated by "Ahh, I cannot take it anymore (and takes a deep breathing)":

Sent:	(	och	tar	ett	djupt	andning	)	.
Gloss:	(	and	takes	a	deep	breathing	)	.
Gold:	i	i	c	c	c	i	i	c
Preds:	c	c	c	i	i	c	c	c
POLMS:	P	P	-	-	-	L	P	-
Subc.:	extra_punct	missing_punct	-	-	-	noun	extra_punct	-
ADR:	D	A	-	-	-	R	D	-

Example 9: "[...] (and takes a deep breath)." Note that certain glosses are simplified in this example and others to ease the reading.

Interestingly, the model counted the tokens "ett" and "djupt" as incorrect, and "andning" as correct. The rationale behind this is logical: if one considers that "andning" is the correct noun to use in this context, then there is a gender mistake in both the determiner and the adjective, and the correct expression should be "en djup andning". However, the human annotator marked both the determiner and adjective as correct, and the noun as incorrect, since the context does not require the use of the word "andning", but rather the more familiar "andetag". This example hints that the model might lack the interpretative power to distinguish between two words that are semantically related, but used in slightly different situations.

The above example concerns a noun, but the majority of cases in the lexicon category concern prepositions or verbs: in Italian, prepositions represent a particularly high number of lexical FNs, while in Swedish, both verbs and prepositions are problematic, with respectively 56 and 51 cases out of 236 in total. Rastall (1994:1) showed the difficulties of English prepositions, which is mainly due to their anomalous nature, and illustrates his point with a striking example: "One may be arrested **for** a crime, accused **of** it and charged **with** it" (Rastall, 1994: 1).

Inspired by Rastall's (1994:1) quotation, one can create a similar sentence in Swedish:

(1) a. *Jag tänker* ***på*** *något, jag pratar* ***om*** *något, men jag reflekterar*
I think on something, I speak about something, but I reflect
över *något.*
over something.

The same example can be translated into Italian and make the exact same point:

(2) a. *Pens-o* ***a*** *qualcosa, parl-o* ***di*** *qualcosa, ma riflett-o* ***su***
Think-1SG about something, speak-1SG of something, but reflect-1SG on
qualcosa.
something

This reflects the somewhat arbitrary nature of prepositional use in both Swedish and Italian, and the difficulties this might cause to both learners and LMs.

In Swedish, the model also missed a nigh number of lexical errors related to verbs, which are listed in the table below, by order of frequency:

Verb	**Translation**	**Occurrences**
vara	be	9
ha	have	6
bli	become	5
tycka	think, have an opinion	3
tänka	think	2
få	get	2
ta	take	1
ska	shall	1
måste	must	1
göra	do	1
finnas	exist	1
befinna	be located	1
borde	should	1
gå	go	1
komma	come	1
prata	speak	1
konservera	preserve	1
anstränga	make an effort	1
läsa	read	1
veta	know	1
vandra	walk	1
träffa	meet	1
bilda	form	1
stå	stand	1
påpeka	impose	1
sluta	stop	1
invänta	wait	1
se	see	1
lägga	lay	1
passa	pass	1
återanmäla	re-enroll	1
anse	consider	1

Table 10: Verbs among lexical FNs. The occurrences also include the inflected forms of the verb, when relevant.

The majority of verbs in this list are modal verbs, or verbs which are widely used and/or polysemous. The Swedish verb "bli" is a good example of this: "bli" can be translated as "become", "get", "would be", and is also used in passive sentences. These uses are illustrated in the examples below:

(3) *Han ska bli ingenjör efter universitet-et.*
He shall become engineer after university-DET.

Translation: "He will become an engineer after university."

(4) *Det blir svårare att hitta ett hus.*
It becomes harder to find a house.

Translation: "It gets harder to find a house."

(5) *Det blir bättre att gå dit först.*
It becomes better to go there first.

Translation: "It would be better to go there first."

(6) *Hans företag blev grundat år 2018.*
His company became founded year 2018.

Translation: "His company was founded in 2018."

Many verbs from table 10 are widely used in Swedish, as well as in the test data. Below is a list of the ten most commonly used verbs from the test data the majority of which (7 out of 10) also stands among verbs that are commonly missed by the model as lexical FNs:

Verb	**Translation**	**Occurrences in the test set**
vara	be	438
ha	have	192
kunna	can	164
ska	shall	77
vill	want	59
tycka	think, have an opinion	58
finnas	exist	53
komma	come	51
bo	live, reside	47
måste	must	30

Table 11: ten most frequent verbs in the test data. The occurrences also include the inflected forms of the verb, when relevant.

The extensive usage of these verbs within the data, encompassing a wide array of meanings and contexts, might pose challenges for the model in discerning between correct and incorrect instances.

In Italian, lexical FNs related to verbs are much less frequent: only five cases. However, they also concern verbs that are widely used, such as "fare"/"do", "andare"/"go", "prendere"/"take" or "avere"/"have". One could therefore hypothesise that XLM-RoBERTa tends to have difficulties with verbs that cover a wide range of meanings and are hence overused in the data - a hypothesis that could be reinforced or disputed with further analysis of the True Positives in future works.

3.2.4 Morphology

Morphology represents 16% of the total number of False Negatives in Swedish; it is about twice as few than Syntax (36%) and Lexicon (31%). Morphology errors are more represented in Italian, with 21% of FNs. The figures below illustrate the distribution of FNs across subcategories for Swedish and Italian:

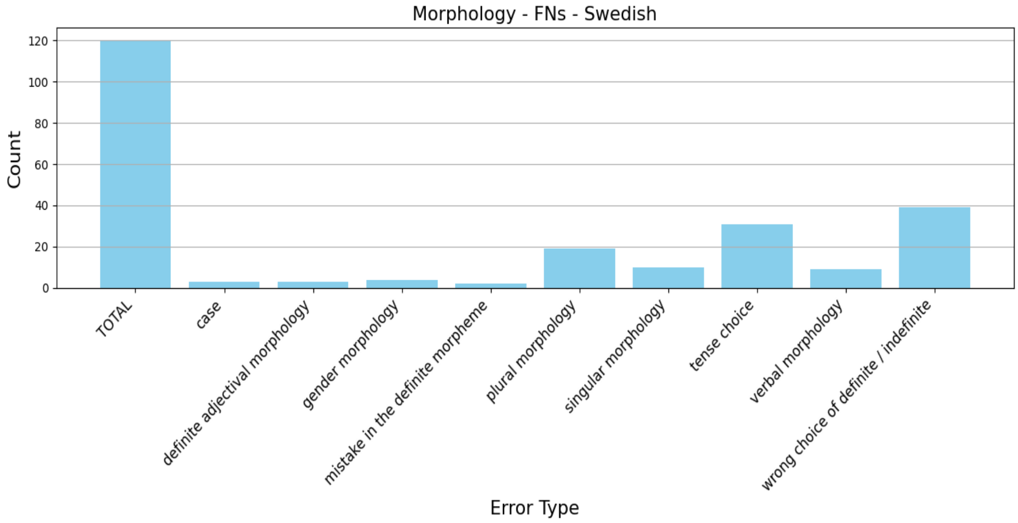


Figure 12: morphology FNs classified into subcategories in Swedish.
Count: 'TOTAL': 120, 'case' (R): 3, 'definite adjectival morphology' (R): 3, 'gender morphology' (R): 4, 'mistake in the definite morpheme' (R): 2, 'plural morphology' (R): 19, 'singular morphology' (R): 10, 'tense choice' (R): 31, 'verbal morphology' (R): 9, 'wrong choice of definite / indefinite' (R): 39

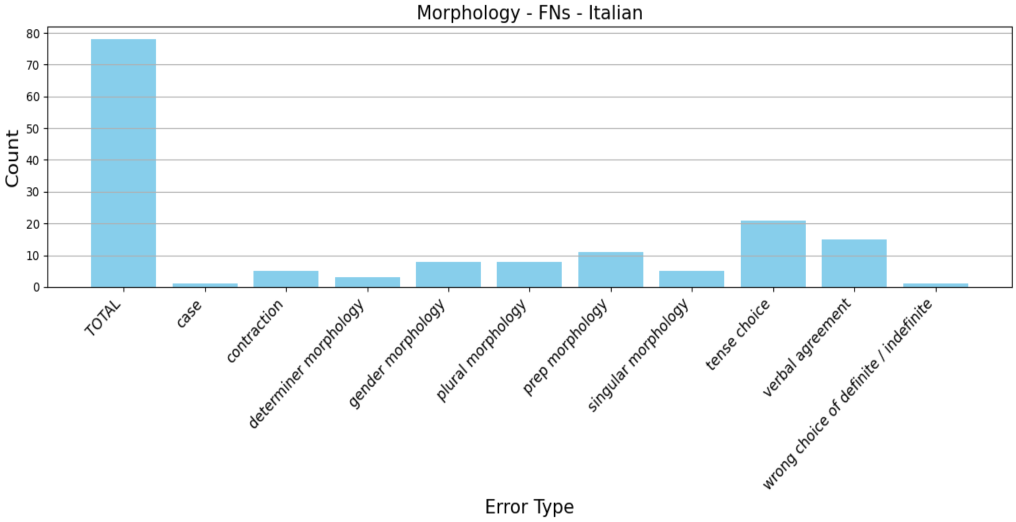


Figure 13: morphology FNs classified into subcategories in Italian.
Count: 'TOTAL': 78, 'case' (R): 1, 'contraction' (R): 5, 'determiner morphology' (R): 3, 'gender morphology' (R): 8, 'plural morphology' (R): 8, 'prepositional morphology' (R): 11, 'singular morphology' (R): 5, 'tense choice' (R): 21, 'verbal agreement' (R): 15, 'wrong choice of definite / indefinite' (R): 1

A first observation is that certain subcategories are completely absent in one language, mostly due to their distinct morphological properties. For example, Swedish has no verbal agreement: the verb does not inflect for person, hence the category "verbal agreement" does not appear in Swedish. Similarly, "mistake in the definite morpheme" can only appear in Swedish, since Italian does not have a definite morpheme: definiteness is indeed expressed with a separate article. Such linguistic differences make the comparison of morphological FNs difficult. However, there are some morphological aspects that are common to both Swedish and Italian among False Negatives. It is the case of tense choice, with 31 cases in Swedish, and 21 in Italian. Proportionally, this represents respectively around 26% and 27% of morphological FNs. Additionally, singular and plural morphology are also problematic on occasion: 29 cases out of 120 in Swedish, 13 cases out of 78 in Italian. On the other hand, certain categories are represented in one language only. It is the case of definiteness and indefiniteness in Swedish: in 39 cases, the learner use the definite form of a noun or determiner, when the indefinite form should have been used (or inversely). In Italian, verbal agreement as well as prepositional morphology seemed to pose difficulties for the model, with respectively 15 and 11 cases out of a total of 78. Each of these subcategories will be analysed, starting with the subcategories that are problematic across languages.

A first subtype of error that was present cross-linguistic concerns tense choice. Tense selection depends on the context of a given sentence: its temporality, as well as the other tenses used within the same sentence. The model did not identify a number of tense coordination errors, as illustrated in the example below, where the writer mixes present and past:

Sent:	Hon	tycker	att	Hans	är	hennes	äkta	kärlek	men	så	var
Gloss:	she	thinks	that	Hans	is	her	true	love	but	so	was
Gold:	c	i	c	c	c	c	c	c	c	c	i
Preds:	c	c	c	c	c	c	c	c	c	c	c
POLMS:	-	L	-	-	-	-	-	-	-	-	M
Subc.:	-	verb	-	-	-	-	-	-	-	-	tense_choice
ADR:	-	R	-	-	-	-	-	-	-	-	R

Sent:	det	inte	,	där	förstår	hon	att	hon	hade	fel	.
Gloss:	it	not	,	there	understands	she	that	she	had	wrong	.
Gold:	c	c	c	c	c	c	c	c	c	c	c
Preds:	c	c	i	i	c	c	c	c	c	c	c
POLMS:	-	-	-	-	-	-	-	-	-	-	-
Subc.:	-	-	-	-	-	-	-	-	-	-	-
ADR:	-	-	-	-	-	-	-	-	-	-	-

Example 10: "She thinks that Hans is her true love, but it was not the case, and there she understands that she was wrong."

In the example above, the learner starts the sentence in the present before switching to the past tense with the verb "var"/"was". The model did not mark any verb as incorrect, which hints that it did not necessarily keep track of all tenses used within the sentence.

Sometimes, a certain tense is grammatically correct, but logically incorrect. In the sentence below, the learner speaks about a memory, and says that it "was" a good memory from their childhood. However, even if the memory occurred in the past, the fact that it is a good memory occurs in the present, hence the present tense should be used:

Sent:	Ibland	tänker	jag	på	honom	och	tycker	att
Gloss:	Sometimes	think	I	about	him	and	think	that
Gold:	c	c	c	c	c	c	c	c
Preds:	c	c	c	c	c	c	c	
POLMS:	-	-	-	-	-	-	-	-
Subc.:	-	-	-	-	-	-	-	-
ADR:	-	-	-	-	-	-	-	-

Sent:	det	var	en	god	minnen	[...]
Gloss:	it	was	a	good	memories	[...]
Gold:	c	i	i	i	i	
Preds:	c	c	i	i	i	
POLMS:	-	M	-	-	-	
Subc.:	-	tense_choice	-	-	-	
ADR:	-	R	-	-	-	

Example 11: "Sometimes, I think about him, and I think that it was a good memories."

In other cases, the choice of the correct tense is hinted by a word or expression within the sequence. However, the model sometimes has difficulties detecting such hint, as illustrated below:

Sent:	Efter	det	hade	vi	aldrig	kontaktat	varandra	.
Gloss:	After	that	had	we	never	contacted	each.other	.
Gold:	c	c	i	c	c	c	c	c
Preds:	c	c	c	c	c	c	c	c
POLMS:	-	-	M	-	-	-	-	-
Subc.:	-	-	tense_choice	-	-	-	-	-
ADR:	-	-	R	-	-	-	-	-

Example 12: "After that, we had never contacted each other."

In the example above, the expression "efter det"/"after this" indicates that the following tense should be the simple past, as in "after that, we never contacted each other". However, the learner uses the past perfect, which is erroneous in this case. Other cases of FNs similar to this include time phrases such as "until today", "now" or "soon", which require specific tense, yet are followed by erroneous tenses. Similarly, certain conjunctions require a specific tense. In Italian, an illustrative example of this is the word "se"/"if": in the vast majority of cases, "se" is followed by the subjunctive, and using the conditional is erroneous. Yet, the model missed the error in the following example:

Sent:	Mi	piacerebbe	lavorare	nell	Vostro	campo	,
Gloss:	me	like	work	in	your	camp	,
Gold:	c	c	c	i	c	c	c
Preds:	c	c	c	i	c	c	c
POLMS:	-	-	-	-	-	-	-
Subc.:	-	-	-	-	-	-	-
ADR:	-	-	-	-	-	-	

Sent:	se	mie	competenze	sarebbero	abbastanze	.
Gloss:	if	my	competencies	be:COND	enough	.
Gold:	c	i	c	i	i	c
Preds:	c	i	c	c	i	c
POLMS:	-	-	-	M	-	-
Subc.:	-	-	-	tense_choice	-	-
ADR:	-	-	-	R	-	-

Example 13: "I would like to work in your camp, if my competencies would be enough."

In the example above, the verb "to be" is conjugated in the conditional "sarebbero", which is ungrammatical in Italian - yet the error is not detected by the model. The Italian data contains four similar cases with the word "se" among the FNs.

Other cases of tense-related FNs leave more room for interpretation. In the sentence below, the learner mixes two tenses: present and past. In fact, a majority of verbs are in the past, and only one is in the present. The human annotator chose to count all the past tenses as incorrect, certainly due to the fact that the wider context of the essay requires the present tense. On the other hand, the model only counts the present as incorrect:

Sent:	Allan	var	ledsen	och	vill	inte	bor	i	äldrebående
Gloss:	Allan	was	sad	and	wants	not	lives	in	retirement.home
Gold:	c	i	c	c	c	c	i	i	i
Preds:	c	c	c	c	**i**	c	i	c	i
POLMS:	-	M		-	-	-	-	L	-
Subc.:	-	tense_choice	-	-	-	-	-	prep	-
ADR:	-	R	-	-	-	-	-	R	-

Sent:	och	leva	mer	,	så	han	bestämde	sig	själv	att	lämna
Gloss:	and	live	more	,	so	he	decided	him	self	to	leave
Gold:	i	c	c	c	c	c	i	c	i	i	c
Preds:	c	c	c	c	c	c	c	c	i	i	c
POLMS:	L	-	-	-	-	-	M	-	-	-	-
Subc.:	conj	-	-	-	-	-	tense_choice	-	-	-	-
ADR:	R	-	-	-	-	-	R	-	-	-	-

Sent:	äldrebånde	och	försvann	.
Gloss:	retirement.home	and	disappeared	.
Gold:	i	c	i	c
Preds:	i	c	c	c
POLMS:	-	-	M	-
Subc.:	-	-	tense_choice	-
ADR:	-	-	R	-

Example 14: "Allan was sad and does not want to live in the retirement home, and live anymore, so he decided to leave the retirement home and disappeared."

In this case, a wider context would have been necessary to determine whether present or past should be used: since this sentence is taken from an essay, it is the temporality of the essay as a whole that should govern the use of tense within this sentence. However, in the absence of context, the language model can hardly decide what tense should be used. In the Swedish data, three sentences are problematic in that regard: they contain FNs related to tense choice, even though it is impossible for the model to identify the correct tense. This emphasises the need of larger contexts when training language models for GED, or more flexibility in the gold standard, so as to accept two possible corrections.

In addition to tense choice, the model also missed errors related to singular and plural morphology. In Italian, nouns and adjectives inflect for gender and number; in the example below, the model fails to identify that the adjective "quant-" should take the plural suffix "-i" instead of singular "-o":

Sent:	Quant-o	anni	hai	?
Gloss:	how.many-SG	years	have	?
Gold:	i	c	c	c
Preds:	c	c	c	c
POLMS:	M	-	-	-
Subc.:	plural_morph	-	-	-
ADR:	R	-	-	-

Example 15: "How old are you?"

Swedish also has plural inflection on nouns and adjectives. In the example below, the plural suffix "-ar" should have been used on the noun "klädstil", but was forgotten by the learner:

Sent:	Då	kan	jag	[...]	pröva	olika	klädstil	.
Gloss:	Then	can	I		try	different	clothes.style	.
Gold:	c	c	c		c	c	i	c
Preds:	c	c	c		c	c	c	c
POLMS:	-	-	-		-	-	M	-
Subc.:	-	-	-		-	-	plural_morph	-
ADR: -	-	-	-		-	-	R	-

Example 16: "Then I can [...] try different style."
Complete sentence: "Then I can maybe be more experimental and try different style."

Let us now focus on subcategories that were specific to Swedish. Within the Morphology category, the greatest number of FNs concerns definiteness: namely, cases where the learner used the definite form of a noun, where an indefinite form would have been correct, and inversely.

In Swedish, definiteness is expressed morphologically: the noun is inflected for definiteness with a definite suffix, which can be either the singular -en, -et or -n, or the plural -na or -en, depending on the gender and phonological properties of the noun. On the other hand, the indefinite noun is not morphologically marked; instead, the indefinite article "en" or "ett" is added in front of the noun, as a separate word. This is illustrated below:

(7) a. *en* *situation*
DET situation

"a situation"

b. *situation-en*
Situation-DET

"the situation"

Interestingly, the model has difficulties distinguishing whether the definite marking or the indefinite determiner should be used in a given context. The sentence below illustrates this type of False Negative:

Sent:	Dessa	ändringar	kan	i	sin	tur	påverka
Gloss:	These	changes	can	in	its	turn	influence
Gold:	c	c	c	c	c	c	c
Preds:	c	c	c	c	c	c	c
POLMS:	-	-	-	-	-	-	-
Subc.:	-	-	-	-	-	-	-
ADR:	-	-	-	-	-	-	-

Sent:	regeringsnivå	och	politisk	situation	.
Gloss:	government.level	and	political	situation	.
Gold:	i	c	i	i	c
Preds:	c	c	c	c	c
POLMS:	M	-	M	M	-
Subc.:	def_indef	-	adj_morph	def_indef	-
ADR:	R	-	R	R	-

Example 17: "These changes can in turn influence government level and political situation." Note that there is also a missing determiner: "och **den** politiska situation"/"and **the** political situation". To ease the reading, only morphological annotations were included on the token "politisk".

The relevant errors were corrected in the following sentence:

(8)
Dessa ändringar kan i sin tur påverka regeringsnivå-n och den
These changes can in its turn influence government.level-DET and the
politisk-a situation-en .
political-DEF situation-DET .

Translation: "These changes can in turn influence the government level and the political situation."

On the other hand, definiteness FNs are much less represented in Italian, where definiteness is expressed with a separate article, placed in front of the noun: only one case among 78 morphology FNs. However, different subcategories were more represented in Italian: among others, prepositional morphology. When a preposition is followed by an article, the article is suffixed to the preposition, which sometimes involved phonological changes:

(9) **su il letto → sul letto*
On the bed → on.the bed.

Translation: "On the bed."

(10) **in la stanza → nella stanza*
In the room → in.the room.

Translation: "In the room."

Prepositional morphology represents 11 FNs out of 78 in Italian. In five cases, the article was completely forgotten, as in the example below:

Sent:	Alla	cortese	attenzione	di	Signora	Gabriella	[...]
Gloss:	to.the	kind	attention	of	Ms	Gabriella	
Gold:	c	c	c	i	i	c	
Preds:	c	c	c	c	c	c	
POLMS:	-	-	-	M	O	-	
Subc.:	-	-	-	prep_morph	cap	-	
ADR:	-	-	-	R	R	-	

Example 18: "To the kind attention of Ms. Gabriella [...]"

The preposition "di" lacks the determiner "la" and should be replaced by the inflected form "della". Three cases present the reversed scenario: there is an article where it is not needed. In the sentence below, "alla" should be replaced by "a" to be grammatical:

Sent:	[R]esto	alla	sua	disposizione	[...]
Gloss:	stay	at	your	disposition	
Gold:	c	i	c	c	
Preds:	c	c	c	c	
POLMS:	-	M	-	-	
Subc.:	-	prep_morph	-	-	
ADR:	-	R	-	-	

Example 19: "[...] I stay at your disposition [...]"

Verbal agreement is also well represented among FNs in Italian. Agreement is a complex part of the Italian grammar, with one different verbal suffix for each person in almost every tense. The example below illustrates a case of incorrect verbal agreement, which is not identified by the model:

Sent:	Potre-bbe	anche	operare	come	guida	perché	parl-o	anche	tedesco	[...]
Gloss:	could-3SG	also	operate	as	guide	because	speak-1SG	also	German	
Gold:	i	c	c	c	c	c	c	c	c	
Preds:	c	c	c	c	c	c	c	c	c	
POLMS:	M	-	-	-	-	-	-	-	-	
Subc.:	verbal_agr	-	-	-	-	-	-	-	-	
ADR:	R	-	-	-	-	-	-	-	-	

Example 20: "He could also operate as a guide because I also speak German [...]"

In the example above, the learner uses the 3rd person agreement on the first verb, but first person agreement on the second verb. Logic dictates that one of these is erroneous. The human annotator identifies the first one as incorrect, certainly based on the previous context, but the model does not identify any error in the present sentence. Logic tells us that both verbs should have the same subject, since one fact (becoming a guide) comes as a consequence the other fact (speaking German). However, one must admit that the sentence is technically grammatical, though logically unacceptable. Such cases are numerous among morphological FNs; let us present another relevant example:

Sent:	Sare-bbero	contento	di	una	Vostra	risposta	[...]
Gloss:	be-3PL	happy	of	a	your	reply	
Gold:	i	c	c	c	c	c	
Preds:	c	c	c	c	c		
POLMS:	M	-	-	-	-	-	
Subc.:	verbal_agr	-	-	-	-	-	
ADR:	R	-	-	-	-	-	

Example 21: "They would be happy to receive a reply from you [...]"

In the sentence above, the learner uses the 3rd person plural agreement "sarebbero" instead of the first person singular agreement "sarei"; the third person plural and first person singular can easily be confused, since they often (but not always) end with the suffix -o, as in "sono"/"I am" and "sono"/"they are", "ero"/"I was" and "erano"/"they were", "sarò"/"I will be" and "saranno"/"they will be", etc. Logically, the verb in the example above should refer to the first person singular, but grammatically, a third person agreement is acceptable. A tentative conclusion is that the model could lack the capacity to do logical deductions. Such difficulty could be reinforced by the fact that Italian is a pronoun-dropping language: in a majority of cases, there is no pronoun in front of the verb to indicates which person the verb should agree with. Hence, only knowledge of context and/or logical thinking can indicate what is the correct subject.

The model identified a total of 15 morphological mistakes that were not identified by the human annotator: eight in Swedish and seven in Italian. The breakdown of these correct predictions is represented in the graphs below:

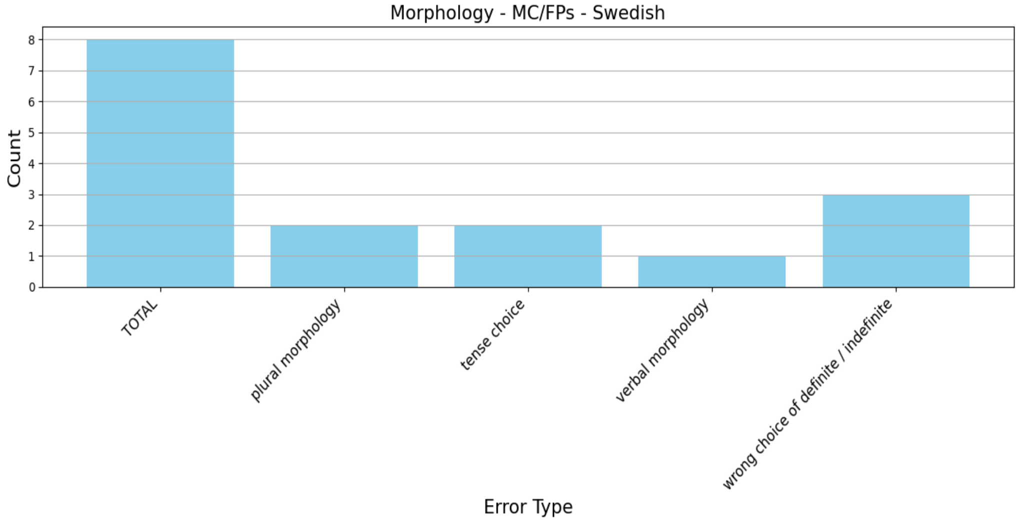


Figure 14: MC/FPs identified in Swedish per morphological subcategory.
Count: 'TOTAL': 8, 'plural morphology' (R): 2, 'tense choice' (R): 2, 'verbal morphology' (R): 1, 'wrong choice of definite / indefinite' (R): 3

Interestingly, these are also the subcategories that have just been identified as problematic for the model in Swedish. A similar observation can be made for Italian:

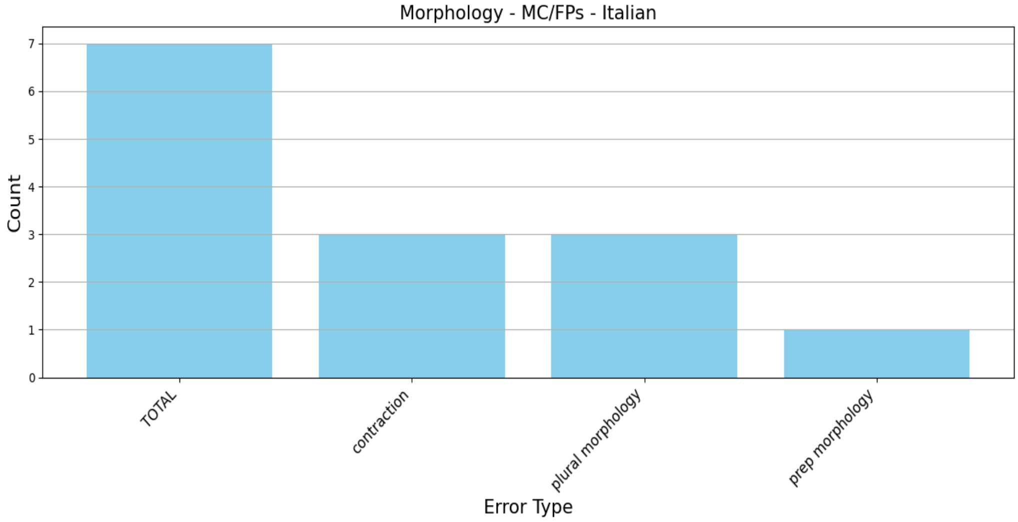


Figure 15: MC/FPs identified in Italian per morphological subcategory.
Count: 'TOTAL': 7, 'contraction' (R): 3, 'plural morphology' (R): 3, 'prepositional morphology' (R): 1

The subcategories "plural morphology" and "prepositional morphology" are both problematic for the model. Yet, the model also manages to identify errors that are not identified by the human annotator - although in limited instances.

Morphology is a complex area of any language's grammar, and poses difficulties in adulthood in particular (Nikolaev et al. 2019). It is hence interesting to see that morphology errors are also numerous among FNs: the False Negatives could indeed reflect the difficulties of the learners. Further studies could carry out analyses of the test data, and analyse how many cases of morphological mistakes the model correctly identifies. This would allow for a better understanding of the model's morphological capabilities. For now, the analysis of morphological FNs hints that the model might lack the capacity to rely on context and knowledge of the world to identify morphological mistakes, such as those related to tense choice or verbal agreement.

3.2.5 Syntax

Syntax is one of the most represented POLMS category among FNs in both Swedish and Italian, representing 36% of the False Negatives in both languages. However, syntax is also well represented among the errors that the model identified, but not the human annotator: 33% in Swedish and 28% in Italian. Taking a closer look at the syntactic error subcategories could cast light on this apparent paradox.

Syntactic FNs can be of four kinds: a missing word, an extra word, an issue with word order or formulation. A missing word requires the addition of a word (A), an extra word requires deletion (D), and word order requires the re-ordering of the words , hence replacement (R). Formulation, on the other hand, concerns more complex errors, often spanning more than one token, and may require all three correction operations (ADR). The distribution of FNs across these categories is illustrated below for Swedish and Italian:

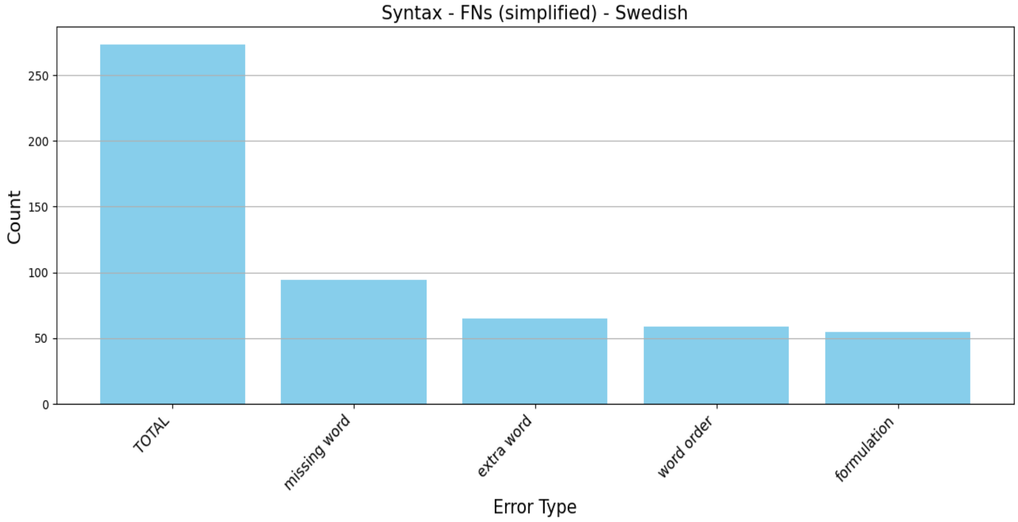


Figure 16: Syntax FNs classified into subcategories in Swedish.
Count: 'TOTAL': 273, 'missing word' (A): 94, 'extra word' (D): 65, 'word order' (R): 59, 'formulation' (ADR): 55.

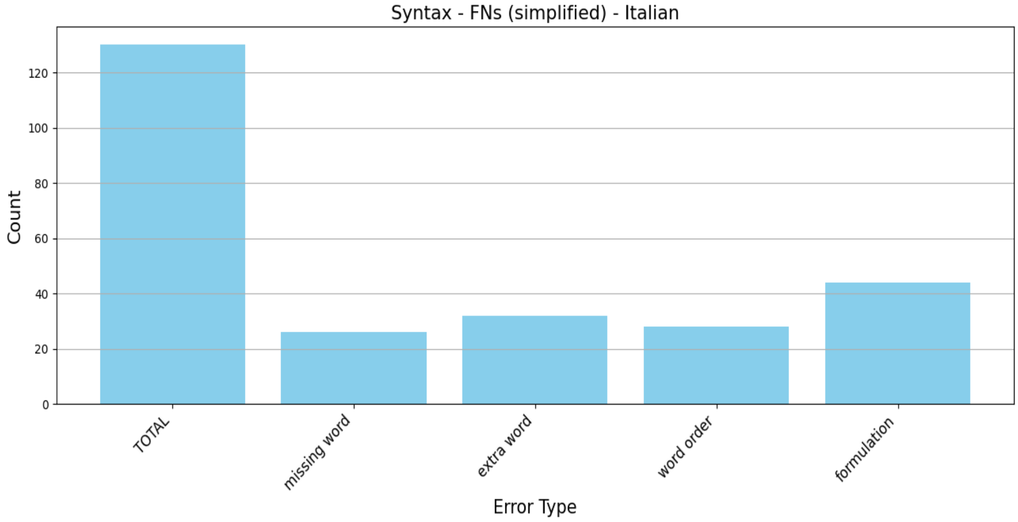


Figure 17: Syntax FNs classified into subcategories in Italian.
Count: 'TOTAL': 130, 'missing word' (A): 26, 'extra word' (D): 32, 'word order' (R): 28, 'formulation' (ADR): 44

Syntax is a particularly challenging category, since it has to deal with tokens that are not technically present in the data. As mentioned earlier, cases of missing tokens are particularly challenging to annotate; in the MultiGED shared task, the following token is counted as incorrect to signal a missing token. This approach, like any other, has limitations, as the language model might have difficulties understanding this guideline. This could explain the high amount of missing tokens, particularly in Swedish, where missing words are the most represented type of syntactic FNs. In Italian, formulation is the most problematic subcategory. This subcategory represents complex sets if errors, since it encompasses cases where the mistake cannot simply be corrected by one single ADR operation. The example below illustrates a formulation error:

Sent:	[...]	nella	città	di	storia	[...]
Gloss:		in:DET	city	of	history	
Gold:		c	i	i	i	
Preds:		c	c	c	c	
POLMS:		-	S	S	S	
Subc.:		-	formulation	formulation	formulation	
ADR:		-	ADR	ADR	ADR	

Example 22: "In the city of history"
Translation (intended): "In the historical city"

The corrected version of the sentence is:

(11) *Nella città storica [...]*
In:DET city historical

Translation: "[I]n the historical city [...]"

To correct the expression above, the syntactic structure of the phrase had to be changed, which entailed a deletion and a replacement.

Overall, we can see that all four syntactic subcategories are well represented among FNs, and it is therefore of interest to break down these subcategories. The graphs below illustrate the subcategories, but this time the "missing word" and "extra word" subcategories have been broken down according to the part of speech of the token they refer to, when possible:

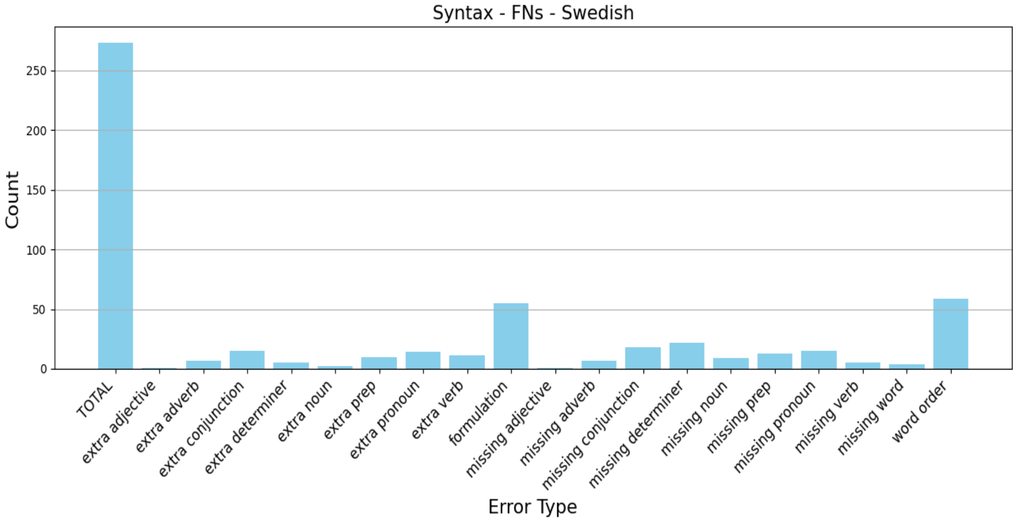


Figure 18: Syntax FNs classified into subcategories in Swedish, including POS.
Count: 'TOTAL': 273, 'extra adjective' (D): 1, 'extra adverb' (D): 7, 'extra conjunction' (D): 15, 'extra determiner' (D): 5, 'extra noun' (D): 2, 'extra prep' (D): 10, 'extra pronoun' (D): 14, 'extra verb' (D): 11, 'formulation' (ADR): 55, 'missing adjective' (A): 1, 'missing adverb' (A): 7, 'missing conjunction' (A): 18, 'missing determiner' (A): 22, 'missing noun' (A): 9, 'missing prep' (A): 13, 'missing pronoun' (A): 15, 'missing verb' (A): 5, 'missing word' (A): 4, 'word order' (R): 59.
Note that in four cases, it was not possible to deduct which POS was missing, and we hence kept the subcategory "missing word".

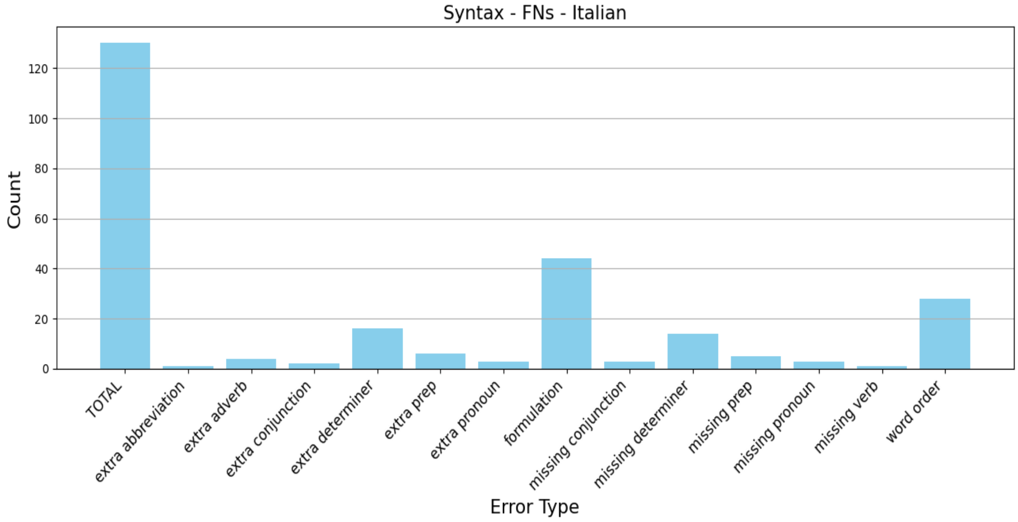


Figure 19: Syntax FNs classified into subcategories in Italian, including POS. Count: 'TOTAL': 130, 'extra abbreviation' (D): 1, 'extra adverb' (D): 4, 'extra conjunction' (D): 2, 'extra determiner' (D): 16, 'extra prep' (D): 6, 'extra pronoun' (D): 3, 'formulation' (ADR): 44, 'missing conjunction' (A): 3, 'missing determiner' (A): 14, 'missing prep' (A): 5, 'missing pronoun' (A): 3, 'missing verb' (A): 1, 'word order' (R): 28

Determiners are the most represented POS within the "missing" and "extra word" categories, with 27 cases in Swedish and 30 in Italian.

The use of determiners in Italian is governed by a relatively complex set of rules. For example, singular possessive determiners are always preceded by the definite article, except when talking about family members. Compare "this is my son" and "this is my cat" in the example below:

(12) a. *Questo è il mio gatto.*
This is DET my cat

Translation: "This is my cat."

(13) a. *Questo è mio figlio.*
This is my son

Translation: "This is my son."

However, the definite article must be placed in front of the noun when the noun is plural:

(14) a. *Questi sono i miei gatti.*
These are DET my cats

Translation: "These are my cats."

(15) a. *Questi sono i miei figli.*
These are DET my sons

Translation: "These are my sons."

Presence or absence of determiner is hence governed by the semantics of the noun it qualifies. Five cases of missing and extra determiners are related to family relationships in our data. Determiners can also be anomalous in other contexts. For example, I "speak Italian", but I "study the Italian":

(16) *Parl-o italiano.*
Speak-1SG Italian

Translation: "I speak Italian."

(17) *Studi-o l'italiano.*
Study-1SG DET-Italian

Translation: "I study Italian."

Our data contains two cases similar to the example above. Additionally, there exists important dialectal variations in the use of determiners in Italian. Cardinaletti and Giusti (2020: 679) showed that there was no less than four ways of using determiners in the sentence "I do not eat meat" across Italy, as can be seen in the example below:

(18) a. *Non mang-io carne.*
Not eat-1SG meat

b. *Non mang-io la carne.*
Not eat-1SG DET meat

c. *Non mang-io di carne.*
Not eat-1SG of meat

d. *Non mang-io de-lla carne.*
Not eat-1SG of-DET meat

e. *Non mang-io certa carne.*
Not eat-1SG certain meat

Translation: "I do not eat meat". Example taken from Cardinaletti and Giusti (2020: 679).

An interesting follow-up question is whether the Common Crawl data on which XLM-RoBERTa was initially trained contains different dialectal variations. This could possibly make it difficult for the model to generalise and identify errors related to determiners. Further studies could investigate this issue and possibly emphasise the importance of dialectal variation when creating and training LMs.

The model does not display such a pronounced tendency towards missing and extra determiners in Swedish, although such cases also exist among Swedish FNs. In some of the cases within the "missing determiner" category, the sentence turns out to be grammatically correct, but does not make much sense logically. Indeed, certain plural words in Swedish have the same form as the singular, and only the presence or absence of the indefinite determiner will hint whether the word is plural or singular:

(19) a. *Hus*
House/house:PL

Translation: "House" or "Houses"

(20) a. *Ett hus*
DET house

Translation: "A house."

An illustrative example from our data concerns the word "bibliotek"/"library" in Swedish. In the following example, the sentence is grammatically correct, but our knowledge of the world tells us that it should have been a singular, and that the indefinite determiner is hence lacking:

Sent:	[A]lla	människor	behöver	bibliotek	i	del	av	plast	som	bor	.
Gloss:	all	people	need	library:PL	in	part	of	plastic	that	live	.
Gold:	c	c	c	i	c	i	i	i	i	i	c
Preds:	c	c	c	c	c	i	i	i	i	i	c
POLMS:	-	-	-	S	-	-	-	-	-	-	-
Subc.:	-	-	-	missing_det	-	-	-	-	-	-	-
ADR:	-	-	-	A	-	-	-	-	-	-	-

Example 23: "Everyone needs libraries in the part of the plastic where they live."
Translation (intended): "Everyone needs libraries in the part of the place where they live."

To correct this error, it is necessary to add the determiner "ett" in front of the word "bibliotek":

(21) a. [A]*lla människor behöver ett bibliotek i del av plast som bor .*
All people need DET library in part of plastic that live .

Translation: "Everyone needs a library in the part of the plastic where they live." Translation (intended): "Everyone needs a library in the area where they live."

It is indeed with knowledge of the world that one may assume that there is usually only one library per area. Without such knowledge, it is natural to count the sentence as grammatically correct.

Overall, function words are more represented than content words among FNs. Taken together, determiners, prepositions, conjunctions and pronouns represent respectively 41% and 40% of the FNs related to syntax in Swedish and Italian. However, this could also be due to the fact that function words are also difficult to use for learners. Further investigations could conduct analyses of both the training and test datasets to determine the distribution of errors in content words and function words.

The model identified a certain number of syntactic mistakes that were not identified by the human annotator: 14 in Swedish and 6 in Italian. Their distribution across subcategories (including POS) is illustrated below:

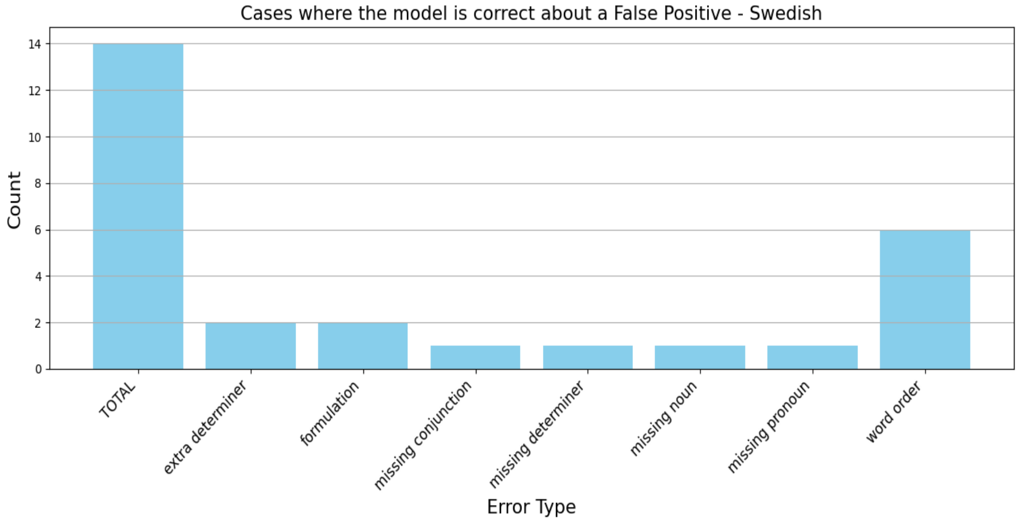


Figure 20: syntax MC/FPs classified into subcategories in Swedish.
Count: 'TOTAL': 14, 'extra determiner' (D): 2, 'formulation' (ADR): 2, 'missing conjunction' (A): 1, 'missing determiner' (A): 1, 'missing noun' (A): 1, 'missing pronoun' (A): 1, 'word order' (R): 6.

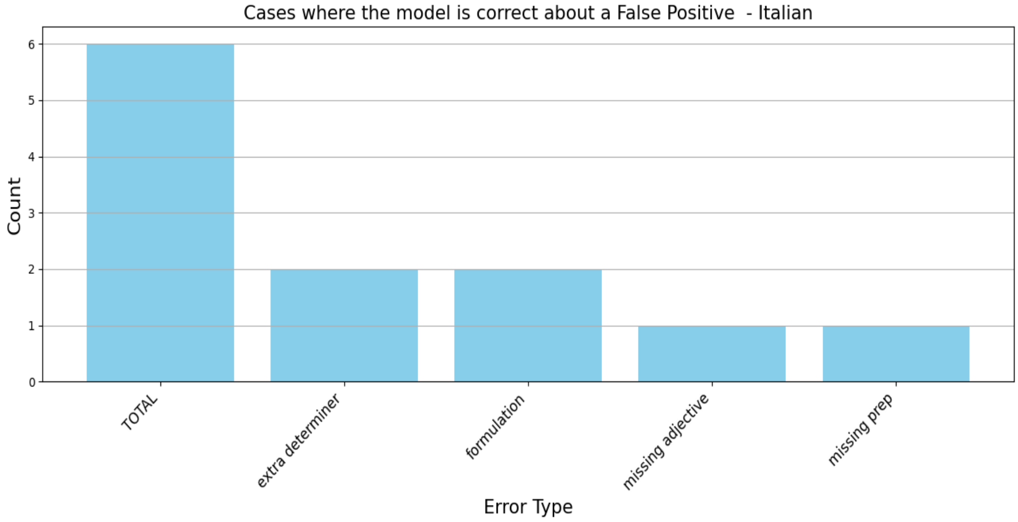


Figure 21: syntax MC/FPs classified into subcategories in Italian.
Count: 'TOTAL': 6, 'extra determiner' (D): 2, 'formulation' (R): 2, 'missing adjective' (A): 1, 'missing preposition' (A): 1.

The errors are well distributed across subcategories, particularly in Swedish where there is often no more than one or two cases per subcategory. Word order is slightly more represented in Swedish, while formulation and extra determiners are the most represented subcategories in Italian. Because of the paucity of cases, one can hardly draw convincing conclusions from these graphs, but it is worth noting that the model still manages to identify missing tokens that were not marked by the human annotator in both languages. This suggests that the model took into account the annotation guidelines mentioned above, at least on these occasions. This shows that there are certain irregularities in the model's learning.

3.3 Summary and conclusion

First and foremost, the present chapter provided a new version of the test dataset, curated for the following aspects: first, certain tokens that were annotated as incorrect given a wider essay context were re-annotated as correct within their limited context. Second, we showed that there are sometimes multiple ways of correcting a sentence, and we reviewed some of the annotations so as to fit the model's predictions, hence avoiding skewing the results at the disadvantage of the model. We then conducted a linguistic analysis of the cases where the model was indeed incorrect, and drew conclusions as to where the strengths and weaknesses of the model could lie.

As previously mentioned, any conclusion that is drawn here is tentative, as there lacks an analysis of the True Positives identified by the model. Let us illustrate this with an example: the model missed a number of errors related to definiteness in Swedish.

This hints that the model has difficulties with this particular area of the grammar. If the analysis of True Positives reveals that the model also correctly identified definiteness errors on numerous instances, then the high number of FNs related to this particular type of error could be due to other factors; for example, an over-representation of this error type in the data. On the other hand, if the model identifies few True Positives related to definiteness, the hypothesis of a difficulty with this specific error type is reinforced. With this in mind, we shall now state our main hypotheses.

Cross-linguistically, the areas of syntax, lexicon and morphology are more represented among FNs than punctuation or orthography. This allows us to make a first tentative conclusion: the analysis of the False Negatives hints that the model could have more difficulties with syntax, morphology and lexicon than punctuation and orthography.

Within the lexicon category, the model did not identify a large number of erroneous prepositions. This could be due to their anomalous nature, and to the fact that knowledge of the world is often needed to correctly use prepositions. Widely used verbs as well as modal verbs are also common among lexical FNs in the Swedish data. Overall, lexical mistakes are often dependent on the context, as well as world knowledge, to be identified. The present data hence suggests that these might not yet be perfectly acquired by the model.

Errors related to tenses were common among morphological FNs. This could hint that the model struggles to make use of the sentence's content to situate an action within its correct temporality. Further, a number of errors related to definiteness, verbal agreement and prepositional morphology were also missed by the model. Such errors can sometimes give rise to sentences that are illogical, but still grammatically correct. The model could therefore lack the interpretative power needed to identify these errors.

Additionally, tokens that should be deleted or added were numerous among syntactic FNs. Annotation guidelines might be partly responsible for these FNs. Analysis of the True Positives would reveal if the model also correctly identified some of these errors, and if so, to what extent, so as to dispute or or further confirm the hypotheses made during the linguistic analysis.

Overall, four main hypotheses can be drawn regarding XLM-RoBERTa as fine-tuned by EliCoDe. First, the model might lack knowledge of the world on certain occasions. Second, it seems like the model does not always take into account the whole context given within a sentences. Third, the model also seems to lack logical thinking in certain contexts. Finally, the model missed a number of redundant and missing words, which could partly be due to data annotation.

Despite these potential difficulties, XLM-RoBERTa still performed well on the provided datasets, as shown by the general results of precision, recall and F0.5 in Table 4 and 5, which were always in the range 66%-86%. Further, the model has more than once identified errors that were not identified by the human annotator. In the next section, we will attempt to use m-DeBERTa-V3, a brand new model release in 2023, for the purpose of GED, and compare its results to the ones of XLM-RoBERTa as fine-tuned by EliCoDe.

4 Using m-DeBERTa-V3 for GED

So far, the state-of-the-art LM for English GED is ELECTRA (Clark et al. 2020, Yuan et al. 2021), while XLM-RoBERTa performs best for multilingual GED (Conneau et al. 2020, Colla et al. 2023). ELECTRA was developped in 2020 by Clark et al. (2020) and presents a key difference with XLM-RoBERTa: while the latter is trained on Masked Language Model (MLM), where the model has to predict a hidden token (Liu et al. 2019: 2), ELECTRA is trained on Replaced Token Detection (RTD). This technique consists in replacing real tokens by synthetically generated tokens, and training the model to identify such tokens (Clark et al. 2020: 1).

ELECTRA outperforms XLM-RoBERTa in English GED, as shown by the results below:

Model	**Precision**	**Recall**	**F0.5**
ELECTRA (Yuan et al. 2021)	82.5	50.49	72.93
XLM-RoBERTa (Colla et al. 2023)	73.64	50.34	67.40

Table 12: Results of XLM-RoBERTa (Colla et al., 2023) and ELECTRA (Yuan et al., 2021) on the English FCE dataset.

According to Yuan et al. (2021: 8730), ELECTRA performs well in GED because RTD is a training technique that is conceptually close to error detection: finding synthetic tokens is indeed similar to finding erroneous tokens.

A question that naturally follows is whether ELECTRA could also become the new SOTA for multilingual GED - a question that cannot be answered for now, since ELECTRA only exists for English at present. However, a new multilingual LM has recently been released, combining the strenghts of both ELECTRA and XLM-RoBERTa: m-DeBERTa-V3 (He et al., 2023). This model is of particular interest in GED, since it combines the strenghts of LMs that perform exceptionally well in error detection.

4.1 The architecture of m-DeBERTa-V3

m-DeBERTa-V3 is the result of several developments, which started in 2021 with the release of DeBERTa, the first version of m-DeBERTa-V3.

DeBERTa was initially released as an improved version of both BERT and RoBERTa, using disentangled attention mechanism and an enhanced masked decoder (He et al., 2021: 1). Disentangled attention mechanism signifies that each token is represented by two vectors, one encoding its position and one encoding its content, while enhanced mask decoder is used to include absolute positions in the decoding layer (He et al., 2021: 1). The pre-training, on the other hand, is similar to the one of Ro-BERTa, since it makes use of Masked Language Model (MLM) to predict masked tokens (He et al., 2021: 2). The third version of DeBERTa, DeBERTa-V3, changes the pre-training procedure by replacing MLM by Replaced Token Detection, the same technique used by ELECTRA (He et al., 2023, Clark et al. 2020). Thanks to these improvements, DeBERTa-V3 became the new SOTA among LMs in numerous NLP tasks, outperforming BERT, RoBERTa and ELECTRA (He et al., 2023: 7). Furthermore, He et al. (2023) also released a multilingual version of DeBERTa-V3, m-DeBERTa-V3, trained on the same 100 languages CommonCrawl data as XLM-RoBERTa (He et al., 2023: 8), which significantly outperformed XLM-RoBERTa on all the tested languages (He et al., 2023: 9) in a wide range of NLP tasks.

m-deBERTa-V3 is of major interest for multilingual GED, since it combines the strenghts of models that have already presented promising results in grammatical error detection. On the one hand, it presents the system specifications that have been proven efficient in RoBERTa, while also improving on them using disentangled attention (He et al. 2023: 4). On the other hand, it is pre-trained using ELECTRA's Replaced Token Detection, a procedure conceptually similar to GED, which has demonstrated state-of-the-art performances for English GED (Clark et al. 2020). The differences and similarities between ELECTRA, XLM-RoBERTa and m-DeBERTa-V3 are presented in the table below:

Model	**Language(s)**	**Training data**	**Pre-training technique**
XLM-RoBERTa (Conneau et al., 2020)	100	CC100 multilingual data	Masked Language Model
ELECTRA (Clark et al., 2020)	1	BookCorpus + English Wikipedia	Replaced Token Detection
m-DeBERTa-V3 (He et al., 2023)	100	CC100 multilingual data	Replaced Token Detection

Table 13: XLM-RoBERTa, ELECTRA, and m-DeBERTa-V3 specifications

The present work proposes to run a first try at using m-DeBERTa-V3 for GED, and discusses possible paths for future improvements. m-DeBERTa-V3 will therefore be trained to perform Grammatical Error Detection in two languages: Italian and Swedish, so as to compare the results with the ones from the participants of the MultiGED-2023 shared task.

After the revision period, the code used to fine-tune m-DeBERTa-V3 will be published on Github.

4.2 Datasets

The present thesis uses the same training, development and test datasets as the teams of the MultiGED-2023 shared task for Swedish and Italian. As mentioned in the previous sections, the datasets are collection of sentences, in random order, taken from learners' essays. Each token is annotated as "correct" or "incorrect". The data is collected from the following corpora:

Swedish	SweLL-gold (Volodina et al., 2019)
Italian	MERLIN (Boyd et al., 2014)

Table 14: Corpora from which the datasets were taken in the MultiGED-2023 shared task. See Volodina et al. (2023) for a detailed description of each dataset.

The results of the present work can therefore be easily compared to the results of Colla et al. (2023) for Swedish and Italian. While other languages could have been tested, the present work limited itself to Swedish and Italian, since they are the smallest datasets and can hence be trained quicker than the other datasets; they are therefore appropriate for a work of this scope.

4.3 Data processing and fine-tuning

To allow m-DeBERTa-V3 to perform GED, the data has to be processed, and the model fine-tuned. The present subsection details the technical processes, starting with data processing.

4.3.1 Processing the data

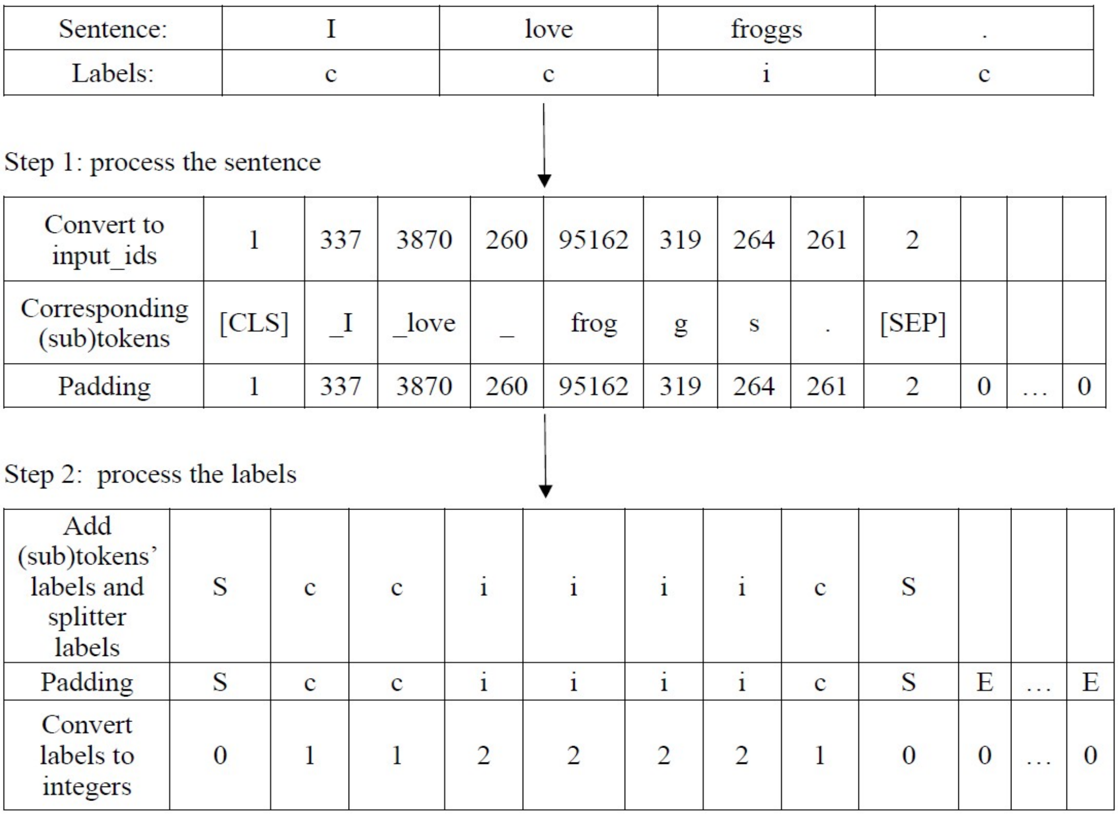
Before being fed to the model, the data undergoes several processing steps. Both the sentences and their corresponding labels undergo processing. First of all, the sentence is tokenized using the m-DeBERTa-V3 tokenizer. This signifies that each sentence is represented by three different tensors, which are the input IDs, the token type IDs and the attention mask.

In the input IDs tensor, each token is represented by an integer, which corresponds to its index in the model's vocabulary. Certain tokens are divided into subtokens by the tokenizer; for example, plural words or inflected verbs are segmented based on their affixes. The sign "_" is added as a separated subtoken and always precedes the main token. For example, the word "frogs" will be represented as [_, frog, s]. In cases where the main token is not divided into subtokens, the sign "_" is directly attached to the token: for example the word "frog" will be represented as [_frog] . The tokenizer also adds an integer "1" at the very start of the tensor to indicate the beginning of the sentence, and a "2" after the last token's input ID to indicate the end of the sentence. Each sentence is then padded to a length of 128, to ensure that every sequence has the same length. In practice, this means that if a sentence is shorter than 128 tokens, the tokenizer adds the input ID 0 for each empty token up until there is a total of 128 input IDs in the tensor. Conversely, sentences that are longer than 128 tokens are divided into separate sentences. While this approach may result in some loss of context, it's important to note that sentences of this length are exceedingly rare, and are hence unlikely to influence the results. In fact, there were only two instances in Swedish and none in Italian.

The token type IDs tensor always contains 0s up until length 128, since all tokens belong to the same sentence. Token type IDs are used to identify whether a certain token belongs to a sequence A or B. Such distinction is useful in NLP tasks that work with more than one segment, such as question answering or natural language inference. In our case, the model works with single segments, and this tensor hence only contains 0s.

Finally, the attention mask discriminates which tokens should receive attention from the model: the tokens that should receive attention are labelled "1" in the attention mask, while tokens that can be ignored are represented as "0". In the present task, every token is represented by a "1", since it is a token classification task: each token should therefore be considered.

Then, the corresponding sequence of labels is also encoded for the training and development data. Since a 1 and a 2 are added to the sentence input IDs to mark the beginning and the end of a sentence, a splitter label "S" is also added at the beginning and end of the sequence of labels. Then, additional labels are added for each subtoken created by the tokenizer: every subtoken is assigned a label, which corresponds to the label of the original token. Finally, an empty label "E" is added up until a length of 128 for labelling sequences that are shorter than 128. Each of the labels is then converted into an integer: {correct/c: 1, incorrect/i: 2, splitter/S: 0, empty/E: 0}. Let us illustrate the data processing with the custom erroneous "I love froggs" in English:



Sentence:	I	love	froggs	.
Labels:	c	c	i	c

Step 1: process the sentence

Convert to input_ids	1	337	3870	260	95162	319	264	261	2			
Corresponding (sub)tokens	[CLS]	_I	_love	_	frog	g	s	.	[SEP]			
Padding	1	337	3870	260	95162	319	264	261	2	0	…	0

Step 2: process the labels

Add (sub)tokens' labels and splitter labels	S	c	c	i	i	i	i	c	S			
Padding	S	c	c	i	i	i	i	c	S	E	…	E
Convert labels to integers	0	1	1	2	2	2	2	1	0	0	…	0

Figure 22: Processing steps for the sentence "I love froggs.".

This process is repeated for every sentence in the datasets and its corresponding label sequence.

4.3.2 Fine-tuning

After processing, the training and development data is passed into the model and trainer. The batch size is set to 2: a small batch size allows the model to adapt its weights every two sequences, thereby improving its learning capacities. The number of epochs is set to 7, so as to have a relatively rapid training while also allowing enough time for the model to learn. The dropout rate is set to 0.3: during each epoch, 30% of the neurons in the model are randomly deactivated. This encourages the model to rely on the entirety of its neurons, rather than becoming overly dependent on a specific subset, thus reducing the risk of overfitting. The learning rate is set to a value that is traditionally used to fine-tune language models, namely 3e-04, and we make use of the CosineAnnealingWarmRestarts scheduler, which optimizes the model by adjusting the learning rate throughout the training. We choose to use the default optimizer AdamW, traditionally used in deep learning. The weights of the encoder are frozen, so as to keep all the knowledge m-DeBERTa-V3 has learnt during its pretraining. Additionally, we use mean pooling, which captures the mean of all token embeddings, thus aggregating information from the whole sequence, while also ensuring that the model always produces a fixed-size output.

When it comes to the loss function, our model uses the column-wise Mean Square Error (MSE) with added weight. Column-wise MSE calculates the difference between the predicted and truth labels across the entire sequence. The formula is illustrated below:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{1}$$

Since the data is imbalanced and contains a lot more correct and empty tokens than incorrect tokens, we added a weight to the function: if the truth label is "2"/"incorrect" but the model predicts "0"/"empty" or "1"/"correct", the error for this given token is multiplied by 50. This allows the model to focus on incorrect tokens, and limits the effect of the data imbalance. The loss function then computes the average error across the entire sequence and squares it. The model was trained on the GPU provided by the University of Gothenburg, MLT GPU.

4.3.3 Predictions processing

In order to compare our predictions to the ones of the MultiGED-2023 teams, the predictions had to be processed.

First, the predictions of subtokens are removed. This implies re-encoding each sentence, and checking whether the token is or starts with the sign "_". If so, it is a token, and its prediction should be kept. If the token does not start with the sign "_", it is a subtoken, and its prediction can be removed from the list of predictions. For example, the word "lämna"/"leave", is divided in the subtokens [_,lämn,a]. Each of these has been assigned a prediction, which we can expect to be identical, since the same labels are assigned to all subtokens in the training process. Only the prediction of the initial token is kept, hence ensuring that each word only receives one prediction.

4.4 Results

So as to compare our results to the ones of EliCoDe, we evaluate the predictions using the evaluation file that was used during the MultiGED-2023 shared task.

The results are calculated based on the ability of the model to recognise incorrect tokens. The number of True Positives, False Positives and False Negatives is used to calculate Precision, Recall and F0.5 (see section 2.2 for a detailled explanation of the formulas).

The table below presents the results obtained by our fine-tuned version of m-DeBERTa-V3 for Italian and Swedish:

	Precision	**Recall**	**F0.5**
Swedish	0.36	0.02	0.08
Italian	0.54	0.01	0.04

Table 15: Results for Swedish and Italian produced by our fine-tuning of m-DeBERTa-V3.

Compared to the teams that participated in the competition, m-DeBERTa-V3 would systematically rank last, except for the Precision score, where it would rank second to last in both languages (see table 4 and 5).

4.5 Discussion

These results are far from equalling the ones of XLM-RoBERTa, in both experimental settings and for both languages. See below the results of XLM-RoBERTa as fined-tuned by EliCoDe:

	Precision	**Recall**	**F0.5**
Swedish	0.82	0.66	0.78
Italian	0.87	0.68	0.82

Table 16: Results of the EliCoDe team for Swedish and Italian (Volodina et al., 2023: 9)

A question that naturally follows is whether our model learns at all, given the low results. To verify this, we plotted the loss obtained at the end of each epoch: if the loss lowers over time, the model is learning, since it makes predictions that are closer to the truth for each epoch. If the loss does not decrease, the model is not learning.

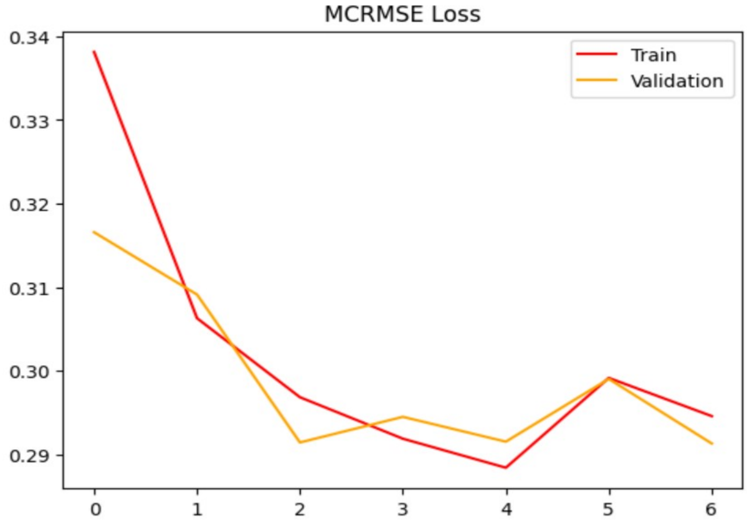


Figure 23: the evolution of Loss using the Column-Wise Mean Squared Error (colwiseMSE) Loss Function across Epochs for the Swedish dataset. The red line represents the progression of the training loss, the orange line represents the evolution of the validation (development) loss.

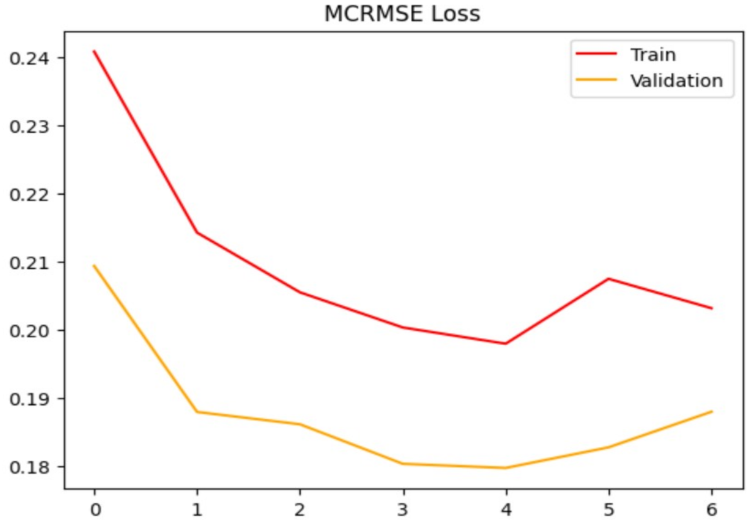


Figure 24: the evoluation of Loss using the Column-Wise Mean Squared Error (colwiseMSE) Loss Function across Epochs for the Italian dataset.The red line represents the progression of the training loss, the orange line represents the evolution of the validation (development) loss.

The loss shows some fluctuations over time, but gradually decreases, thus signifying that the model is progressively learning. A closer look at the predictions given by the model brings insights as to what the model does and does not learn.

To do so, we plotted the truth values for the Swedish and Italian test datasets into a graph. The x-axis represents the position of each token in a sequence, from 0 to 128. The y-axis shows the count of each label. The coloured lines represent the different numerical labels. These graphs hence illustrate the count y of each label at position x in every sentence of the test dataset:

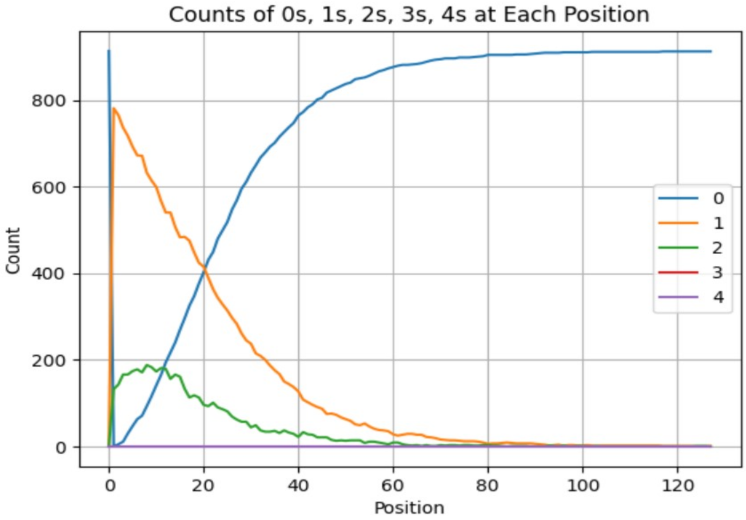


Figure 25: plotted truth values for the Swedish dataset.

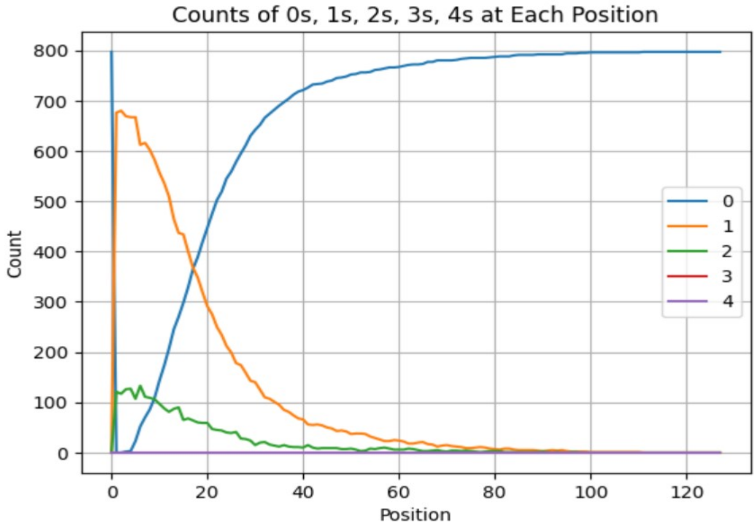


Figure 26: plotted truth values for the Italian dataset.

At position 0, the label is 0 in more than 800 sequences in both datasets. In other words, the token at position 0 in every single sentence of both datasets is always empty. During the data processing, we indeed added a splitter token at the start and end of each sentence. At position 1, the token is almost systematically correct: in around 800 sentences in Swedish and almost 700 in Italian, the first word is correct. On the other hand, the picture is more varied for the 20th token in each sentence: in about 100 sentences, the 20th token is incorrect in Swedish, against 50 cases in Italian. In 400 Swedish sentences, the 20th token is correct, and in 400 other sentences, the 20th token is simply empty. Let us now plot the same graph, but with the predictions for Swedish and Italian:

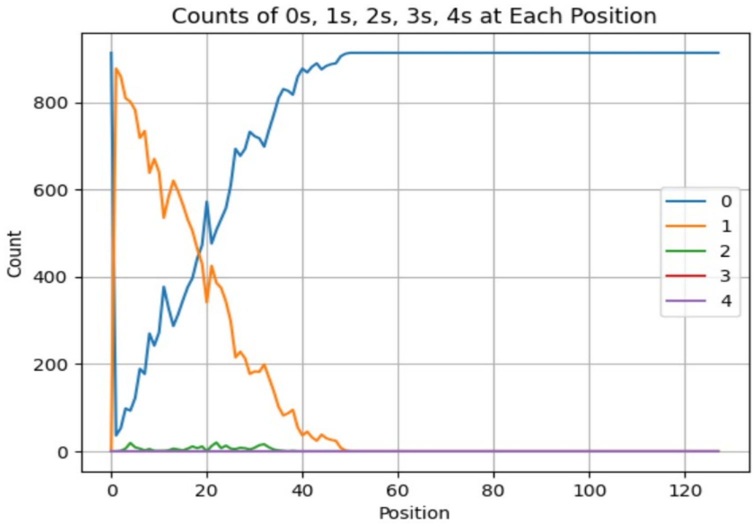


Figure 27: plotted predicted values for the Swedish dataset.

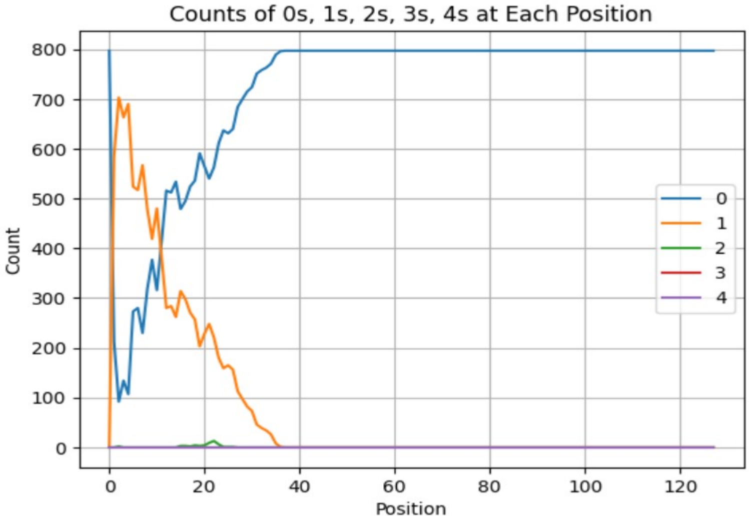


Figure 28: plotted predicted values for the Italian dataset.

These graphs shows that the model is learning about correct and empty tokens: both curves are similar to the ones from the plotted truth values. However, the curve for the incorrect tokens "2" does not follow the pattern of the truth values. Our model barely identifies any incorrect tokens. The data contains a lot more correct and empty tokens than incorrect tokens, and our model's difficulties seem to stem from the imbalance in the data, despite the added weight of the loss function.

There can be numerous reasons to explain the poor results of m-DeBERTa-V3 in GED, compared to XLM-RoBERTa. To start with, the architecture of each model is different, and the fact that m-DeBERTa-V3 uses Replaced Token Detection instead of Masked Language Model might hurt the performance, contrary to what was initially suggested. If this is true, then the excellent performances of ELECTRA in English could be due to a different factor than its RTD pretraining. This, however, seems unlikely, given the drastic difference between ELECTRA and m-DeBERTa-V3 in scores, but great similarity in pretraining procedure. Hence, our model's poor performance could be due to other factors.

First of all, there are differences in the experimental settings used to fine-tune XLM-RoBERTa compared to our model. In our experiment, the same model is used for both Italian and Swedish, while EliCoDe fine-tuned a model for each language (Colla et al., 2023: 24). EliCoDe converted the task into a Name Entity Recognition task and made use of the ClinicalTransformerNER package to fine-tune this model (Colla et al., 2023: 28), while the present experiment does not use such package[3]. If one looks into the code of the EliCoDe team[4], one notices that a wide range of optimization processes are used: adversial training, Xavier and Kaiming Initialisation, shared dropout, etc. No such

[3] It is worth noting that the ClinicalTransformerNER package is not available for m-DeBERTa-V3 as of June 2023. See https://github.com/uf-hobi-informatics-lab/ClinicalTransformerNER

[4] The code of the EliCoDe team is available on Github: https://github.com/davidecolla/EliCoDe

processes are used in the present code.

It is also worth noting that external factors also limit the possibilities when it comes to training and improving the performances of m-DeBERTa-V3. To start with, time is an important factor. Training the present model can go fast (around one hour) but only on a GPU while using an accelerator. Training the model on MLT GPU is only possible when the server is not used for other tasks by other users - a factor which is nearly impossible to control. If the server is already used, the training time can be multiplied by 10 or more. In numerous cases, the CUDA ran out of memory due to other tasks and projects being run on MLT GPU simultaneously. In such cases, we coded the model so that it would start training as soon as the server is available again. It is not rare to have to wait one day or more to start training. This therefore limited the number of possibilities we had to train the model, and the scope of improvements we could test.

This suggests that the performances of m-DeBERTa-V3 could be greatly improved if one could make numerous attempts and try different optimization techniques. The subsection below shall therefore discuss paths for improvement.

4.5.1 Paths for improvement

A first possible improvement concerns the loss function. The column-wise MSE is used in the present experiment. This function takes the mean of the differences between the truth and the prediction for each column in the tensor, and squares this difference. However, such a loss function is more commonly used with continuous values: since the difference is squared, it heavily penalises scores that are numerically further away from the truth. Hence, if the model predicts that an empty token (0) is in fact an incorrect token (2), the mean square error will be larger than if the model predicts that it is a correct token (1). However, we work with categorical values (empty, correct, incorrect), with the goal of finding incorrect tokens. The attributed integers are only used to make the labels interpretable by the computer, and do not hold numerical significance. Therefore, a more suited loss function is the cross-entropy loss, which assigns a probability of each class (empty, correct, incorrect) for a given token. We therefore attempted to use cross-entropy loss on the Swedish data, with the same hyperparameters. The results are presented below:

	Precision	**Recall**	**F0.5**
Swedish	0.28	0.02	0.07

Table 17: Results for the Swedish data using the cross-entropy loss function.

The results show that using the cross-entropy loss function did not improve the perfomance of m-DeBERTa-V3. Compared to the teams from the MultiGED shared task, m-DeBERTa-V3 would once more systematically rank last, except for the Precision score, where it would rank second to last (see table 4 and 5). A closer look at the progression of the losses, as well as the predictions, could cast light on these results.

The losses are much higher for each batch than with the column-wise MSE: starting at 283 in cross-entropy against 0.34 with column-wise MSE. Despite this high loss, the model is learning, since the learning rate decreases over time, despite some fluctuations:

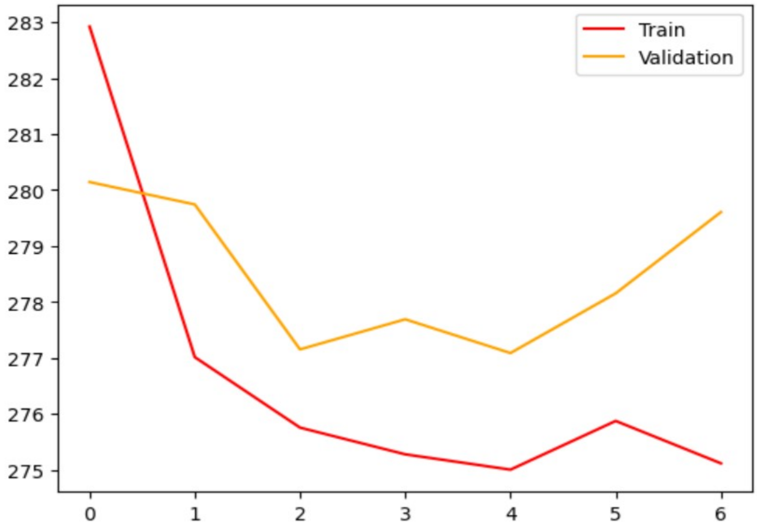


Figure 29: the evolution of loss using the cross-entropy loss function across epochs for the Swedish dataset. The red line represents the progression of the training loss, the orange line represents the evolution of the validation (development) loss.

However, the plotted predictions clearly show that the model is under-performing, as the curves do not correspond to the ones of the truth values:

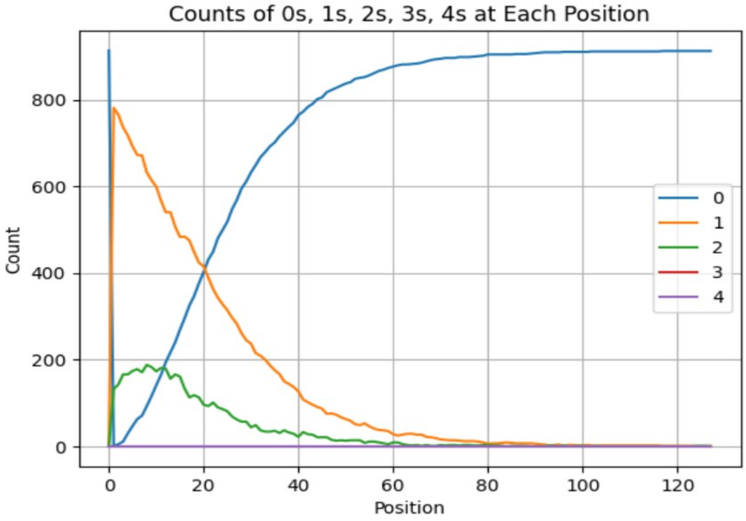


Figure 30: plotted truth values for the Swedish dataset.

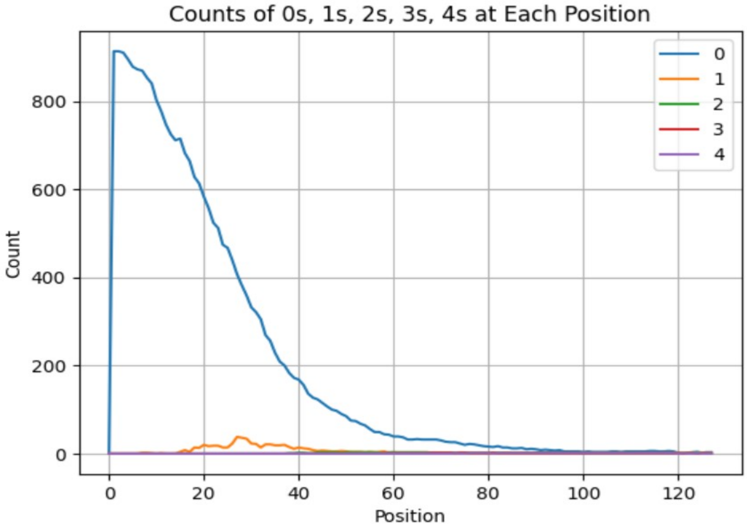


Figure 31: plotted predicted values for the Swedish dataset, using cross-entropy loss.

Interestingly, the curve of the predicted empty label "0" follows the curve of the correct truth labels "1". This shows that the model only predicts empty tokens, with a minority of correct tokens mid-sequences. In fact, the majority of the models's predictions are negative numbers, which do not appear on the present graph. Once more, this seems to indicate that the model is overwhelmed by the number of empty tokens, and only predicts 0, or negative numbers.

Future studies could attempt to improve these results in several ways. The cross-entropy loss could still be an appropriate loss function to use in the present task, but certain modifications could be necessary to make it work correctly. A possibility is to add weight to the cross-entropy loss function, so as to prevent it from being influenced by the imbalance of data. Another avenue for improvement could involve preprocessing the input tensors. We fed the model tensors of length 128 containing a prediction for each token. An alternative strategy could involve creating tensors of length 3 for each label, with each element representing the probability of a given class.

Other possibilities for improvement include training the model on a higher number of epochs. The present work trained the model on 7 epochs, so as to limit both the training time and the computational power used. Further studies could try to train the model on 10, 20 or even 100 epochs to see how the performances improve. These parameters, however, come at the cost of time and computational power.

Other hyperparameters are also worth looking into. For example, future studies could try a different learning rate and scheduler, a higher or lower dropout rate, as well as a different maximum length. In other words, paths for improvements are incredibly numerous.

4.6 Summary and conclusion

The present work makes a first attempt at using m-DeBERTa-V3 for the purpose of Grammatical Error Detection. The model was fine-tuned for the purpose of GED. The model was trained on Swedish and Italian data, the same data that was used in the MultiGED shared task.

The model's performances were poor for both languages: Recall and F0.5 were systematically lower than any other team's for both languages. Precision, on the other hand, ranked second to last. Overall, all the other language models used in the task outperform m-De-BERTa-V3 for every metrics, and in every language.

From there, several hypotheses can be made to explain such low scores. A first possibility is that the architecture of the model is not appropriate for the task. However, this is unlikely, given the fact that similar models performed exceptionally well in GED. Another hypothesis is that the model could perform much better with different optimization processes, a different loss function, or a better tuning of the hyperparameters. While exploring these possibilities is out of the scope of the present thesis, future studies could keep on exploring the use of m-DeBERTa-V3 for GED, and cast light on whether it could become the new SOTA in error detection.

5 Conclusion

The general goal of the present thesis is to contribute to improving the field of Grammatical Error Detection. To do so, we focused on two different methods: first, the analysis of the output of the state-of-the-art system for GED, and second, the testing of a new language model for the purpose of GED.

Our first contribution is to show that XLM-RoBERTa misses more errors related to lexicon, morphology and syntax than punctuation or orthography. Numerous False Negatives concern the choice of preposition, the choice of tense, as well as missing or extra tokens. A tentative conclusion to be drawn here is that XLM-RoBERTa might lack world knowledge, interpretative power and the ability to use context to correctly identify these errors. Future studies could test these hypotheses by analysing the True Positives related to lexicon, morphology and syntax in the test set.

Our second contribution is to provide a curated version of the Swedish and Italian test sets used in the MultiGED-2023 shared task. The present thesis showed that certain errors are context-dependent, and that some of the annotations that are made in the context of an essay benefit from being reviewed once the sentences are extracted from the essay. The curated test sets will be shared with the organisers of the MultiGED-2023 shared task. The updated test sets could be used in future studies on GED, and hence have a long-lasting effect on the benchmarking of GED tasks in Swedish and Italian.

Finally, our last contribution is to evaluate the model m-DeBERTa-V3 in GED by fine-tuning it for the purpose of Grammatical Error Detection. The fine-tuning of m-DeBERTa-V3 shows that despite its promising architecture, m-DeBERTa-V3 seems to underperform in GED. More precisely, the model seems to suffer from the imbalance in the data, and mostly learns to recognise correct tokens as well as empty tokens. The present thesis offers a critical analysis of the results and proposes paths for improvements: for example, modifying the tensors' shape, utilising a weighted cross-entropy loss function, or try different hyperparameters, such as a higher number of epochs. Future studies could make use of these suggestions to further improve the results of m-DeBERTa-V3 in GED.

6 References

Bell, S., Yannakoudais, H., Rei, M. 2019. Context is key: grammatical error detection with contextual word representation. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications.* 103–115.

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* Association for Computing Machinery. 610-623.

Boyd, A. 2018. Using Wikipedia Edits in LowResource Grammatical Error Correction. In: *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text.* Brussels, Belgium. Association for Computational Linguistics. 79–84.

Boyd, A., Hana, J., Nicolas, L., DetmarMeurers, Wisniewski, K., Abel, A., Karin-Schöne, Štindlová, B., Vettori, C. 2014. The MERLIN corpus: Learner Language andthe CEFR. In: *Proceedings of the Ninth Interna-tional Conference on Language Resources and Eval-uation (LREC'14).* Reykjavik,Iceland. European Language Resources Association (ELRA). 1281–1288.

Bryant, C., Felice, M., Briscoe, T. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Vancouver, Canada. Association for Computational Linguistics. 793–805.

Bryant, C., Qorib, M.R., Ng, H.T., Yuan, Z. Cao, H., Briscoe, T. 2022. Grammatical Error Correction: a survey of the State of the Art. Association for Computational Linguistics. 1-61.

Bungum, L., Gambäck, B., Næss, A.B-D. 2023. NTNU-TRH System at theMultiGED-2023 Shared Task on Multilingual Gram-matical Error Detection. In: *Proceedings of the 12th Workshop on NLP for Computer Assisted LanguageLearning (NLP4CALL).* 17-23.

Cardinaletti, A., Giuliana, G. 2020. Indefinite determiners in informal Italian: A preliminary analysis. In: *Linguistics*, 58(3): 679–712.

Clark, K., Luong, M-T., Le, Q.V., Manning, C.D. 2020. ELECTRA: Pretraining Text Encoders as Discriminators Rather Than Generators. In: *International Conference on Learning Representations.* 1-18.

Colla, D., Delsanto, M., Di Nuovo, E. 2023. EliCoDe at MultiGED2023: fine-tuning XLM-RoBERTa for multilingual grammatical error detection. In: *Proceedings of the 12th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2023).* 24-34.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics. 8440-8451.

Devlin, J., Chang, M-W., Lee, K., Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of NAACL-HLT 2019.* 4171-4186.

Guu, K., Lee,K., Tung, Z., Pasupat, P., Chang, M-W. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In: *Proceedings of the 37 th International Conference on Machine Learning.* 1-10.

He, P., Liu, X., Gao, J., Chen, W. 2021. DeBERTa: decoding-enhanced BERT with disentangled attention. Published as a conference paper at ICLR 2021. 1-23.

He, P., Gao, J., Chen, W. 2023. DeBERTa-V3: improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. Published as a conference paper at ICLR 2023. 1-16.

Kurfalı, M., Östling, R. 2023. A distantly supervised Grammatical Error Detection/Correction system for Swedish. In: *Proceedingsof the 12th Workshop on NLP for Computer AssistedLanguage Learning (NLP4CALL).* 35-39.

Kochmar, E., Briscoe, T. 2014. Detecting Learner Errors in the Choice of Content Words Using Compositional Distributional Semantics. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.* 1740-1751

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In: *Proceedings of the 20th Chinese National Conference on Computational Linguistics.* 1218-1227.

Madi, N., Al-Khalifa, H.S. Grammatical Error Checking Systems: A Review of Approaches and Emerging Directions. In: *2018 Thirteenth International Conference on Digital Information Management (ICDIM).* 142–147.

Moner, J.C. Multiclass grammatical error detection. 2022. Master's thesis in Language Technology, University of Gothenburg.

Moner, J.C., Volodina, E. 2023. Swedsh MuClaGED: A new dataset for Grammatical Error Detection in Swedish. In: *Proceedings of the 11th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2022).* Linköping Electronic Conference Proceedings 190. 36–45.

Náplava, J., Straka, M., Straková, J., Rosen, A. 2022. Czech grammar error correction with a large and diverse corpus. In: *Transactions of the Association for Computational Linguistics* 10, 452–467.

Ngo, T.Q., Nguyen, T.M.H., Le-Hong, P. 2023. Two neural models for multilingual grammatical error detection. In: *Proceedings of the 12th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2023).* 40-44.

Nikolaev, A., Ashaie, S., Hallikainen, M., Hänninen, T., Higby, E., Hyun, J., Lehtonen, MiM.nna, Soininen, H. 2019. Effects of morphological family on word recognition in normal aging, mild cognitive impairment, and Alzheimer's disease. Cortex, 116. 91-103.

Pires, T., Schlinger, E., Garrette, D. 2019. How Multilingual is Multilingual BERT? In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019).*4996-5001.

Rastall, P. 1994. The prepositional flux. In: *Notes and discussions* 32(3), 229-242.

Rei, M., Yannakoudakis, H. 2016. Compositional Sequence Labeling Models for Error Detection in Learner Writing. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).Berlin, Germany. Association for Computational Linguistics.* 1181–1191.

Rudebeck L., Sundberg, G. 2021. SweLL correction annotation guidelines. Forskningsrapporter från institutionen för svenska språket, Göteborgs universitet. 1-78.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017).* 1-15.

Vinogradova O., Lyashevskaya, O. 2022. Re-view Of Practices Of Collecting And AnnotatingTexts In The Learner Corpus REALEC. In: *Text,Speech, and Dialogue: 25th International Confer-ence, TSD 2022.* Berlin, Heidelberg. Springer-Verlag. 77-88.

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L.,Schenström, C.-J., Sundberg, G., Wirén, M. 2019. The SweLL Language Learner Corpus:From Design to Annotation. In: *Northern European Journal of Language Technology.* 67-104.

Volodina, E., Bryant, C., Caines, A., De Clercq, O., Frey, J-C., Ershova, E., Rosen, A., Vinogradova, O. 2023. MultiGED-2023 shared task at NLP4CALL: Multilingual Grammatical Error Detection. In: *Proceedings of 12th workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2023).* Linköping Electronic Conference Proceedings 197. NEALT Proceedings Series 53.

Yang, X., Bian, J., Hogan, W.R., Wu, Y. 2020. Clinical concept extraction using transformers. In: *Journal of the American Medical Informatics Association.* 1935-1942.

Yannakoudakis, H., Briscoe, T., Medlock, B.2011. A New Dataset and Method for Automati-cally Grading ESOL Texts. Inö *Proceedings of the49th Annual Meeting of the Association for Com-putational Linguistics: Human Language Technolo-gies.* Portland, Oregon, USA. Association for Computational Linguistics. 180–189.

Yuan, Z., Bryant, C. Taslimipoor, S., Davis, C. 2021. Multi-Class Grammatical Error Detection for Correction: A Tale of Two Systems. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, November 2021.* 8722-8736.

A Appendix: annotation framework

To facilitate reading, the glosses have been simplified, and only the relevant token(s) have been corrected.

A.1 Punctuation

Subcategories:

- Missing comma: a comma missing in front of the current token. A comma should be added. (A)

(22) a. **Jag inser att kläder är viktiga för att det är hur man*
I realise that clothes are important for that it is how one
presenterar sig själv för andra ***[missing comma]*** *men jag har*
presents them selves for others **[missing comma]** but I have
också insett att kläder kan vara en problem [...]
also realised that clothes can be a problem

"I realise that clothes are important because they are how one represents themselves in front of other but I have also realised that clothes can be a problem [...]"

b. *Jag inser att kläder är viktiga för att det är hur man presenterar*
I realise that clothes are important for that it is how one presents
sig själv för andra ***,*** *men jag har också insett att kläder kan*
them selves for others , but I have also realised that clothes can
vara en problem [...]
be a problem

"I realise that clothes are important because they are how one represents themselves in front of other, but I have also realised that clothes can be a problem [...]"

- Extra comma: the comma should be deleted. (D)

(23) a. **Stud-io tre lingue* ***,*** *alla scuola [...]*
Study-1SG three languages , at school

"I study three languages , at school [...]"

b. *Studio tre lingue alla scuola [...]*
Study-1SG three languages at school

"I study three languages at school [...]"

- Wrong punctuation: the learner used the wrong punctuation mark. The current punctuation mark should be replaced by another punctuation mark. (R)

(24) a. **Omsorg ingår hur kan ta sina barn till skolan* ***?***
Care include how can take their children to school ?

"Care includes how one takes their children to school ?"

b. *Omsorg ingår hur kan ta sina barn till skolan* ***.***
Care include how can take their children to school .

"Care includes how one takes their children to school."

- Missing punctuation: a punctuation mark (excl. commas) is missing in front of the current token and should be added. (A)

(25) a. **Finalmente* ***[miss. punct.]*** *cred-o che l' insegnante sia*
Finally **[missing punctuation]** think-1SG that the teacher is
nel centro [...]
in.the center

"Finally I think that the teacher is in the center [...]"

b. *Finalmente* ***,*** *cred-o che l' insegnante sia nel centro [...]*
Finally , think-1SG that the teacher is in.the center

"Finally, I think that the teacher is in the center [...]"

- Extra punctuation: the punctuation mark should not be used in this context and should be deleted. (D)

(26) a. **När man pratar om klädelsen* ***,*** *kan man se den som en helhet .*
When one speaks of clothing , can one see it like a whole .

"When one speaks of clothing, one can see it as a whole."

b. *När man pratar om klädelsen kan man se den som en helhet .*
When one speaks of clothing can one see it like a whole .

"When one speaks of clothing one can see it as a whole."

- Punctuation/conjunction: the punctuation mark should be replaced by a conjunction. (R)

(27) a. **[C]i sono degli francesi , degli inglesi* ***,*** *anché dei*
There are DET French.people , DET English.people , also DET
Tedeschi .
German.people .

"There are French people, English people, also German people."

b. *Ci sono degli francesi , degli inglesi* ***e*** *anché dei*
There are DET French.people , DET English.people and also DET
Tedeschi .
German.people .

"There are French people, English people and also German people."

- Conjunction/punctuation (Swedish only): the conjunction should be replaced by a punctuation mark. (R)

(28) a. **När läkare , ekonomer* ***och*** *IT-tekniker och professorer*
When doctors , economists and IT.technicians and professors
sammanstrålar [...]
come.together

"When doctors, economists and IT technicians and professors come together [...]"

b. *När läkare , ekonomer* ***,*** *IT-tekniker och professorer sammanstrålar*
When doctors , economists , IT.technicians and professors come.together
[...]

"When doctors, economists, IT technicians and professors come together [...]"

A.2 Orthography

Subcategories:

- Spelling: the token contains a spelling mistake. (R)

(29) a. ****Arrividerci*** *e fino alla prossima volta !*
Goodbye and until DET next time !

"Goodbye and see you next time!"

b. ***Arrivederci*** *e fino alla prossima volta !*
Goodbye and until DET next time !

"Goodbye and see you next time!"

- Word concatenation: the token is concatenated with another token, yet should not be concatenated. Inversely: the token should be concatenated with another token, yet is not concatenated. (R)

(30) a. ****Sta mattina*** *siamo partiti a nuotare [...]*
This morning are gone to swim

"This morning, we went swimming [...]"

b. ***Stamattina*** *siamo partiti a nuotare [...]*
This.morning are gone to swim

"This morning, we went swimming [...]"

- Capitalisation: the first letter of the word should be capitalized, or lowered. (R)

(31) a. **_*det_** _är olika ._
It is different .

"it is different."

b. **_Det_** _är olika ._
It is different .

"It is different"

A.3 Lexicon

Subcategories:

- Adjective: the wrong adjective is chosen for the given context. (R)

(32) a. _*[S]o_ **_många_** _folk passera och gå upp trapperna ._
So much people pass and go up stairs:DET .

'[S]o a much people pass and go up the stairs.'

b. [S]_o_ **_mycket_** _folk passera och gå upp trapperna ._
So many people pass and go up stairs:DET .

"[S]o many people pass and go up the stairs."

- Adverb: the wrong adverb is chosen for the given context. (R)

(33) a. _*[N]ågra år_ **_sedan_** _började mobiltelefoner att användas ._
Few years after start:PST phones to be.used .

"A few years after, mobile phones starte to be used."

b. [N]_ågra år_ **_senare_** _började mobiltelefoner att användas ._
Few years later start:PST phones to be.used .

"A few years later, mobile phones starte to be used."

- Conjunction: the wrong conjunction is chosen for the given context. (R)

(34) a. _*Globalisering kan både vara ett hot **som** en fördel [...]_
Globalisation can both be a danger like an advantage

"Globalisation can be a danger like an advantage"

b. _Globalisering kan både vara ett hot **och** en fördel [...]_
Globalisation can both be a danger and an advantage

"Globalisation can be a danger and an advantage [...]"

- Noun: the wrong noun is chosen for the given context. (R)

(35) a. _*Forse uno degli amici virtuali è un **gente** molto pericoloso [...]_
Maybe one of.the friends virtual is a people very dangerous

"Maybe one of the virtual friends is a dangerous people."

b. _Forse uno degli amici virtuali è un **persona** molto pericoloso [...]_
Maybe one of.the friends virtual is a person very dangerous

"Maybe one of the virtual friends is a dangerous people."

- Preposition: the wrong preposition is chosen for the given context. (R)

(36) a. _*[V]i auguro del cuore per il vostro matrimonio **il** novembre ._
We wish from hear for DET your wedding DET november .

"We wish you the best for your wedding in November."

b. [V]_i auguro del cuore per il vostro matrimonio **a** novembre ._
We wish from heart for DET your wedding DET november .

"We wish you the best for your wedding in November."

- Pronoun: the wrong pronoun is chosen for the given context. (R)

(37) a. **Hör av* ***mig*** *snart .*
Hear of me soon .

"Hear from me soon"

b. *Hör av* ***dig*** *snart .*
Hear of you soon .

'Hear from you soon.'

- Verb: the wrong verb is chosen for the given context. (R)

(38) a. **Jag tror att religion* ***har*** *ingen roll [...]*
I think that religion has no role

"I think that religion has no role [...]"

b. *Jag tror att religion* ***spelar*** *ingen roll [...]*
I think that religion plays no role

"I think that religion plays no role [...]"

- Determiner (Swedish only): the wrong determiner is chosen for the given context. (R)

(39) a. **Paris har en flygplats , men* ***det*** *är också ganska lång från stan .*
Paris has an airport , but it is also quite far from city:DET .

"Paris has an airport, but it is quite far from the city."

b. *Paris har en flygplats , men* ***den*** *är också ganska lång från stan .*
Paris has an airport , but it is also quite far from city:DET .

"Paris has an airport, but it is quite far from the city."

- Auxiliary (Italian only): the wrong auxiliary is chosen for the given context. (R)

(40) a. *****[S]on****o gia lavorata per una azienda di moda .*
Be:1SG already worked for a company of fashion .

"I have already worked for a fashion company (intended)"

b. ***Ho*** *gia lavorata per una azienda di moda .*
Have:1SG already worked for a company of fashion .

"I have already worked for a fashion company."

- Expression: the wrong expression (ie. group of words) is chosen for the given context. (R)

(41) a. **Pens-o che ogni giorno parl-iamo di bene in meglio .*
Think-1SG that every day speak-1PL of good in better .

"I think that today, we speak better and better. (intended)"

b. *Pens-o che ogni giorno parl-iamo sempre più bene.*
Think-1SG that every day speak-1PL always more good.

"I think that today, we speak better and better."

A.4 Morphology

Subcategories:

- Wrong choice of definite / indefinite. (R)

→All tokens that should have been definite, but are indefinite.

(42) a. **Att en säkning av* ***ålder*** *skulle legitimera bra [...]*
Hat a securisation of **age** would legitimate good

"That a securisation of age would legitimate good [...]"

b. *Att en säkning av* ***ålder-n*** *skulle legitimera bra [...]*
That a securisation of age-DET would legitimate good

"That a securisation of the age would legitimate good [...]"

→All tokens that should have been definite, but are indefinite.

(42) a. _*De som sitter i **rullstol-en** [...]_
Those who sit in wheelchair-DET

"Those who sit in the wheelchair [...]"

b. _De som sitter i **en** rullstol [...]_
Those who sit in DET wheelchair

"Those who sit in a wheelchair [...]"

- Tense choice:

(43) a. _*Allan var ledsen och vill inte **bor** i äldrebående [...]_
Allan was sad and wants not live:PRES in retirement.home

"Allan was sad and does not want to lives in a retirement home [...]"

b. _Allan var ledsen och vill inte **bo** i äldrebående [...]_
Allan was sad and wants not live in retirement.home

"Allan was sad and does not want to live in a retirement home [...]"

- Plural morphology. (R)

→ the plural morphology should have been used, but singular morphology is used instead:

(44) a. _*De ska följa samma **regel** [...]_
They shall follow same rule

"They shall follow the same rule"

b. _De ska följa samma **regl-er** [...] ._
They shall follow same rule-s

"They shall follow the same rules."

→ error in the plural morpheme

(45) a. **[F]lera andra* ***plattform-er*** *.*
Several other platform-s .

"Several other platforms."

b. [F]*lera andra* ***plattform-ar*** *.*
Several other platforms-s .

"Several other platforms."

- Singular morphology. (R)
→ the singular morphology should have been used, but the plural morphology is used instead

(46) a. **hur är det med* ***din-a*** *familjan ?*
How is it with your-PL family

"How is it with your family?"

b. *hur är det med* **din** *familjan ?*
How is it with your family

"How is it with your family?"

- Case (R)

→ Forgotten case marking:

(47) a. ****mammorna*** *rörelse .*
mothers concern .

"The mothers concern."

b. ***mammornas*** *rörelse .*
mothers:GEN concern .

"The mothers' concern."

→ Incorrect case used:

(48) a. **Är han inte kär i* ***hennes*** *längre .*
Is he not in.love in her:GEN anymore .

"He is not in love with her anymore."

b. *Är han inte kär i* ***henne*** *längre .*
Is he not in.love in her:ACC anymore .

"He is not in love with her anymore."

- Gender morphology: the gender morphology is incorrect or lacking. (R)

(49) a. **Questa settimana era* ***fantastic-o*** *!*
This week was fantastic-MASC !

"This week was fantastic !"

b. *Questa settimana era* ***fantastic-a*** *!*
This week was fantastic-FEM !

"This week was fantastic !"

- Contraction (Italian only): the word should have been contracted following morphophonological rules. (R)

(50) a. **Era* ***una*** *esperienza molto interessante .*
Was an experience very interesting .

"It was a very interesting experience."

b. *Era* ***un'*** *esperienza molto interessante .*
Was an experience very interessting .

"It was a very interesting experience."

- Determiner morphology (Italian only): the morphology on the determiner is incorrect. (R)

(51) a. **Se hai tempo e* ***un*** *spazio per noi per due notte [...]*
If have time and a space for us for two nights

"If you have time and space for us for two nights [...]"

b. *Se hai tempo e* ***uno*** *spazio per noi per due notte [...]*
If have time and a space for us for two nights

"If you have time and space for us for two nights [...]"

- Prepositional morphology (Italian only): incorrect prepositional morphology. (R)

(52) a. **[L]' annucio pubblicato che ho letto* ***su*** *vostro sito Internet .*
The announcement published that have read on your site internet .

"The published announcement that I read on your website [...]"

b. [L]*' annucio pubblicato che ho letto* ***sul*** *vostro sito Internet .*
The announcement published that have read on:DET your site internet .

"The published announcement that I read on your website [...]"

- Verbal agreement (Italian only): the verb does not agree with the subject. (R)

(53) a. **Mi* ***chiam-a*** *Maria Michele .*
1SG call-3SG Maria Michele .

"Me calls Maria Michele. (literal translation)"

b. *Mi* ***chiam-o*** *Maria Michele .*
1SG call-1SG Maria Michele .

"I call myself Maria Michele (i.e. my name is Maria Michele)."

- Mistake in the definite morpheme (Swedish only): the definite morpheme contains an error. (R)

(54) a. **Hoppas att du kan klara dig bra med* ***intervju-en*** *[...]*
Hope that you can cope you good with interview-DET

"I hope you can manage the interview well."

b. *Hoppas att du kan klara dig bra med* ***intervju-n*** *[...]*
Hope that you can cope you good with interview-DET

"I hope you can manage the interview well."

- Definite adjectival morphology (Swedish only): incorrect definite adjectival morphology[5]. (R)

(55) a. **Dessa ändringar kan i sin tur påverka regeringsnivå och* ***politisk***
These changes can in its turn influence government.level and political
situation [...]
situation [...]

'These changes can then influence government level and political situation.'

b. *Dessa ändringar kan i sin tur påverka regeringsnivå och*
These changes can in its turn influence government.level and
politisk-a *situation [...]*
political-DEF situation .

"These changes can then influence government level and political situation."

- prep morphology (Italian only): misformed prepositional morphology

(56) a. **[P]oss-o andare* ***a*** *voi matrimonio .*
Can-1SG go a your wedding .

"I can go to your wedding."

b. [P]*oss-o andare* ***a-l*** *voi matrimonio .*
Can-1SG go a-DET your wedding .

"I can go to your wedding."

[5] There is no indefinite adjectival morphology.

- Verbal morphology (Swedish only): tense is misformed. (R)

(57) a. **Fatima rigt till hennes mamma och berättade hur Sofia må och vad*
Fatima calls to her mom and told how Sofia is and what
som ***händt*** *.*
that happened

"Fatima calls her mom and told how Sofia is doing and what happened"

b. *Fatima rigt till hennes mamma och berättade hur Sofia må och vad*
Fatima calls to her mom and told how Sofia is and what
som ***hände*** *.*
that happened

"Fatima calls her mom and told how Sofia is doing and what happened"

A.5 Syntax

Subcategories:

- Missing word: a word is missing in front of the current token. (A)
Divided into further subcategories corresponding to the POS of the missing word: adverb, conjunction, determiner, noun, preposition, pronoun, verb.

(58) a. **Hoppas att du mår bra och* ***[missing word]*** *allt blir som*
Hope that you do well and [missing word] everything becomes like
du vill .
you want

"I hope you are doing well and everything will be as you want. (litteral)"

b. *Hoppas att du mår bra och* **att** *allt blir som du vill .*
Hope that you do well and that everything becomes like you want

"I hope that you are doing well and that everything will be as you want."

- Extra word: the current token should be deleted. (D)
Divided into further subcategories corresponding to the POS of the extra word: abbreviation, adjective, adverb, conjunction, determiner, noun, preposition, pronoun, verb.

(59) a. **Miei genitori sono stati li* ***un*** *mezzo anno fa [...]*
My:PL parents are stayed there a half year ago

"My parents were here half a year ago [...]"

b. *Miei genitori sono stati li mezzo anno fa [...]*
My:PL parents are stayed there half year ago

"My parents were here half a year ago [...]"

- Word order: the token is at the wrong position in the sentence. (R)

(60) a. **Där* ***Riad nämner*** *Mirja Saaris [...]*
There Riad names Mirja Saaris

"There, Riad names Mirja Saaris [...]"

b. *Där* ***nämner Riad*** *Mirja Saaris [...]*
There names Riad Mirja Saaris

"There, Riad names Mirja Saaris [...]"

- Formulation: the group of tokens should be reformulated. Usually entails several ADR operations. (ADR)

(61) a. **[V]i auguro* ***del cuore*** *per il vostro matrimonio* ***il*** *novembre .*
We wish from hear for DET your wedding in november .

"[W]e wish you from the heart for your wedding in November."

b. [V]*i auguro* ***dal profondo del cuore*** *per il vostro matrimonio a novembre .*
"[W]e wish from bottom of heart for DET your wedding in november .

"We wish you the best for your wedding in November from the bottom of our hearts."

B Appendix: cross-referencing the SweLL taxonomy

SweLL taxonomy (Volodina et al. 2019)	Current taxonomy	Comments
O	O – spelling (R)	-
O-Cap	O – capitalization (R)	-
O-Comp	O – word concatenation (R)	-
P-M	P – missing punctuation (A)	-
P-R	P – extra punctuation (D)	-
P-W	P – wrong punctuation (R)	-
P-Sent	-	This subcategory is not present in the test data that we analysed.
L-Der	L - POS (R)	-
L-FL	-	This subcategory is not present in the test data that we analysed.
L-Ref	L - POS (R)	-
L-W	L - POS (R)	-
M-Adj/adv	-	This subcategory is not present in the test data that we analysed.
M-Case	M – case (R)	-
M-Def	M – wrong choice of definite/indefinite (R) M - definite adjectival morphology (R) M - mistake in the definite morpheme (R)	-
M-F	M – any subcategory	Depending on the token, the SweLL subcategory can correspond to different morphological subcategories of the present taxonomy.
M-Gend	M – gender morphology (R)	-

M-Num	M – singular morphology (R) M – plural morphology (R)	-
M-Other	M – any subcategory	Depending on the token, the SweLL subcategory can correspond to different morphological subcategories of the present taxonomy.
M-Verb	M – verbal agreement (R) M – tense choice (R) M – verbal morphology (R)	-
S-Adv	S – word order (R)	-
S-Comp	S – formulation (ADR)	-
S-Clause	S – formulation (ADR)	-
S-Ext	S – formulation (ADR)	-
S-FinV	S – word order (R)	-
S-M	S – missing POS (A)	-
S-MSubj	S – missing POS (A)	-
S-Other	S – any subcategory	Depending on the token, the SweLL subcategory can correspond to different subcategories of the present taxonomy.
S-R	S – extra POS (D)	-
S-Type	S – formulation (ADR) L – POS (R)	-
S-WO	S – word order (R)	-
C	Any subcategory	Depending on the token, the SweLL subcategory can correspond to different subcategories of the present taxonomy.
Cit-FL	-	This subcategory is not present in the test data that we analysed.
Com!	-	This subcategory is not present in the test data that we analysed.
OBS!	-	This subcategory is not present in the test data that we analysed.

X	Any subcategory	Depending on the token, the SweLL subcategory can correspond to different subcategories of the present taxonomy.

Table 18: Correspondence between SweLL (Volodina et al. 2019) and the current taxonomy