unibz
Freie Universität Bozen
Libera Università di Bolzano
Università Liedia de Bulsan

FREE UNIVERSITY OF BOZEN-BOLZANO

FACULTY OF EDUCATION

Master in Applied Linguistics
Automatic Language Analysis

AEFLL: Automatic Evaluation of Foreign Language Learning

An interdisciplinary project for English, Italian and German at the Free University of Bozen-Bolzano

Supervisor

Dr. Brutti Alessio

Co-Supervisor

Prof. Vietti Alessandro

Submitted by

Schmalz Veronica Juliana

Keywords: Automatic Language Assessment, CEFR scoring, English, German, Italian, L2, Second Language acquisition

Summer session

Academic year: 2020-2021

Abstract

The aim of this thesis project is to investigate neural architectures based on BERT models to automatically assess the competences of adult language learners. Given the exponential growth of the latter worldwide and the increasing adoption of computer-assisted language examinations, these automated systems could facilitate the objective scrutiny of numerous tests, reducing the biases of human evaluators while providing cross-validly efficient and detailed assessments. Our research addresses the three languages officially employed at the Free University of Bozen-Bolzano, namely English, German and Italian. We combine analysis and assessment methods within machine learning, natural language processing, language acquisition and development to correct and classify both written and oral examinations of adult language learners following the principles of the Common European Framework of Reference.

For our analysis we use written open-source datasets of English proficiency tests, which are more numerous and accessible, of Italian and German, which are less numerous and available. To train our models, we conduct different experiments alternatively using the original written texts from the students, human corrections, when available, and the automatic corrections provided by LanguageTool, a computerized language checker tool. In this way the BERT model provides an embedding representation that not only describes the text content but at times also accounts for rule violations and other errors. A multi-layer perceptron is then used to map the embedding text representation into the related CEFR levels. We evaluate the performance of our architecture on each dataset training a language specific model, achieving extremely high proficiency prediction in all cases.

In addition, we received a narrow dataset of oral examinations for B2 English exams from the University Language Centre on which we conduct a separate case study. We applied the pretrained English models on the oral exams, previously transcribed using an automatic speech recognition engine.

Finally, we consider linguistic aspects related to the written and spoken language of the learners of different languages and possible features to be added to the models to possibly improve their performance.

Acknowledgements

I would like to express my gratitude to the people who contributed and supported me carrying out this thesis project. First, I would like to sincerely thank my supervisor, Dr. Alessio Brutti, for his expert guidance and advice during the research and development process of this dissertation. Along with him, I would like to extend thanks to the members of the SpeechTek Research Unit of the Fondazione Bruno Kessler in Povo, Trento. In particular, I would like to thank Dr. Marco Matassoni and Stefano Bannò for their assistance in the project.

Second, with regard to the data collection of the Free University of Bozen-Bolzano, I would like to express my sincere appreciation to my co-supervisor, Prof. Alessandro Vietti, and the Head of the Testing Unit of the Language Centre, Chistoph Nickenig. I would further like to thank not only the former for his critical contribution with regard to the analysis of the linguistic aspects, but especially Dr. Luca Ducceschi for his assistance and support.

Finally, I would like to thank my family, especially my parents, for the sacrifices they have made and continue to make in order to provide me with a high-level education and allow me to pursue my career goals.

Table of contents

List of tables

List of Figures

CHAPTER ONE: INTRODUCTION

1. Introduction

In recent years, the numbers of foreign language learners in the world and in Europe have grown exponentially. Indeed, more and more children and adults decide to engage in the learning and acquisition of a new language. According to the data released by Eurostat (2020), the majority of children in Europe from Primary to Secondary school learn English, French and German. From upper secondary general education onwards, students also explore other languages such as Spanish and Italian. Thus, the percentages of language learners appear relatively high among young and older learners, ranging between 70% and 100% among the students of most member states according to Eurostat statistics (2020). As a consequence, in the field of education, schools and institutions dealing with language learning have had the need to improve and create new teaching materials suitable for the effective acquisition of a language other than their first one (L1), as well as to design and make use of valid and objective assessment methods for a large number of proficiency tests.

Especially given that Europe is a fertile location for language learning, the EU soon recognised the requirement for a common framework of material for language learning and assessment (Council of Europe 2001). In 2001, the Common European Framework (CEFR) was established by the European Council with the aim of collaborating internationally on the development of teaching and assessment methods for modern languages and promoting plurilingualism. This system was intended to be a context-free independent framework based on theories of language acquisition and competence while remaining clear and user-friendly. It consists of a six-level scale, namely A1, A2, B1, B2, C1 and C2 in ascending order, each characterised by relevant and distinct orientation points (see Fig.2 Appendix A). To date, this scale is still used across Europe and beyond to measure language proficiency in several languages. Moreover, it represents a system applied to evaluate both students' skills and teachers' methods, apart from constituting an orientation structure for the world of linguistic education and related fields.

Since the establishment of the CEFR, the challenge has been to create language testing systems that are both objective in their assessment and comparable to other languages evaluated on the same scale. In addition, given the constant increase in the number of learners of all ages, methods were sought that could be used for universal large-scale tests. In an attempt to meet these demands, numerous researchers dealing with Second Language Acquisition (SLA) have begun to design and study systems for the automatic evaluation of linguistic proficiency. This topic has long been tackled by a variety of scholars, who in turn focused on the technical aspects of the matter, trying to design efficient systems for corrections and scoring (Yannakoudakis et al. 2011, Chapelle & Voss 2008), on the cognitive aspects and the mental mechanisms of language learning and processing (Jang 2017), but also on the linguistic aspects, regarding errors or lexicon (Richards 2015, Saito 2017), as well as the sociolinguistic ones about the contact between first languages and second languages and the interaction with other speakers (Cohen 2016). However, especially in recent times, thanks to technological advances in Natural Language Processing (NLP), machine learning (ML), AI, corpus linguistics, computer-assisted language testing (CALT) and learning (CALL), automatic language assessment has gained momentum.

Despite the fact that automatic language assessment has been the focus of several cross-disciplinary studies, still few data and results are available in its regards. For example, there have been some attempts to assess both spoken and written language of English Language Learners (ELL). In either case, the competence of human learners has been usually measured by means of a classifier adopting a standard scale (cf. Bernstein et al. 1990, Cucchiarini et al. 2002, Yannakoudakis et al. 2011). Nonetheless, challenges still exist regarding not only the choice of methods and principles to adopt but also the absence of publicly available data suitable for this task.

In order to assess the language learners' competences several aspects related to cognition, interpretation and observation must be considered (Pellegrino 2001), besides the performance in the assessment tests and the obtained scores. Indeed, these elements represent how the learners not only perform a task during the assessment but also how they develop competence and how they process language during the learning process. For this purpose technological tools have proved particularly effective. Unlike human evaluators, who often tend to make holistic judgements and do not manage to extemporaneously assess different aspects of proficiency, noting errors and suggestions with precision, automated systems have the ability to combine technological efficiency with precision and detailed analysis (Chapelle & Voss 2016). In order to be accurate, however, tests need to be modelled to assess specific linguistic features which can be systematically measured. This can be improved by means of learner corpora with data from different groups of students and evidence on empirical and systematic language analysis. Unfortunately, however, few resources are freely available online, especially those containing detailed annotations of the different aspects to be considered.

Moreover, recent advances in spoken and text NLP have exploited Error Analysis (EA) to create test correction strategies capable of not only detecting different error types but also providing detailed personalised feedback (Naber 2003, Miłkowski 2010), not always offered to language learners by human evaluators. Thanks to improvements in the fields of machine learning, AI, applied linguistics and corpus linguistics, taking advantage from the EA theories for language assessment has been possible. Indeed, according to the latter, errors of different nature relate both to the speakers' competence and to their acquisition considering their native language and the system of mental rules they create.

Therefore, on the one hand, the automatic assessment tools' scoring indicators must be somewhat related to those followed by language experts, yet they must also be applicable to an automatic scoring system. On the other hand, they must take into account the different aspects encompassed by language, especially those related to the characteristics of foreign language learners and their L1s, which are at times not so evident as to be automatically detected. For this reason, the type of data these systems are designed with is also crucially important.

In this project, we explore an innovative method for the automatic language assessment of three languages, namely English, Italian and German. We do this using learner corpora relatively to written and oral language tests carried out by adults and university students with different native languages. We compare the results obtained with state-of-the-art automatic models, as well as with the tests evaluations according to the CEFR principles made by human language experts. In order to attempt to fully automate the process of language assessment, we also employ a language checker to detect and correct errors in the original students' texts. Ultimately, we consider the errors found in the different levels and languages to look for possible correlations between their typologies across the various levels, as well as languages.

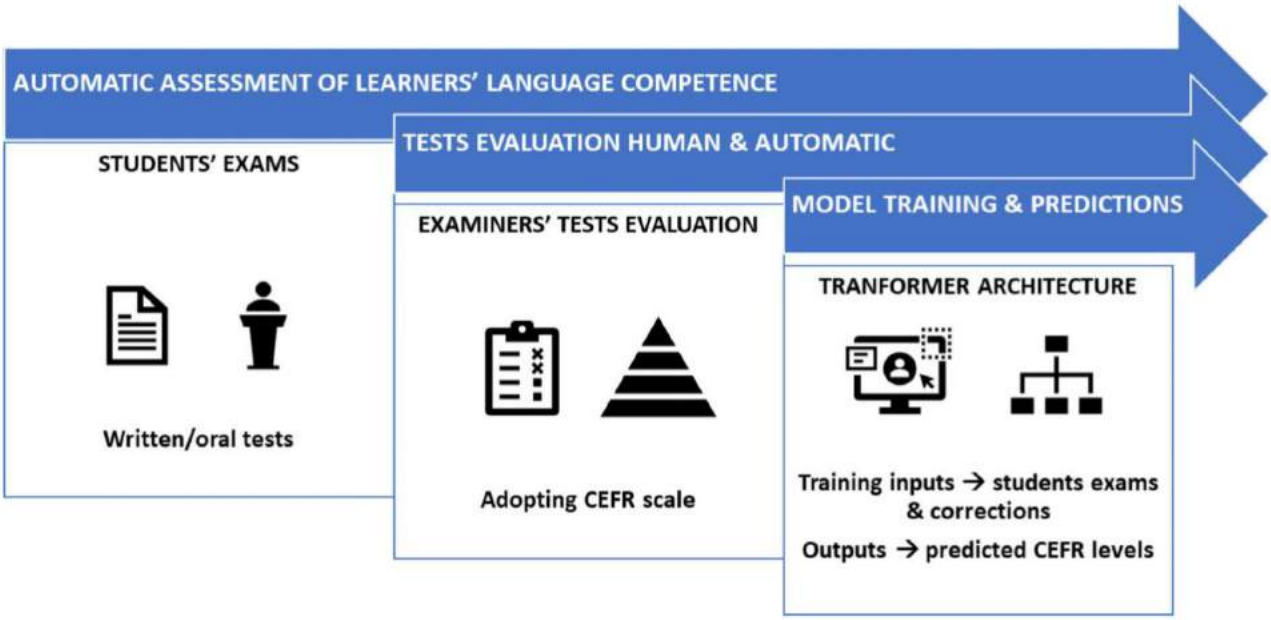


Figure 1: Pipeline of our project to automatically assess students' competences.[1]

More specifically, in the next section, namely chapter two, we present the related works and the problems addressed in the current project. In chapter three, we describe the data used for the automatic assessment of written texts in English, German and Italian. In chapter four, we introduce the methodology adopted, together with the tools for the automatic correction and the model for the classification of the exams. In chapter five, we provide a description of the different experiments carried out, while the subsequent sections report on the results obtained. Chapter six is devoted to a previous analysis of the errors found and numerically measurable linguistic features. Chapter seven, on the other hand, is dedicated to a small case study we carried out on oral examinations for English language from the Free University of Bozen-Bolzano. In the last chapter, namely chapter eight, we discuss the results obtained and compare the written and oral modalities as well as the different analysed languages.

[1] The pipeline in Figure 1 shows the procedures adopted in our project to automatically assess students' exams according to the CEFR by employing a neural architecture. First, the students' oral or written tests are received and corrected by human assessors. The original texts or those transcribed via ASR are provided to the system together with automatically generated LanguageTool corrections. The neural model classifies the texts by level, and in some experiments even attempts to evaluate them with a score.

CHAPTER TWO: LITERATURE REVIEW

2.1. Research Background

Automatic second language (L2) assessment concerns the development of tools and methods for the evaluation of learners of different ages, genders, languages of origin and skills. The principal aim is the creation of effective, unbiased and cross-linguistically valid systems that can both simplify assessment and render it objective. However, achieving such results represents a complex task that researchers have been addressing for years while experimenting with several methodologies and techniques.

From the beginning, the evaluation of L2 proficiency appeared to be deeply connected to the evolution of corpus linguistics and automated texts scoring. Indeed, in order to design automatic tools for this purpose, both a broad range of learner data and automatically applicable assessment techniques are required. In the 1960s, early research in this area focused on the contrastive aspects of language learning. According to various scholars (Lado 1961, Sheen 1996 & James 2005), the key components for measuring competence in a foreign language would be the differences between it and the native language. In this perspective, learners' errors could be traced back to their first language, or L1, and therefore be caused by the clash of the two language systems (Lado 1961). In this respect, the interlanguage (IL) emerged while establishing intermediate elements between the target language and that of the learner, has been since then the focus of a considerable deal of research in the field of second language acquisition (Selinker 1972, Bialystok & Sharwood Smith 1985, Taguchi 2017, Hao et al. 2021). According to the latter, indeed, assessing and analysing language proficiency regards also the consideration and quantification of the distance between a well-defined standard target language to be acquired and that of the learner's variety, less marked and defined. In the 1970s, nonetheless, more attention was paid to the typology of errors made by language learners. In particular, researchers investigated the distinction between *mistake* and *error*, establishing that *errors* mainly concerned competence as a whole, independently from the unique learner characteristics, while *mistakes* involved the performance of a specific linguistic act of the language learner (Corder 1967). However, further studies (Richards 1971) soon revealed that learners' errors do not only result from the influence of their L1, but may also be determined by the following additional factors:

- *Overgeneralization*, when the leaners create a biased language structure that does not apply to the target L2;
- *Ignorance of rule restriction,* when the learners do not observe L2 sets of rules ignoring or failing to recognize them;
- *Incomplete rule application,* when the learners have not fully acquired or developed rules knowledge*;*
- *False concept hypothesis, when* the learners have not clearly understood concepts in the L2.

Given the above, errors could be used as methods to identify challenging areas of language learning and acquisition, to improve language teaching strategies, find insights common to different language learners and improve the development of better learning and testing materials. In the following years, Dell Hymes started to develop the theory of communicative competence, according to which competence cannot be simply translated into

its linguistic component, since it broadly encompasses the abilities of comprehension and appropriateness of language use (Hymes 1972). Following these principles, in the 1980s a new framework for the evaluation of L2 learning and teaching competences began to emerge (Canale and Swain 1980). It combined four distinct language users' sub-competences to assess the communicative competences introduced the years before, mainly:

- *Grammatical competence*, or the ability to build and utter grammatically correct linguistic constructs;
- *Sociolinguistic competence*, or the ability to generate utterances fitting the sociolinguistic context in which they are to be used;
- *Discourse competence*, or the ability to construct cohesive and coherent discussions;
- *Strategic competence*, or the ability to plan communicative utterances and resolve related issues.

These perspectives appear to be at variance to some extent with Chomsky's dual distinction between *competence* and *performance* (Chomsky 1965). According to the latter, a differentiation is drawn between the knowledge of abstract linguistic structures, and the practical realization of these in the context of the communicative act. However, in the field of education and assessment of foreign language learning, it is often difficult to separate these two aspects. In fact, in order to establish the successful acquisition of a language other than the native one, having learned its abstract system of rules and principles on their own is insufficient since one must also and above all be able to practically use it to communicate. Therefore, from the 1990s onwards, researchers and educators gradually shifted their focus from assessing the formal correctness to a set of linguistic, cognitive and communicative skills acquired by the foreign language learner. For example, Bachman and Palmer wrote a guide for the design of language tests which allowed to assess the competences of ELLs taking into account the necessary correspondence between language use and the learner's evaluated qualities, clearly and precisely defined by the examiners (Bachman & Palmer 1990). Moreover, in 2000 new outlined systems for structuring and modelling the language classroom interaction emerged (Lee 2000). According to these, teaching activities for language learning need to be focused on real communication to encourage both information exchange and vocabulary and grammatical development. The adoption of these techniques compared to the typical theoretical lesson would allow learners to gradually strengthen their skills by bringing them closer to the concrete experience rather than abstracting them into a distant dimension. During the same period, the European Union introduced the CEFR to monitor learners' language skills according to common guidelines valid in all the Member states. This framework was developed to map the competences and sub-competences in various languages spoken in the EU, based on shared principles and skills to be acquired to attest a level among the six predetermined ones, namely A1, A2, B1, B2, C1 and C2 (cf. Council of Europe 2001).

In the meantime, technology and computational power were expanding as well. Their development allowed the promotion of the first automated language proficiency scoring systems. Page's *Project Essay Grade* (1968) was among the first capable of evaluating written essays after they had been manually entered into a computer. This was followed by increasingly advanced systems, such as the *e-rater*® (Burstein 2003) and the *Intelligent Essay Assessor* (Landauer 2003). The former exploits Part-of-Speech (PoS) tagging to grade students' essays and provide feedbacks for improvement. The latter, namely IEA, instead, is based on latent semantic analysis and evaluates both the text and its meaning, providing also short responses for electronically submitted essays. In conjunction, there were also significant improvements in the field of spoken language assessment. In fact, systems focusing

primarily on the evaluation of the quality of pronunciation were implemented, for example starting from simple speaking tasks or reading texts by non-native English or Dutch speakers (cf. Bernstein et al. 1990 and Cucchiarini et al. 2002). However, since not merely pronunciation constitutes an important element in the evaluation of the competence in a spoken foreign language, fluency started also to be evaluated. In this context, though, also typical phenomena of spontaneous speech, such as pauses, hesitations and the like are to be taken into account. In this regard, the best systems in use are *Versant* (Townshend et al. 1998) and *SpeechRater* (Xi et al. 2008), applied in Educational Testing Services (ETS) exams and Test of English as a Foreign Language (TOEFL) speaking tasks. The former considers how the learner accesses language constituents and builds a discourse, measuring the clarity of expression, pronunciation, fluency and speed. Differently, the latter, i.e. the *SpeechRater*, focuses on the learner's ability to immediately interact with an interlocutor, considering also topical coherence and specific non-native speech aspects.

Recently, the use of word embedding techniques and deep neural networks has been introduced to perform automatic language assessment of spoken and/or written language. From these new approaches end-to-end systems have been developed that outperform those previously described. For example, BERT (Bidirectional Encoder Representations from Transformers), a powerful Transformer model, allows to bidirectionally read a sequence of words at a time and uses different training strategies other than distributional word embeddings, such as *word2vec* (cf. Mikolov et al. 2013) and *GloVe* (cf. Pennington et al. 2014). This particular type of model, along with the large amount of data on which it has been trained, allows for high accuracy in both automatic scoring of spoken and written language (cf. Devlin et al. 2019).

2.2. Open issues in automatic language assessment

The design of properly functional and adequate automated language assessment systems is usually correlated to a series of issues concerning both the datasets to use and the principles to follow in order to operate with them. First, the amount of open learner corpora is still limited and the majority of the data available are restricted to the English language. The reasons behind this may be primarily related to digitalization and privacy issues, apart from the fact that English is the first foreign language spoken all over the world. For the latter cause, in fact, many researchers and projects have invested in the English language rather than in other less widespread ones, such as Italian or German. Moreover, since the sort of content needed to train and evaluate the automated architectures is produced for language proficiency tests, it was unusual until a few years ago for these to be carried out in a computer-based modality. The standard procedure involved having students write texts and carry out monologues and conversations to be assessed by language experts. Therefore, this requires that the materials made available are either manually transcribed or, in the case of speech, using Automatic Language Recognition (ASR) systems. The latter, however, are generally not intended to recognise non-native accents and may therefore introduce additional errors.

Second, the corpora with which to train the automated models must already have been evaluated by human examiners and thus have correctly assigned labels to be fed into the machine learning framework. In the case of a macro-system applicable to several languages, i.e. trilingual, these labels must be the same cross-linguistically or at least in the same scale in order to be comparable. This represents yet another reason why this project appears to be challenging. There are not numerous ready-to-use datasets that have adopted the same evaluation metrics for

Italian, German and English. Those that do exist also may not present sufficient data to be compared with those for the predominant English language.

The third problem is the identification of objective evaluation metrics that can be applied to different groups of learners. The system to be implemented must be able to identify errors within written and oral productions, correct them and quantify them in order to map them to a grade and a language level of the CEFR. However, as each learner and each language are different, there are many, sometimes highly variable, aspects to consider. The measurement of competence does not merely require checking the formal correctness of learners' texts, otherwise only the form and not the actual use of language for communication would be considered. Aspects such as syntactical discourse structure, style, appropriateness of register, fluency and speed for spoken language must be taken into account and quantitatively rated.

Finally, especially with regard to oral language examinations, bias can sometimes arise due to a learner's accent, gender, or other characteristics that diverge from the standard fluent language learner. In addition, oral examinations are those that usually receive an immediate score if taken face-to-face, or if taken online require the examiner more time to carefully listen to them. Automating their assessment would be a great step forward facilitating not only the examiners but also the examinees. An automated system could, for example, measure aspects that are not relevant or objectively easy for a human to quantify and offer feedback targeted at each individual. However, constructing end-to-end systems for automated oral exams scoring is a considerable challenge requiring a large amount of high-quality and meaningful spoken data, difficult to collect, apart from a high level of abstraction and accuracy.

The open issues concerning automatic language assessment that are encountered and partly addressed in this research project can be summarised as follows:

- The scarcity of corpora of language learners for languages other than English and in digital format, especially for oral examinations;
- The almost absence of publicly available labelled data adopting universally valid methodologies and scales;
- The complexity of identifying objective but effective assessment metrics for different languages as well as for oral and written examinations;
- The quantification of non-strictly linguistic and subjectively highly variable factors such as accent and style;
- The presence of bias due to occasionally inaccurate and non-objective human evaluation.

2.3. Research Objectives

The goal of this project is the automatic evaluation of written, and in part also spoken, language proficiency exams among English, Italian and German language learners. The steps in which the execution of the task is structured include the following:

1) We run a set of experiments using a Transformer model on different datasets to assess proficiency by classifying the students' exams according to the CEFR competence levels.

a) We test matched data, namely data selected from the initial corpus used to design the starting architecture.
b) We make trials with mismatched datasets, meaning other than the initial one, applying our model on language exams from different learners' groups with distinct labelling systems.

2) We then create a system that automatically detects the errors and assigns a numerical class or score to the written and/or spoken language inputs.
   a) We map the scores to a given level of the CEFR.
   b) We track the foreign language learners' errors, both to understand prediction errors from the automated architectures and to attempt to correlate specific types of errors to given scores and/or levels.

The goals of this research project relate both to the data and the structures necessary for automatic language assessment, as well as to the development of an efficient automatised method to achieve an effective competence evaluation. They can be summarised as below:

- Development of a cross-linguistically valid generalisable methodology for the documentation of assessed language exams, both written and oral.
- Creation of a pipeline for the automatic correction of adult language learners' exams detecting errors and correcting them.
- Design of an automatic architecture employing Deep Neural Networks to grade and classify students' exams within the CEFR for language learning.
- Introduction of an unbiased system to automatically assess multilingual language competences considering objective and quantifiable learners' skills.

2.4. Research questions

The research questions that this project intends to tackle are deeply linked to what has been described in the previous two sub-sections related to problems and objectives (§ 2.2. and 2.3.). They can be clearly formulated as:

- After having identified open-source available datasets, is it possible to classify the contained examinations applying comparable labels to different languages despite the uniqueness of the languages and the language learners considered?
- Does a multilingual system capable of identifying and quantifying errors in the students' texts exist? What kind of errors are to be found? Do they characterise distinct levels or languages differently?
- Can we assess oral and written examinations using the same or highly similar metrics?
- Can neural architectures actually assign a proficiency level or grade to learners' productions that correspond with those that would have been assigned by humans? Do the aspects taken automatically into account by them match with the human-established ones?

2.5. Significance of the study

The contribution of this project is two-fold as it concerns both possible useful applications in the field of linguistic competence assessment in public and or private contexts, and advances in the field of language learning and testing research. As a matter of fact, like mentioned in the section dedicated to the statement of the problem (cf. § 2.2.), the creation and the use of an automatic language assessment system would give numerous advantages to those who employ proficiency tests, while at the same time it would increase the number of results and materials available concerning foreign language learners' peculiarities.

Firstly, there would be the advantages of a reduced workload for the test examiners, as well as of speeding up the scoring process itself. A linguistic model trained with sufficient data could rapidly process the digitised texts or oral examinations, detect errors, quantify them, correct them and finally assign a grade to the learner. The latter could also be mapped to CEFR levels systematically, providing a clear classification of language skills.

Secondly, automated architectures such as end-to-end or deep neural models would have the advantage of pursuing consistent evaluation metrics. These could assess different learners of an extremely diverse sample, such as students with different native languages, ages or genders in an objective manner. Furthermore, this would additionally contribute to the reduction of the biases at times present in the case of human examiners, who may be conditioned by variable personal factors like gender, accent, age of the language learner.

Thirdly, automated systems for correcting and assessing the proficiency of language users could offer the possibility of personalised feedback provided to each learner on the basis of their individual knowledge and skills extracted from the analysed texts. In addition to this, the critical areas in which the most errors were found could be identified and any generalisable patterns could be explored, whereby teachers and educators could strengthen their teaching and exercising in this regard. Moreover, language experts could also benefit from corpora containing these detailed annotations, which could offer interesting insights into the stages and patterns of foreign language learning.

Finally, the use of automatic systems for the evaluation of language competences would increase knowledge about the effectiveness and functioning of these systems. To date, few large test centres use these tools, but considering the increasing popularity of online language examinations and the rapid advancement of technologies, we can expect them to become mainstream. Therefore, we may start to think about both designing these systems and improving the existing ones, without losing sight of the fact that learners are human beings and consequently the metrics to be assessed cannot be completely error-free but must, on the contrary, be meticulously crafted.

2.6. Critical aspects and limitations

Given the different objectives of this multidisciplinary research, various critical aspects and limitations must be taken into account, especially with regard to the content to be analysed and the applied methods.

Firstly, given the small amount of data available for the study of language learners other than English, namely for Italian and German, there is the risk of obtaining results that do not fully reflect the actual conditions of the learners of these languages. Also regarding the unbalanced nature of our datasets for the different classes of the CEFR scale, it is possible that some results may suffer from this. The fact that the various used corpora were constructed from

materials of different learners and varying scoring and assessment methodologies also constitutes a limitation within which we can compare the three languages and results.

Secondly, concerning the methods applied in our research, both from the point of view of linguistic analysis and error classification, as well as the design of automatic assessment systems, there are diverse crucial aspects to consider. The linguistic analysis and quantification of elements relating to the lexical richness and syntactic complexity of the learners, for example, has been conducted on the basis of metrics that are commonly used to study either literary texts of significant length or native writers. However, ours is an attempt to extend existing techniques to texts of language learners in order to additionally ascertain their potential applicability and contemporarily raise the issue of the need for objective but customised metrics for this under-explored type of content. As far as error classification is concerned, there are few open-source systems available that allow several texts in different languages to be comparably processed. Given the open nature of these systems, however, which provides for their improvement through the employment of users, our results could be influenced by the fact that for languages such as Italian, there are still few researchers who have actually employed this system for the analysis of similar texts to ours and that have contributed to its implementation for this language.

Finally, the limited amount of data available and our constrained time frame prevented us from drawing robust and definitive conclusions in some cases, for example with regard to oral English examinations. If we had had more resources, we could have shed light into the various aspects that affect language learning from the theoretical sphere and from the field of neurolinguistics, as well as on typical factors that characterise the written texts and oral productions of foreign language learners derived from the interlanguage phase. In addition, we could have conducted other types of experiments, also using different inputs and variable architectures. Nevertheless, given the limitations mentioned above, this project identified interesting trends and promising research directions in the domain of automatic competence assessment among adult language learners of English, Italian and German.

CHAPTER THREE: RESEARCH DATA TO ASSESS WRITTEN PROFICIENCY

In this chapter we will describe in detail the datasets used within our project for the English, Italian and German languages, respectively. In the search for automatic language assessment resources available for foreign language examinations we noticed that most of the available data are predominantly related to the English language. There are two main reasons for this. The first reason concerns the fact that English is the foreign language with the highest number of speakers in Europe (EUROSTAT 2020) and in the world. The second reason, instead, regards the fact that the few systems implemented for the automatic assessment of both oral and written examinations have been applied on ELLs, such as e-rater® (Burstein 2003) or SpeechRater (Xi et al. 2008). Given these motivations, as discussed in the following sections, not only is the data for the Italian and German languages minor in numerical terms but also more imbalanced across the different proficiency levels and L1.

3.1. Datasets for English

The two English corpora we employed in our project are the English First Cambridge Open Language Database (Geertzen et al. 2014) and the Cambridge Learner Corpus (Yannakoudakis et al. 2011). The underlying ground for considering these two in particular over other English learners' examination data, such as the Pittsburgh English Language Institute Corpus (Juffs et al. 2020), is the availability of language exams information intended to measure the competence of learners following the CEFR framework in a mostly similar way. In the following sections more details are provided in this regard.

3.1.1. The English First Cambridge Open Language Database (EFCAMDAT)

The English First Cambridge Open Language Database (EFCAMDAT) arises from the collaboration of the Department of Theoretical and Applied Linguistics of the Cambridge University and the organization Education First (EF). It is a partially error-tagged large-scale longitudinal learner corpus covering 172 ELLs' nationalities and 16 levels of competence mapped to the 6 CEFR standard levels according to the correspondences indicated in Table 1 below.

EFCAMDAT levels	CEFR mapped levels
1-3	A1
4-6	A2
7-9	B1
10-12	B2
13-15	C1
16	C2

Table 1: EFCAMDAT mapping system in relation to the levels of competence defined by the CEFR

This updated version of the dataset consists of 1,180,310 essays about different topics submitted to the online school of EF, named *Englishtown* (Geertzen et al. 2014). The corpus has been tagged and parsed using NLP tools, namely the PoS dependencies from the Penn Treebank (Taylor et al. 2003) to perform the assignment of tags and the Stanford parser (Manning et al. 2014). Each essay has been corrected and evaluated by teachers. Learners also receive comments for their writing and can access the subsequent levels of competence upon positive evaluation. Additionally, the errors occurring in each text have been annotated using a set of error markup tags provided to the teachers. They mainly concern local morphosyntactic, spelling, subcategorization, word order and semantic errors. The following Figure is a pie chart representation of the distribution of ELLs by nationality within the EFCAMDAT corpus.

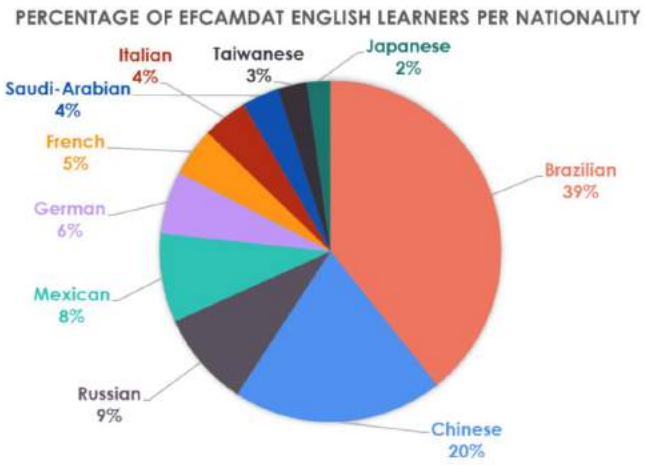


Figure 2: Percentage of English Language Learners within the EFCAMDAT corpus divided per nationality.

The content written by language learners varies according to the levels of competence they were supposed to demonstrate. Considering the presence of 16 proficiency levels (see Appendix B Figure 3) and eight units per each of them, the main topics and tasks covered are the following listed in Table 2:

Essay topic	CEFR level	Essay topic	CEFR level
Introduce yourself by email	A1	Give instructions to play a game	B1
Write an online profile	A1	Review a song for a website	B1
Describe your favorite day	A1	Write an email of apologies	B1
Tell someone what you are doing	A1	Write a movie review	B2
Describe your family's eating habits	A1	Turn down an invitation	B2
Reply to a new penpal	A2	Give advice about budgeting	C1
Write about what you do	A2	Cover a news stories	C1
Write a resume	A2	Research a legendary creature	C2

Table 2: List of the assignments of EFCAMDAT essays divided by CEFR levels of competence

This corpus constitutes one of the most complete corpora available for automatic language assessment research, as the data is clearly labelled, and errors are also carefully marked. Moreover, it represents a valuable resource for this field of research as it is being constantly updated.

3.1.2. Cambridge Learner Corpus for the First Certificate in English exam (CLC-FCE)

The Cambridge Learner Corpus for the First Certificate in English (CLC- FCE) exam is a collection of texts produced by ELLs for English as a Second or Other Language (ESOL) examinations generated by the collaboration between Cambridge University Press and Cambridge Assessment. In particular, these contents are part of the First Certificate English written exam to attest a B2 CEFR level, which consists of two tasks for which students need to write a text between 200 and 400 words. Those productions have been evaluated with a score between 1-40 and the errors contained in them have been classified in around 80 distinct classes (Yannakoudakis et al. 2011). The average score of the two texts produced by the learners has then been partially mapped to CEFR levels. In fact, since the student's writing skills were exclusively taken into account, based on only part of the exercises assigned for the FCE examination, the correspondence of the scores from 1 to 5 in this corpus is not exactly an accurate representation of the CEFR levels. The total number of learners considered is 1,238, between 16 and 30 years of age and from different nationalities. The number of exams present in the corpus is instead 2,469.

Figure 3 below represents the distribution of the native languages of the language learners and the number of examinations available per level within the different idioms.

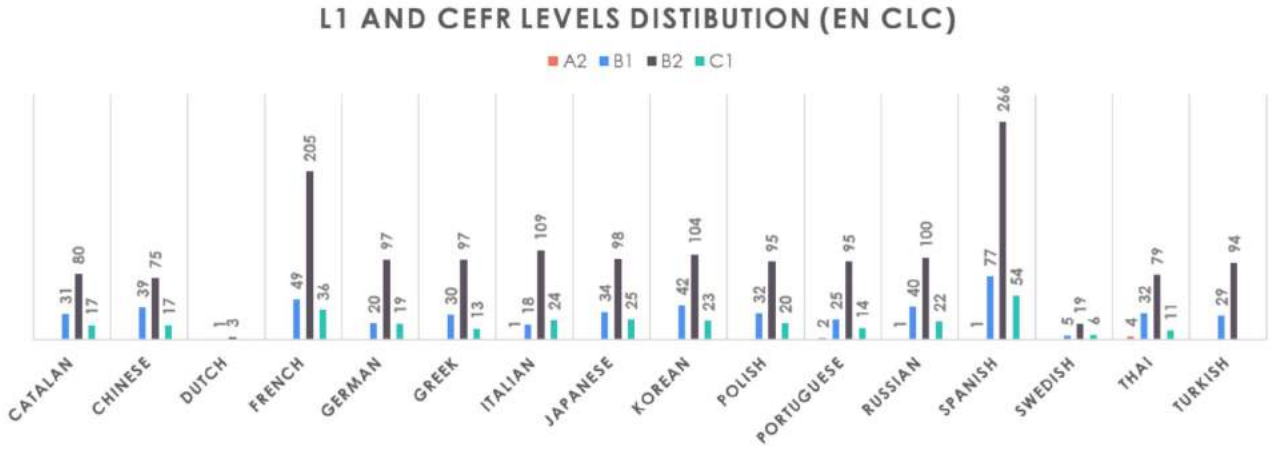


Figure 3: Distribution of L1 and CEFR levels within the examinations included in the CLC-FCE corpus

The format is in line with the requirements of the FCE writing task, which consists of a first part, i.e. an essay of approximately 140-190 words, and a second part with a choice between an e-mail, a letter, an article or a report of the same length of the first.

In comparison to the EFCAMDAT dataset, this one appears to be much less numerous. However, despite its small size, this dataset represents an important resource for the automatic assessment of competences, especially if a systematic analysis is to be made of the errors that have been precisely annotated in it and the corrections provided.

3.2. Dataset for Italian & German: MERLIN

For the Italian and German language we were able to access an open-source corpus containing empirical examples of learners' texts rated according to the CEFR sub-competences indicators. This resource was created within the

MERLIN project, funded by the European Commission during the years 2013 and 2015 thanks to the collaboration of the University of Technology of Dresden, the European Academy of Bozen, the Charles University of Prague, *telc* (the European Language Certificates) of Frankfurt/Main and the Eberhard Karls University of Tübingen with the European Centre for Modern Languages in Graz, the Ministry of Education, Youth and Sports of Prague. Together these research partners carried out and collected 2,265 written exams from learners of Italian, German and Czech. First of all, they had to electronically transcribe the original hand-written texts of language students applying shared transcription conventions. Then, they proceeded to re-correct them considering the general principles indicated by the CEFR (cf. Appendix A Fig. 2) and more particular language-dependent sub-competences related to lexicon, grammar, coherence, socio-linguistic appropriateness and orthography. They assigned a level to each of the latter and allocated a holistic level between A1 and C2 to the entire written exam.

The reasons that led us to the selection of this dataset as the source material for our research are the consistency in cross-linguistic assessment, the clear assignment and compliance with the CEFR indications for language competence evaluation and the availability of metadata about learners. The latter concern the age, the gender and the task assigned to each language learner. Each examination, whether in Italian, German or Czech, is evaluated according to the above-described linguistic parameters. Their transcripts are sufficiently clear and contain annotations for errors according to different categories and subfields (see Table 3). In addition, a human-corrected version, referred to as target text, is at times provided by one or two annotators.

Linguistic Field	Subfield	Phenomen
Ortography	Grapheme	Transposition Accent
	Word boundary	Split Merge
Grammar	Negation	Double negation
	Verb	Tense Voice
Vocabulary	Formulaic sequence	Collocation Idiom
	Form	Deviation Composition
Coherence/cohesion	Coherence	Text structural means
	Connectors	Accuracy
Sociolinguistic Appropriateness	Letter text type	Greetings/farewells Opening/closing formulae
Pragmatics	Request	Direct request Indirect request
General	Text intelligibility Sentence intelligibility	

Table 3: A section of the MERLIN annotation scheme for errors' phenomena[2].

[2] Adapted from Boyd et al. (2014), pp. 1284.

3.2.1. MERLIN Italian

MERLIN's Italian language section contains a total of 813 exams assessed as belonging to the CEFR levels between A1 and B2. More information about the exact number of exams per level can be found in Table 4 below.

Language to assess	LEVEL TESTED	N. OF RATINGS PER LEVEL	TOTAL N. TEXTS PER LANGUAGE
ITALIAN	A1	29	813
	A2	381	
	B1	394	
	B2	2	
	Unrated	6	
	No score	1	

Table 4: Distribution of number of exams collected per level in Italian MERLIN.

As can be seen, MERLIN Italian presents an unbalanced distribution of data for the lowest level, i.e. A1, and the highest attested level, i.e. B2. The total number of examinations may not meaningfully sufficient to train an automatic four-output grading system, as the classes are not equally represented.

The learners belong to different categories both for their age, ranging from 15 up to 74 years, and their native languages. The number of L1s per level is depicted in Figure 4 and Figure 5 below. From the pie chart we can observe more than 12 distinct languages of origin, with a visible prevalence of German and Hungarian learners over the rest. There are various languages representing less than 5% of the tests' material, for example Czech, Arabic, but even Italian and English are not numerous L1 categories.

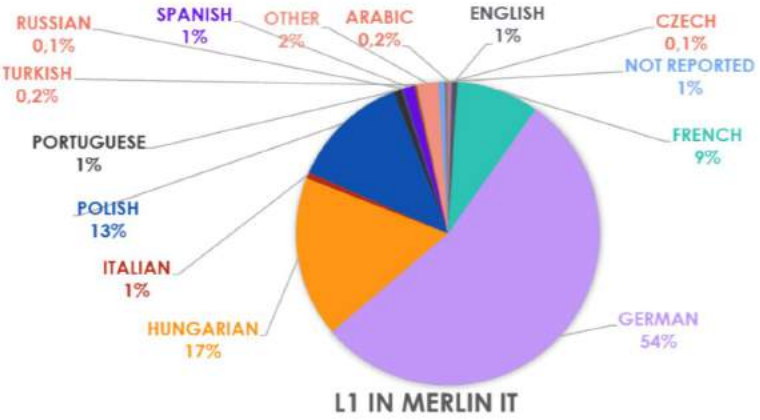


Figure 4: Percentages of Italian language learners in MERLIN by their first language

Additionally, the distribution of learners per CEFR level is not balanced as well. The highest number of examinations available considering also the L1 concerns the A2 level for German native speakers (245), followed by the B1 level exams (177) for the same group of speakers and right after by Hungarian learners for the same level of competence (122). The least numerous native languages are Czech, Russian and Arabic and Turkish, while the

most numerous apart from German are Hungarian and Polish (see Figure 4). These aspects, although significant, will not be specifically addressed in the current project.

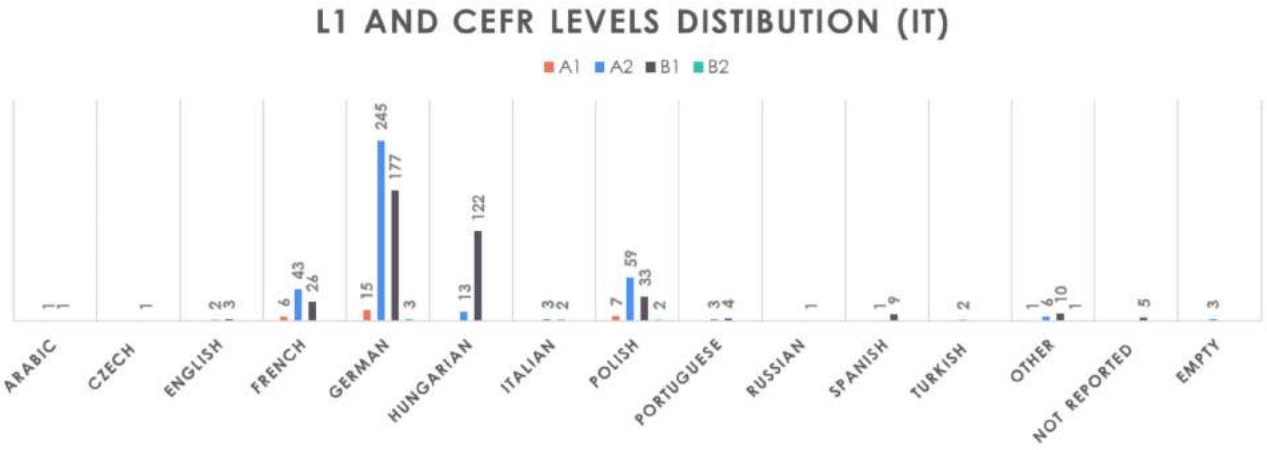


Figure 5: Distribution of L1 and CEFR levels within the examinations included in the MERLIN Italian corpus

Every language learner had to write a different assignment considering the level for which they took the test (see Table 5). Depending on the competences to be demonstrated the tasks varied from formal or informal emails, to essays, reviews and invitations. The minimum text length found is 15 words, while the maximum length corresponds to 308 words.

Tasks for ITALIAN texts (emails/letters to friends or institutions)	Texts per task	Level
answer to a wedding invitation	99	A2/B1
apply for internship in company	16	A2/B1
apply for internship in fashion sector	40	B1
ask for information about International Cooking Evenings	9	A2/B1
complain against a hotel	32	A2/B1/B2
contact a friend after a long time	65	A1/A1/B1
cook with teacher	2	B1
describe experiences with language learning	39	A2/B1
go see a friend	45	A1/A2/B1
help a friend who is looking for work	88	A1/A2/B1
help a friend who is looking for work after school-leaving exam	78	A2/B1
help someone who has problems with chats	12	A1/B2
inform friends about language course	81	A1/A2/B1
inform oneself about an aid project	33	B1
inform oneself about language course	23	A2/B1
reschedule an appointment	108	A1/A2/B1

Table 5: Assignments for the Italian written tests contained in MERLIN divided by CEFR level

Despite the non-homogeneity of data by language level and the modest size of the corpus, in general MERLIN Italian constitutes one of the few open-source resources with metadata available and downloadable for free dedicated to Italian language learners. This renders it a valuable resource of data from foreign language learners.

3.2.2. MERLIN German

MERLIN's German language section contains a total of 1,033 exams between the A1 and C2 CEFR levels of competence. The exact number of texts for each of the six levels can be found in Table 6 below.

Language to assess	LEVEL TESTED	N. OF RATINGS PER LEVEL	TOTAL N. TEXTS PER LANGUAGE
GERMAN	A1	57	1033
	A2	306	
	B1	331	
	B2	293	
	C1	42	
	C2	4	

Table 6: Distribution of number of exams collected per level in German MERLIN

From the numbers above we can get a clear understanding that although there are more levels than for the Italian language, again the number of exams per level is rather unbalanced. With the exception of the A2, B1 and B2 levels, the other classes contain less than 100 exams, with the C2 one having only 4 exams to represent it. Compared to the examinations collected for the Italian language, we find also more native languages per speakers, namely 16 distinct idioms (see Figure 6 and 7).

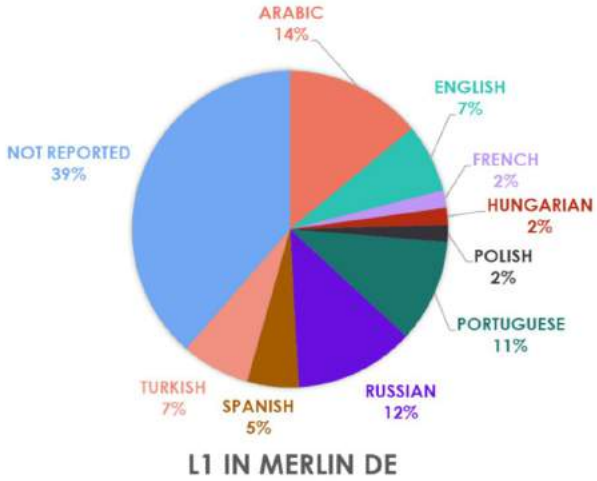


Figure 6: Percentages of German language learners in MERLIN by their first language

The pie chart above highlights that for a significant percentage of data, namely 39%, the language of origin of the participants is not known. Besides that, however, the most numerous groups of language learners are Arabs, Russians and Portuguese. On the contrary, the least numerous are the French, Hungarians and Polish.

Figure 6, instead, illustrates in more detail the distribution of the different levels within each distinct group of the examined language learners. We can notice from the bars of the histogram that taking into account the non-reported group, the most numerous exams are those concerning the A2 and B1 level. Beyond this category, on the other hand, the most numerous exams in order of the CEFR levels are exams of Russian learners for A1, A2 and B1, other and Russian again for B2 and C1. Besides these groups, Spanish and Polish learners also represent significant groups.

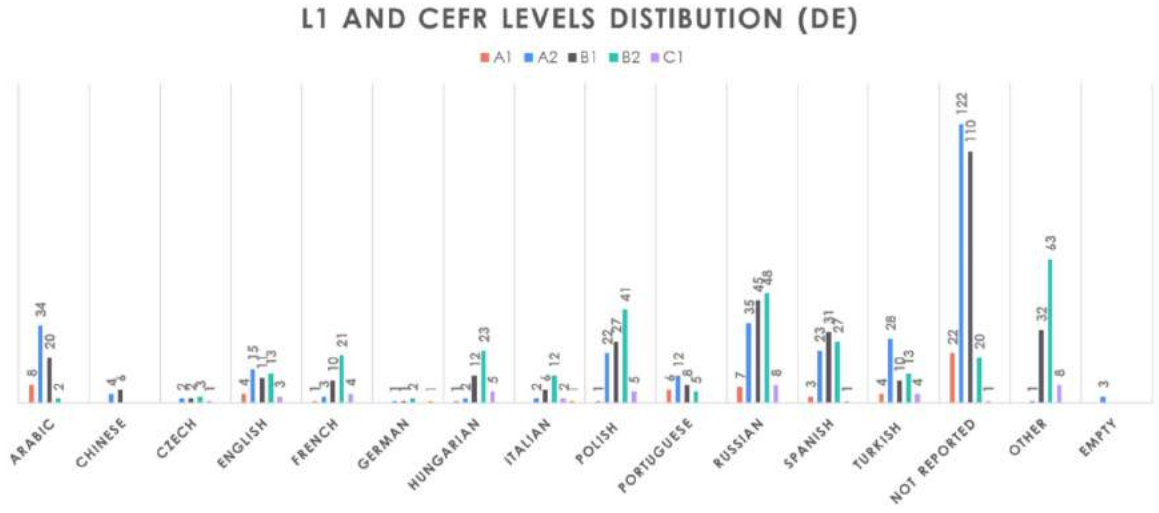


Figure 7: Distribution of L1 and CEFR levels within the examinations included in the MERLIN German corpus

For each level MERLIN German collects several tasks that the learners had to complete in order to attest their competence. As for Italian, they vary according to the level. We found a maximum length of 366 words. The register also varies from informal to formal as the level of competence rises. Therefore, they include emails, letters but also letters of reference or complaint, as well as essays and articles (see Table 7).

Tasks for GERMAN texts (emails/letters to friends or institutions)	TEXTS PER TASK	CEFR LEVEL
arrange appointment with friend to go swimming together	56	A2/B1/B2
ask friend for help for finding apartment	77	A1/A2/B1
congratulate to birth of child	74	A1/A2/B1
Unknown	315	A2
ask friend for help for finding apartment	766	A2
ask friend to take care of pet	72	A2/B1
letter to housing office	70	A1/A2/B1/B2
offer a ticket not used to a friend	65	A1/A2/B1/B2
write a letter for New Year to a friend	73	A1/A2/B1/B2
write a letter to a friend (birthday)	70	A2/B1/B2
write a letter to a friend (visit)	67	A1/A2/B1/B2
Au pair writes letter of complaint to Agency	70	A2/B1/B2/C1
apply for internship in sales department	72	A2/B1/B2/C1
ask for information at Au pair Agency	65	B1/B2
essay about why it's of value to learn German	42	B1/B2/C1/C2
reporting about the housing situation	72	B1/B2/C1/C2
write article about sticking to one's traditions and 'assimilation' in a new environment	90	B1/B2/C1/C2

Table 7: Assignments for the German written tests contained in MERLIN divided by CEFR level

CHAPTER FOUR: METHODOLOGY

In this chapter we provide a systematic description of the approach used in our project. Therefore, we first present the tools employed both for the automatic text correction of the original students' examinations, mainly LanguageTool, and the system and architecture we used for their classification, meaning a BERT-based Transformer model. These two components are necessary for performing the multi-class classification, i.e. for classifying the different examinations according to the five distinct CEFR levels of proficiency, corresponding to the neural networks' final layer logits. The distinguishing characteristic of this task compared to other types of classification, for example binary or multi-label classification, is the assignment of different elements within more than two classes according to the principle that each one can exclusively belong to a single class. It is precisely for this reason that we adopted this system of classification in our project, whereby each learner text can only be categorized within one of the several existing classes, corresponding to the level of competence assigned according to the label provided in the original dataset.
In the following sections we provide a more comprehensive description of the functioning of LanguageTool, the specifications of the Transformers models, the architecture used in this project and the data processing of each dataset.

4.1. LanguageTool

In this project, to perform the automatic correction of the texts originally written by language learners, we used an automatic open-source multilingual style and grammar checker known as LanguageTool. It is an instrument that, starting from a raw text, is able to correct it systematically identifying different categories of errors, namely:

- **Spelling errors** identified by comparing the words of the original text with a wide source;
- **Structural and non-structural grammatical errors** identified by taking into account both context information and a set of rules for a given language;
- **Stylistic errors** identified by considering the presence of ambiguities and the clarity of the text, as well as the context of use for which it is intended;
- **Semantic errors** identified in case of exclusion from the previous classes.

Fundamentally, this Java-written system operates by means of rule-based checking. The rules in the initial implementation of LanguageTool were manually developed (Naber 2003), unlike statistical checking when errors are detected by considering the sequences of PoS tags and their more or less frequent occurrences in a language. Since the project's launch, the number of supported languages has expanded to more than 30, including major languages like Italian, German, six English variants, but also less resourced languages such as Catalan, Tamil, Esperanto and Tagalog. The automatic checker is based on surface text processing (see Figure 8), does not use a deep parser and does not require a fully formalised grammar (Miłkowski 2010). Furthermore, LanguageTool allows rules to be modified or built semi-automatically using corpora. To date, it is possible to use it through Python, but it

is supported by more than 25 software including search engines such as Chrome, Firefox, as well as writing programs such as Google Docs, OpenOffice, Emacs and Word. The languages involved in this project, namely Italian, German and English, are supported by PoS taggers and lemmatizers. Respectively for each of them they are Morph-it, Morphy and open-source data for English together with WordNet (Miłkowski 2010).

Figure 8: LanguageTool text processing and correction pipeline.[3]

The pipeline that the LanguageTool system follows first performs sentence and word tokenization. After this, the parts of speech are tagged. A process of disambiguation and chunking is applied to the words and their corresponding tags. Subsequently, the rule matching process begins, which takes into account a number of rules currently available in Java or XML format. Following this stage the user can choose whether to display all error categories, exclude some from being employed or apply them to the whole text generating a new version of it.
LanguageTool constitutes a very powerful instrument that can offer a complete view of errors in texts, even in the case of content from foreign language learners. The possibility of being applied directly on the original texts and the support of numerous languages makes it particularly suitable for our project.

4.2. BERT-model

The main model used for the implementation of our multi-class classification task belongs to the set of models with deep neural networks (DNNs) inspired by BERT (cf. Devlin et al. 2019) from Google Research. These types of architectures, characterised by efficient parallel training and the capacity of capturing long-range sequence features, distinguish themselves for model size and training data. Being pretrained on generic large corpora, they can be conveniently used for a wide range of specific task, including text classification, language understanding and machine translation (Wolf et a. 2019). The training strategies used in order to train such models, which are often referred to as Transformers (Vaswani et al. 2017), differ from those typically used by the previous systems concentrated on the single words. They are Masked Language Modelling (MLM) and next sentence prediction (NSP). The first, MLM, consists in using 15% of the words of a given text replaced by the so-called MASK token to predict the original value of the substituted words based on the sequence of the remaining present words. The second strategy, NSP, consists of the model's ability to predict whether a given sentence is the match against a half-sentence provided by the text. That is, the model receives 50% of the training set made of pairs of sentences and learns to predict for another 50% of the set the missing half by extracting it from the reference corpus. In the following sections we first report more details about the Transformer models and their functioning in respect to

[3] Adapted from Miłkowski (2010), pp. 549.

other NLP systems. We then provide a description of our applied architecture and its main characteristics, explaining why they make it particularly adapt to our task of automatic language assessment.

4.2.1. Transformers

In 2017 a new type of deep learning model for NLP entered the scene, namely the Transformer. Before its introduction, the systems used to process human language were based on recurrent neural networks (RNNs) built with an addition of attention mechanisms to reproduce human cognitive attention. The application of this procedure is typically performed in order to highlight desired sections of the input data in opposition to the rest. In this manner the neural network can employ more computational power on those selected parts instead than on a larger amount of data (Vaswani et al. 2017). Additionally, since they proceed non-sequentially, the design of Transformers appears to be optimal when processing sequential data, like the words of a language. For these reasons they have overcome other RNN-based models, becoming the preferred and most employed architectures to solve NLP problems.

Prior to the release of the Transformer, researchers would usually rely on Long-Short-Term Memory (LSTM) and gated recurrent units (GRUs) systems. The first, namely LSTM, are generally composed of various cells, containing an input and output gate and a so-called forget gate. One can establish a time interval during which allowing the cells to let information flow. The idea of such systems originated in a partial attempt to solve the vanishing gradient problem in classic RNNs, found since backpropagation and learning methods based on gradients may at times inhibit the weights of the networks from updating their values or even stop the networks' training process (Hochreiter & Schmidhüber 1997). LSTM networks have proven their efficiency on time serial data such as speech and handwritten texts. For this reason, with the years they have improved and have been applied for example by Google and Apple for Google Voice search (Sek et al. 2015) and Siri (Smith 2017), apart than for translation tasks, for example by Facebook (Ong 2017). The second systems, namely GRUs, created as alternative RNN Encoder-Decoder models, present a similar structure than LSTM but lack the output gate and have an added hidden unit to improve memory capacity (Cho et al. 2014). The two parts, namely encoder and decoder are trained in concomitance maximizing the conditional probability of a variable-length target sequence of words or other type of sequential input. Their accuracy has reached comparable results to LSTM in linguistic tasks like speech signal processing and NLP (Ravanelli et al. 2018). However, both LSTM and GRUs need the introduction of additional attention mechanisms when they have to process long input sequences, since the models' results tend to be less precise. On the contrary, Transformers being attention-based from their construction can achieve significantly good results on long sequential data without having to use RNNs. This renders them efficient architectures that are able to parallelly process large amounts of linguistic data tokenizing and estimating attention weights between them in a not necessarily Sequence to Sequence manner (Vaswani et al. 2017).

4.2.1.1 General Transformer architecture

Similarly to the other RNNs architectures before them, Transformers are based on an encoder and a decoder unit which operate together. The first processes the input in an iterative manner and each of its layers consist of

encoding modules which can be stacked on top of each other for a desired number of times. The second unit, namely the decoder, presents a more numerous number of layers that encode the output received from the encoder after being processed (see Figure 9).

Each layer of the model employs attention mechanisms that weight the relevance of each received input since no recurrent networks being able to remember the order in which the sequences have been fed into the model are available in this case. Therefore, each word in the case of language input is assigned a relative position in a given sequence which is added to an embedded representation of the words themselves. Feed-forward neural networks for additional output processing are generally present both in the encoder and decoder layers. More details about their structure and units are to be found in the following two subsections.

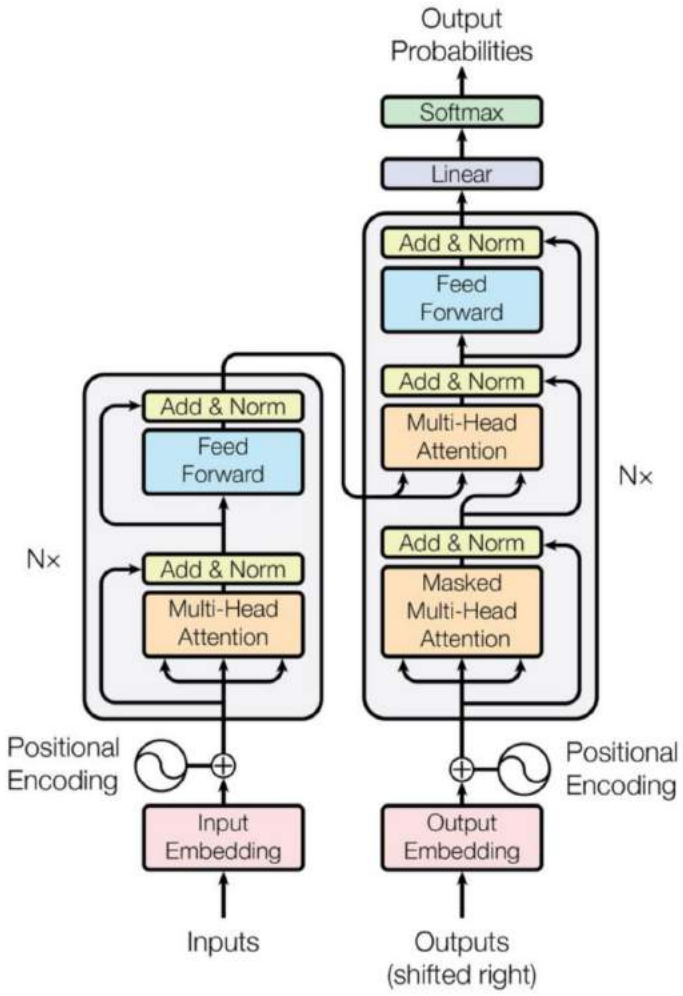


Figure 9: The architecture of a Transformer model[4]

[4] Image taken from "Attention is all you need", Vaswani et al. (2017), pp. 3.

4.2.1.2. Encoder

The encoder module of a Transformer model usually consists of a Multi-Head Attention and Feed Forward component. In the case of the model depicted in Figure 9 there are N 6 identical layers stacked on top of each other which contain the above-described components.

Multi-Head attention consists of parallelly running attention layers that operate following this equation:

$$\text{Attention(Q,K,V)} = softmax\left(\frac{QK^T}{\sqrt{d_k}}V\right)$$

Therefore, the obtained attention value is a value between 1 and 0, given the application of the softmax function on the operation of the Q, K,V values and the different learned linear projections d. Q represents a matrix containing the vector representation of a sequence of words, K are, instead, all the vector representations of the words or keys within that sequence and V are the values of the latter.

The encoder Multi-Head Attention layers interpret V as the same word sequence than Q. The V values are then multiplied and summed to attention weights which result from the influence of each word in the Q sequence on the other sequence K.

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O \\ \text{where head}_\text{i} &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{aligned}$$

The mechanisms of attention are repeated several times allowing the system to learn different representations of Q, K and V. The latter differ for each differently placed attention modules in the architecture.

The Feed Forward component of the encoder, on the contrary, is fully connected to the previous layers with residual connection, followed by layer normalization. It consists of the following transformations:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

The formula above indicates two linear transformations with *ReLU* (Rectified Linear Unit) activation in between them. Overall the Transformer encoder uses learned embeddings to convert the input transformed in tokens to vectors.

4.2.1.3. Decoder

The decoder of the Transformer model contains the same number of stacked layers as the encoder, namely six. However, in addition to Multi-Head Attention and Feed Forward components, it has a third sublayer employing Multi-Head Attention over the encoder's output that it receives as input. These layers are connected in the same way as described for the encoder. Differently to attention mechanisms used in sequence to sequence (Seq2Seq) models, the Transformer's decoder employs the scaled-dot product attention in the input of the softmax of Multi-Head attention layers. From the output received sequence the decoder takes positional data and word embeddings as inputs. In order to prevent reverse information flow the sequences of outputs need to be in part masked. As can

be observed in Figure 9, last layer of the decoder presents a linear or softmax layer outputting the probabilities in the vocabulary.

4.2.2. Our architecture

The main model used for the implementation of our multi-class classification task regarding the CEFR levels of competence belongs to the set of models with deep neural networks (DNNs) inspired by BERT (cf. Devlin et al. 2019) from Google Research. These types of architectures, characterised by efficient parallel training and the capacity of capturing long-range sequence features (§ 4.2.), distinguish themselves for model size and training data. Being pretrained on generic large corpora, they can be conveniently used for a wide range of specific task, including text classification, language understanding and machine translation (Wolf et a. 2019).
The employed model for the language exams' multi-class classification task, namely *bert-base-uncased,* belongs to the Hugging Face Transformers Library released in 2019 (Wolf et al. 2019). It is an Open Source and general-purpose library which focuses on the diffusion of Machine Learning models for NLP tasks and contains not only models' architectures, but also tools that facilitate the training and development stages. The architecture we used consists of three principal building blocks:

- A **Tokenizer**, which translates raw text strings into sparse index encodings;
- A **Transformer**, which transforms the previously generated sparse indices into contextual embeddings;
- A fixed **Head**, in our case BERT, which uses contextual embeddings to generate the specific predictions for text classification task.

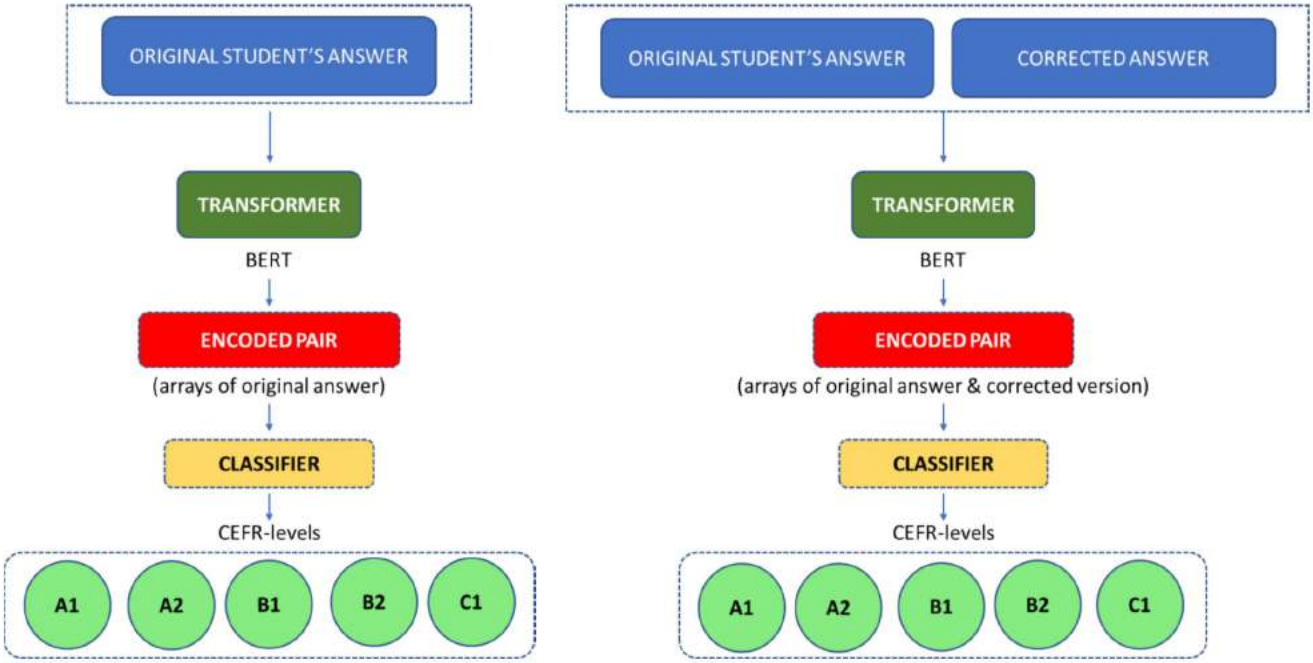


Figure 10: BERT-based model with only original students' texts (left) vs original students' texts and human/automatic corrections (right) employed in multi-class classification task

As displayed in Figure 10, our architecture can exclusively take as input the students' answers to the different language exam assignments or those followed by either a manual or automatic correction generated using LanguageTool, an open-source spellchecker. The system employs a BERT-based model to derive a compact representation of each or both texts. The principal layer, meaning the Head, is frozen, therefore, its parameters do not have to be re-trained. Following the BERT-layer, there is a Dense layer consisting of 768 units, on which the rectified linear activation function, *ReLu*, is applied with a dropout rate of 0.2. The *ReLu* function takes a single number as input and returns 0 if the number is negative, or the value of the original input if it is positive. Another Dense layer with 128 units succeeds the previous one with the same applied activation function and dropout rate. Generally depending on the dataset used, the final layer exhibits 5 units, each one corresponding to a CEFR level (A1, A2, B1, B2, C1) with *softmax* as activation function. This function transforms a vector of numbers into a vector of proportional probabilities in respect to the relative scale of each of the vector's values. More in detail, in

$$\text{Loss} = -\sum_{i=1}^{\substack{\text{output}\\ \text{size}}} y_i \cdot \log \hat{y}_i$$

our model the softmax function is applied to normalize the outputs and return probabilities that summed together return a value of one.

The loss function we selected corresponds to the *categorical cross-entropy function*, typically used for multi-class classification tasks where each input exclusively belongs to one category, as in our case. As can be deduced from the formula, the categorical cross-entropy loss function computes the loss by calculating the sum of the scalar values, resulting from the model outputs, and their respective target values, while considering the number of scalar values contained in the output.

In order to be able to classify the students' texts according to the proved CEFR level using the above-described architecture, we first need to tokenize the texts. The process of tokenization refers to dividing the sentences of a text into individual words. More precisely, when dealing with sequences of strings in Python, we split the longer strings into sub-word tokens. To tokenize the students' exams and their corrected versions provided by LanguageTool, we will be using the BERT Transformer's AutoTokenizer. The Transformer *bert-case-uncased* model will receive as inputs the texts' arrays and not the original strings. The reason for this is given by the fact that arrays are more efficient and less complex to use data structures for scientific and engineering applications, especially when it comes to natural language and large datasets. The single encoded original text or the encoded pair, consisting of the original exam and its corrected version, will be forwarded to the classifier, of which the final layer will output the corresponding CEFR levels of competence.

4.3. Data processing

In order to start working with the English, Italian and German language datasets described in detail in section 3, we first had to create from the original texts the LanguageTool corrected versions, from which we also obtain the error count and categories for each learner text in each language. The procedures for this purpose were not exactly the same since both the data format as well as the correction grids and assigned scores were somewhat dissimilar

across the languages. In the following sections we describe in detail the processes we followed for English, Italian and German corpora, respectively.

4.3.1. Data preparation for English EFCAMDAT

The EFCAMDAT corpus with its 1,180,310 essays constitutes the larger dataset used in our project. It comes with the original students' essays, for which there are corresponding ID codes, the anonymised identities of the authors, their nationality and L1. Moreover, since the original texts have been evaluated by experts, we are provided with the levels to which they have been assigned. However, since the original levels of competence were distributed along 16 classes, we had to map those to the six classes provided by the Common European Framework of Reference for foreign language learning (see Table 1).

As previously stated, the model we intend to use typically takes the automatically corrected version and the original learners' texts as input. In order to obtain the first, namely the LanguageTool checked version, we had to process the data and feed only the texts into the Python pipeline. After having selected the desired language, namely the available British English variant, we applied the pre-defined set of rules to our textual data. The tool returns a series of *Match* objects from which one can derive the position of the error within the text, the *Rule_ID,* a message explaining the error or offering a hint and a suggestion for a possible correction. Additionally, there is the possibility whether to apply the suggestions to the text (see Figure 11).

Original EFCAMDAT text extract (A2):
"Hi Everyone, Were having a Reems birthday party on Thursday, May 6th at 1:00PM Place: No.5 Yello w street. Please come. Dont worry."

LanguageTool error matches:
Offset 12, length 2, **Rule ID**: WHITESPACE_RULE
Message: Possible typo: you repeated a whitespace
Suggestion:
Hi Everyone, Were having a Reems birthday party on Th...
^^
Offset 14, length 4, **Rule ID**: WERE_WE_RE
Message: Did you mean "We're"?
Suggestion: We're
Hi Everyone, Were having a Reems birthday party on Thursd...
^^^^
Offset 120, length 4, **Rule ID**: EN_CONTRACTION_SPELLING
Message: Possible spelling mistake found
Suggestion: Don't

Figure 11: Example of LanguageTool Python interface output on a EFCAMDAT English text extract

For our purposes, we were also interested in obtaining the number of words contained in a text, the number of found errors and the errors' percentage obtained with the previous two data. Finally, we created a Tab-Separated Values (*tsv*) file containing the information for all the texts. Using this file we have subdivided the texts for the training and testing of the BERT model. After removing the texts containing less than 20 words and those written

in the students' L1, we obtained 723,282 texts, from which we separated 1,447 exams as a test set for future experiments.

4.3.2. Data preparation for English CLC- FCE

Apart from the EFCAMDAT dataset, we also used another corpus of English language learners for the experiments of our project, namely CLC- FCE. This corpus is much less substantial, comprising only 2,469 examinations, 12 of which contain no assigned scores and, therefore, had to be removed. Nevertheless, it features useful information for comparison with the other corpus. The material initially available consists of a Comma Separated Values (*csv*) file containing the texts written by the different language learners, their anonymised IDs, their language of origin and a score assigned by human examiners ranging from 0 to 40. The latter was then transformed into another score, marked as *exam score* indicating the written part of the B2 English tests from which they were originally taken, which values range between 0 and 5.3. We used the latter to map the texts to CEFR levels according to the corresponding system depicted in Table 8.

CEFR LEVELS	CLC- FCE MAPPED SCORES
A2	0.0 – 1.1
B1	1.2 - 2.3
B2	3.1 – 4.3
C1	5.1 – 5.3

Table 8: Mapping system for CLC- FCE assigned exam scores to CEFR levels of competence

Similarly to what we described in the previous section, we proceeded to use the LanguageTool checker on Python with the source student texts. The system then detected the different types of errors, listed them and corrected them, generating a new version to eventually feed the Transformer model as input. In total, we estimated a total of 1,997 applied rules. We counted them together with the number of errors and calculated the percentage of errors per text. A detailed analysis of the errors found in each dataset and CEFR level is available in section 5. Finally, we created a *tsv* file containing the information for all the 2,469 English texts. We proceeded to the creation of the partitions needed to train, test and validate the future model to be designed for the automatic classification of the exams based on the original and LanguageTool text versions. The obtained training set consisted of 2,017 texts, the validation set of 159 texts and the test set of 194.

4.3.3. Data preparation for MERLIN datasets

The MERLIN datasets appear differently structured in respect to the English corpora. The learners' exams are made freely available in form of individual text files, which also contain metadata, namely the author ID, the information concerning the language of the test and the level for which the exam was taken, the age, mother language and gender of the examinee. In addition to this, also six distinct ratings are provided in relation to the CEFR evaluation principles (see Appendix B Fig.1 and 2), namely:

- Grammatical accuracy
- Orthography
- Vocabulary range
- Vocabulary control
- Coherence/cohesion
- Sociolinguistic appropriateness

Extracting the mean between the previously listed sub-competence areas, they obtained an overall CEFR rating per student, which at times appeared to be higher than the tested level.

4.3.3.1. Data preparation for MERLIN Italian

The Italian dataset considered from MERLIN contains 813 student texts of different typologies, length and level. The metainformation concerning the language learners would have been particularly interesting if only it had come with associated numerical scores assigned to each area of competence. However, for our experiments the original transcribed student texts needed to be considered sufficient, since this resource is one, if not the only, easily consultable corpus of foreign Italian language learners.

After having collected all the student texts and information in a unique tab-separated values file and having applied the Python string *strip()* method on each of them to avoid the propagation of annotation mistakes, we proceeded to process the data with LanguageTool. We processed each of the learners' content using the Tool on Python and extracted the identified errors, counted them and calculated the percentage of errors per text considering the length of each. In total, the system applied 166 distinct rules on the texts. Below there is an example of the obtained results with an Italian learner production (see Figure 12):

Original MERLIN IT text extract (A2) :
"Nel nostro ristorante abbiamo avuto molto lavoro restarativi le ultimi mese. Ma l'architecto, che a planado tutto, costa molto soldi."
LanguageTool error matches:
Offset 49, length 11, **Rule ID**: MORFOLOGIK_RULE_IT_IT
Message: Trovato un probabile errore di battitura.
Suggestion: restaurativi; re starativi
...o ristorante abbiamo avuto molto lavoro restarativi le ultimi mese. Ma l'architecto, che a ...
^^^^^^^^^^^
Offset 82, length 10, **Rule ID**: MORFOLOGIK_RULE_IT_IT
Message: Trovato un probabile errore di battitura.
Suggestion: Architetto; Architettò; architetto; architettò
...lavoro restarativi le ultimi mese. Ma l'architecto, che a planado tutto, costa molto soldi...
^^^^^^^^^^
Offset 100, length 7, **Rule ID:** MORFOLOGIK_RULE_IT_IT
Message: Trovato un probabile errore di battitura.
Suggestion: planalo; planando; planano; planato; planavo; plana do
... le ultimi mese. Ma l'architecto, che a planado tutto, costa molto soldi.
^^^^^^^

Figure 12: Example of LanguageTool Python interface output on a MERLIN Italian text extract

Finally, we stored the obtained results in a *csv* file and created the partitions necessary to train a dedicated model. Our partitions were primarily made for the application of cross-validation techniques since the available data were limited. Therefore, we divided the original dataset in three sub-datasets of 269, 268 and 269 texts each.

4.3.3.2. Data preparation for MERLIN German

The German dedicated dataset from MERLIN is constituted of 1,033 texts in total from all the six CEFR levels of competence from A1 to C2. As in the case of the Italian portion of the dataset, the content was made available in 1,033 single text files containing the learners' original texts, the metadata associated to each participant of the examination and the related obtained scores assigned by human examiners.

Before starting the experiments for a dedicated German Transformer model that is able to classify the texts in a class corresponding to the CEFR levels of competence, we had to re-organize the data, grouping them in a single file from which we could extract paired original exam texts, assigned levels and metainformation. Therefore, we automatically checked the correctness of the manual electronic transcriptions of the handwritten original learners' texts applying the *strip()* function in Python. After that, we proceeded to the creation of a data frame with columns dedicated to each person metadata, like gender, age, native language, tested language and level, assigned overall and individual scores. Subsequently we used the LanguageTool German checker on each text and updated the data frame with the found error typologies, their number, their percentage in the text and an automatically generated corrected version of the original text, created by accepting the tool's suggestions. In total, LanguageTool applied 1,100 distinct rules to the considered learners' texts. Below in Figure 13 we can observe some of the individuated rules applied to a text excerpt:

Original MERLIN DE text extract (A2):
"Seit ein Jahr Wir suchen 4 Zimmer Für meine Familie, weil meine alte Wohnung zu Klein, Wir brauchen ungever 90 Q Mette Mit Warm, ich bitte mein Damen-Herren."

Offset 14, length 3, **Rule ID**: DE_CASE
Message: Außer am Satzanfang werden nur Nomen und Eigennamen großgeschrieben.
Suggestion: wir
Seit ein Jahr Wir suchen 4 Zimmer Für meine Familie, weil...
^^^
Offset 34, length 3, **Rule ID**: DE_CASE
Message: Außer am Satzanfang werden nur Nomen und Eigennamen großgeschrieben.
Suggestion: für
Seit ein Jahr Wir suchen 4 Zimmer Für meine Familie, weil meine alte Wohnung ...
^^^
Offset 139, length 17, **Rule ID**: DE_AGREEMENT
Message: Möglicherweise fehlende grammatische Übereinstimmung des Kasus (Fall: Wer/Was, Wessen, Wem, Wen/Was - Beispiel: ‚das Fahrrads' statt ‚des Fahrrads') und Genus (männlich, weiblich, sächlich - Beispiel: ‚der Fahrrad' statt ‚das Fahrrad') und Numerus (Einzahl, Mehrzahl
Beispiel: ‚das Fahrräder' statt ‚die Fahrräder').

Figure 13: Example of LanguageTool Python interface output on a MERLIN German text extract

With the aim of designing a model dedicated to the written German language we also separated this dataset into three parts for the purpose of training, testing and validating it. Therefore, we divided the corpus content into 775 texts for training, 130 for testing and 131 for validation. However, given the imbalanced composition of the corpus, we also divided it into three equal parts for cross-validation in case they were needed.

CHAPTER FIVE: AUTOMATIC CORRECTION WITH LANGUAGETOOL

This section is focused on the linguistic and error analysis of the datasets for English, Italian and German used in this project. More specifically, we describe how we quantified and differentiated the errors automatically detected by LanguageTool after feeding it the learners' original texts. Furthermore, since some datasets, such as CLC- FCE and MERLIN Italian, contained corrections annotated by the examiners who assessed the learners' proficiency, we compared them to the ones of the automatic checker. Given the systematic nature, consistency and accuracy of LanguageTool compared to human examiners' diverse evaluations, a detailed analysis and the comparison of errors between different datasets was possible.

Finally, we considered additional features extracted using *Stanza*, a collection of tools available in Python for NLP in more than 60 languages (Qi et al. 2020), and *textcomplexity*, an implementation of complexity measures to assess lexical richness and syntactic complexity in written texts (Proisl & Schöch 2020). Among the considered aspects regarding the latter, we mainly focused on:

- number of unique lemmas in text,
- vocabulary lexical diversity (HD-D),
- textual lexical diversity (MTLD),
- average sentence length,
- average dependency distance,
- number of dependents per word.

The purpose of this analysis was mainly checking the actual correspondence between the different proficiency levels indicated in the datasets and these latter objectively quantifiable values indicating the texts' lexical density and syntactic complexity.

5.1. English corrections

We used LanguageTool, the spelling and grammar checker described in section 4.1., to automatically process and correct the learners' original written tests. In particular, we employed *language_tool_python*, a LanguageTool Python wrapper allowing the selection of different languages, classes and rules and their application on numerous files simultaneously. Therefore, we started processing the English texts from the EFCAMDAT and CLC- FCE datasets. A detailed description of the obtained outputs and subsequent analysis are to be found in the following two subsections.

5.1.1. EFCAMDAT LanguageTool errors and linguistic features

We first started by working with the EFCAMDAT dataset, or rather a portion of it containing 1,447 files selected from a total of over one million essays. Hence, we retrieved the original writings and processed them by means of LanguageTool, deciding to apply the entire repertoire of rules designed for the English language, which consists of 4,864 different elements. The tool systematically determines the sentences and splits the tokens in each text. Through a matching procedure it identifies the parts of speech corresponding to the violations covered in the rules (Miłkowski 2010).

After this process, we counted a total of 6,022 different rules' violations. However, each level of competence presented different typologies and total numbers of detected errors according to the defined set of rules. For this reason, we conducted a more in-depth analysis of error percentages and their typologies described later in this section.

On the one hand, we compared the number of errors detected by LanguageTool along the different levels with those identified by the examiners. Our findings are represented in the following scatter plots.

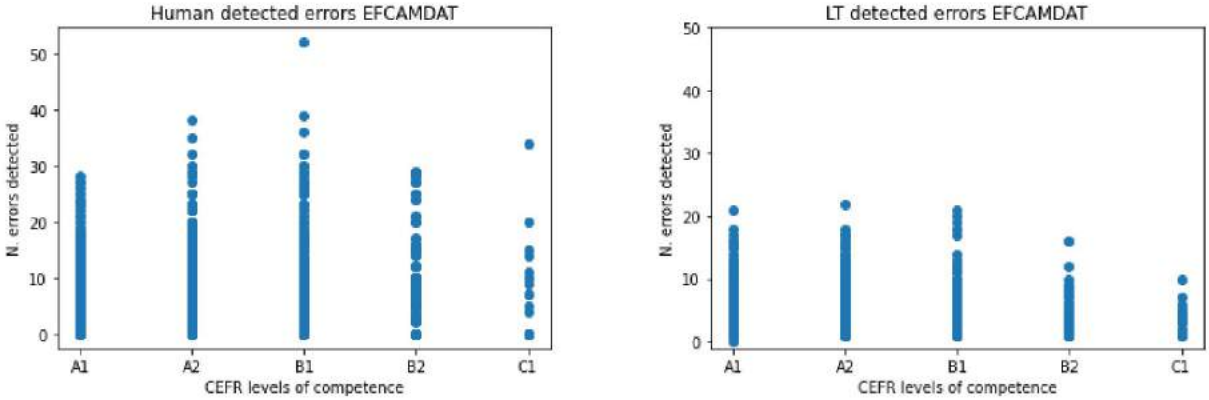


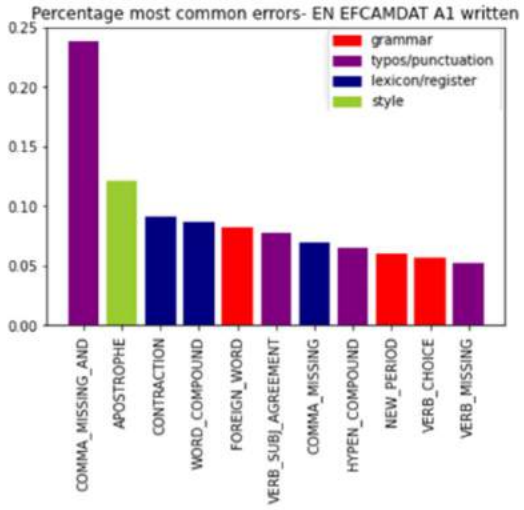
Figure 14: EFCAMDAT human (left) vs LanguageTool (right) detected errors per level

As can be seen in Figure 14, the number of errors detected by humans is up to a maximum of more than 50, while those automatically extracted by LanguageTool are generally less, although distributed in a more orderly pattern. This indicates that LanguageTool is probably more systematic and coherent in the annotation of errors, so much so that it is possible to identify a trend whereby generally the lowest proficiency levels, on the y-axis, exhibit more errors than the others. In the case of human annotations, on the contrary, we can notice that, for example, level A1 presents fewer errors than levels B2 and C1, so that with these annotations alone assessing the learners' competence would not be feasible.

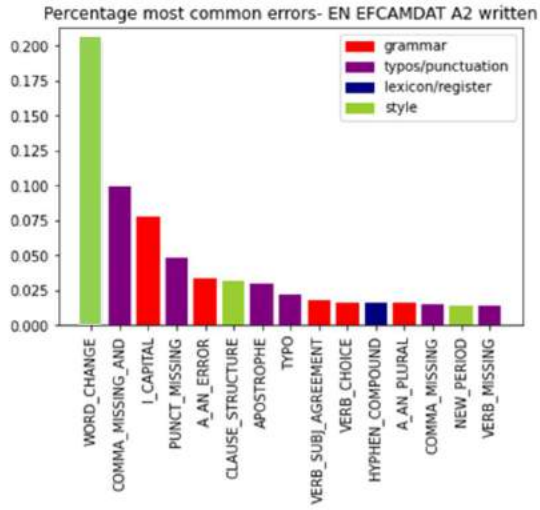
On the other hand, we considered in more detail the errors identified by LanguageTool for each level of competence. We then created macro-categories within which to place them according to the linguistic aspects they concerned. These are the following:

- grammar,
- punctuation and typos,
- register and vocabulary,
- style.

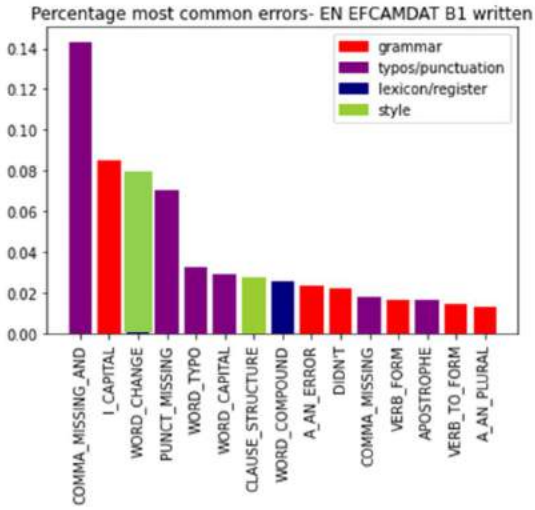
The figures below display histograms containing the percentage of major errors found in the A1 (1), A2 (2), B1 (3), B2 (4) and C1 (5) levels.



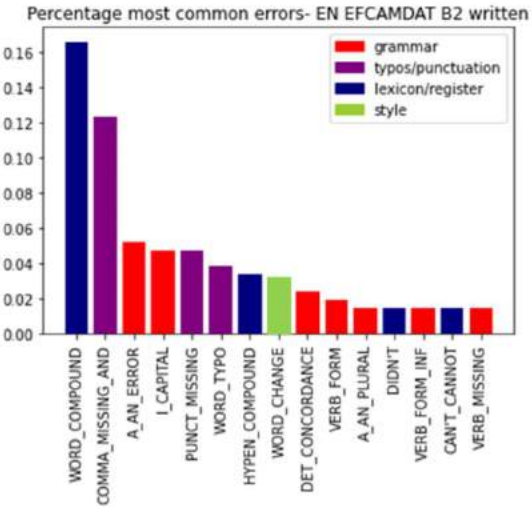
1) A1



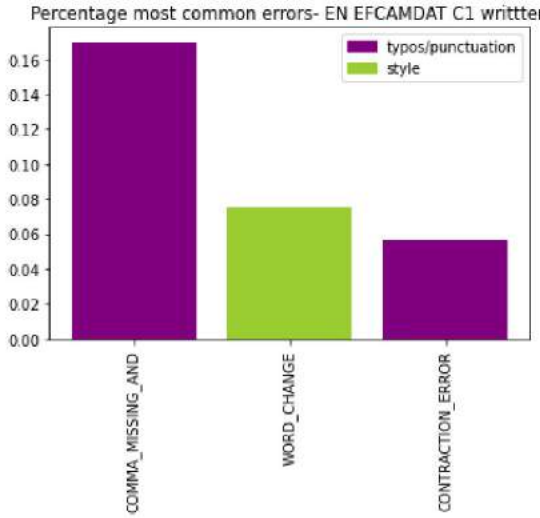
2) A2



3) B1



4) B2



5) C1

Figure 15: Automatically detected errors in EFCAMDAT levels of competence

As can be noticed in the A1 level, the highest percentages of errors concern punctuation, apostrophes and contractions. Grammatical errors, such as the use of foreign words (code-switching) and the need to separate new sentences due to their length, are slightly less common. These types of errors could be traced back to the L1 influence of the learners and the not fully acquired L2 language system during the *interlanguage* phase (Ullman 2005), including elements of style and grammar. At the next level, A2, the most common errors still concern style and punctuation, i.e. the repetition of the same lexical elements in subsequent sentences, the lack of commas and full stops, or spelling mistakes regarding the first person singular. These could be equally attributable to the initial processes of foreign language learning and acquisition.

For the intermediate proficiency levels, namely B1 and B2, on the other hand, the percentages of errors of style and grammar decrease, especially in the latter, while the most frequent are related to vocabulary, together with typos and punctuation. Among the lexical errors, those concerning compound words are often encountered, while among those in the punctuation category, the serial comma before the "and" conjunction persists, especially in the case of two consecutive independent phrases. The latter, however, rather than an error represents a warning inviting the user to double-check whether or not a comma is necessary.

The categories of errors found in the Advanced level, C1, decrease as do their percentages. The most common, as can be observed from the fifth bar plot in the Figure, are punctuation errors, again relating to the "and" conjunction, as well as stylistic suggestions concerning the use of synonyms.

On the basis of the findings regarding the categories of errors contained in the EFCAMDAT corpus, it would seem plausible to assume that since their types and quantity change in accordance with the proficiency level, they represent potential indicators to be considered in the evaluation of learners' competences.

Afterwards, we conducted a more detailed analysis of the linguistic features contained in the learners' texts. In the latter, we analysed the lexical richness, applying TTR, HD-D and MTLD (McCarthy & Jarvis 2010) measures, and the syntactic complexity, considering the average sentence length in words and the dependency distance, together with the number of dependents per word unit. As a first step we tokenized, lemmatized and parsed the texts and stored the results using the *CoNLL-U* format. This allows us to analyse word form, lemma, PoS tags, morphological features and dependency graphs linearly and contemporarily on a single file. The latter files were used with *textcomplexity*, providing a window of 50 words per metric, to extract the subsequently illustrated values.

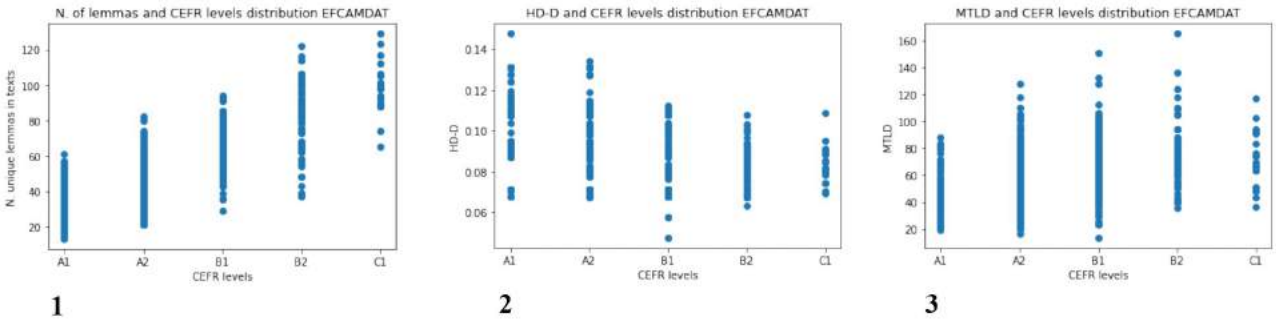
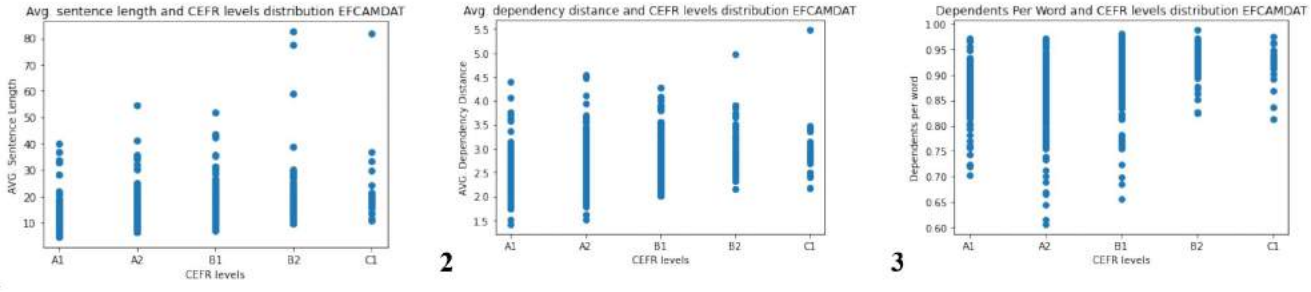


Figure 16: Unique lemmas in text (1), HD-D (2) and MTLD (3) divided per level (EFCAMDAT)

As can be seen from the scatter plots in Figure 16, similarly to our expectations, whereby in theory more competent individuals would use a higher number of different lemmas, it is possible to notice that there is an increment in the amount of unique lemmas (1) as the level of competence increases. On the other hand, considering a metric such as

HD-D (2), similar to type token ration (TTR), except that it does not consider sequential segments but random word samples, we notice that the highest values of lexical diversity are contained in texts of lower levels (ex. A1 and A2), contrary to our expectations. However, the different types of assignments for each level (see Table 2) and the fact that this metric is effective mostly on literary texts should also be taken into account. In contrast, the results of applying MTLD to measure textual lexical diversity starting from the text length in number of words and segments, are rather predictable, meaning that higher values tend to be assigned to better proficiency levels, given that the students' answers were supposed to respect a minimum text length. An exception is class C1, for which however there were fewer samples than for the rest.

Moreover, we also aimed to measure the syntactic complexity of the learners' texts, therefore we calculated the average sentence length, the average dependency distance and the number of dependents per word.



1 2 3

Figure 17: Average sentence length (1), average dependency distance (2) and dependents per word (3) in each level (EFCAMDAT)

From the plots in Figure 17, we note that the average length of the sentences (1) does generally indeed increase as the students' level of proficiency increases. This trend, nonetheless, is not as visible if we consider the average dependency distance (2) as a measure of lexical complexity, since at A2 level we find some outliers that exceed B1 values. Similarly, the number of dependents per word (3) reveals that rather high values are reached even at the lowest proficiency levels. In spite of this, however, higher intervals are noticeable only for the advanced proficiency levels, namely B2 and C1.

5.1.2. CLC- FCE errors and linguistic features

With the other corpus of English language examinations, namely CLC-FCE, we proceeded similarly to EFCAMDAT. Although the sample of learners is different (see Figure 2 and Figure 4) we expected to find comparably related results. Therefore, we first counted the errors detected by LanguageTool, compared the various levels and then more closely analysed the errors for the different categories previously defined. After that, we compared the errors reported by human examiners with those detected by LanguageTool. Finally, we considered in more detail linguistic features such as lexical richness, vocabulary variety and number of sentences produced.

As a first step we calculated the total number of unique violations present in the dataset. Considering the 2,467 examinations contained in CLC- FCE, we counted 1,713 violated rules. Nevertheless, given the unequal distribution of the number of examinations for the different categories of competence, they had to be considered in more detail within each level. However, before this was done, since this corpus also contained human-corrected

versions, we decided to compare the number of errors extracted automatically with the number of errors annotated manually.

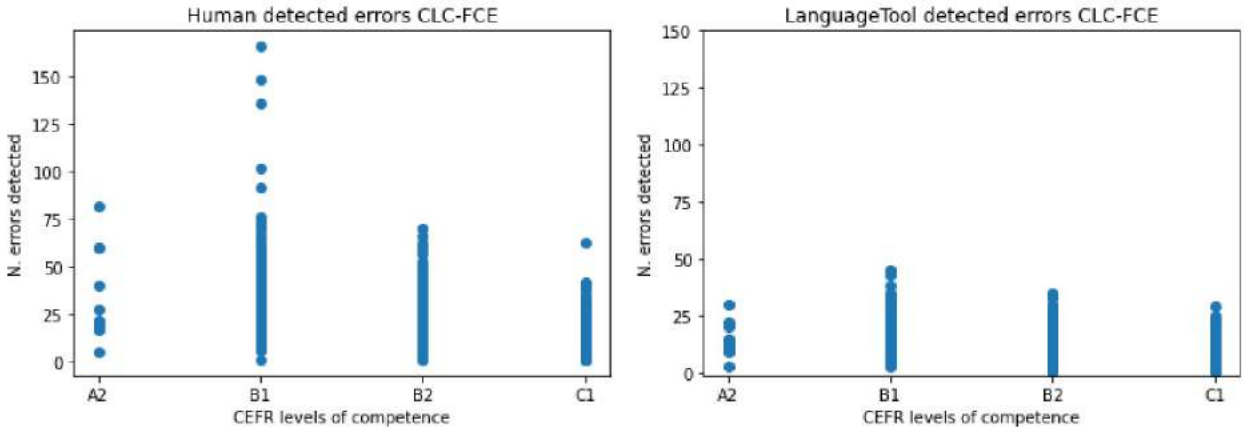
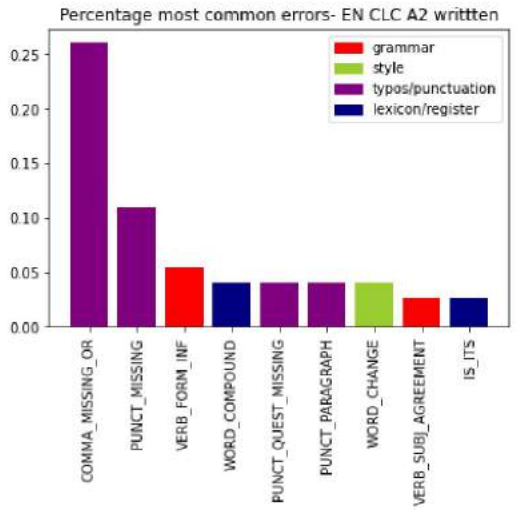


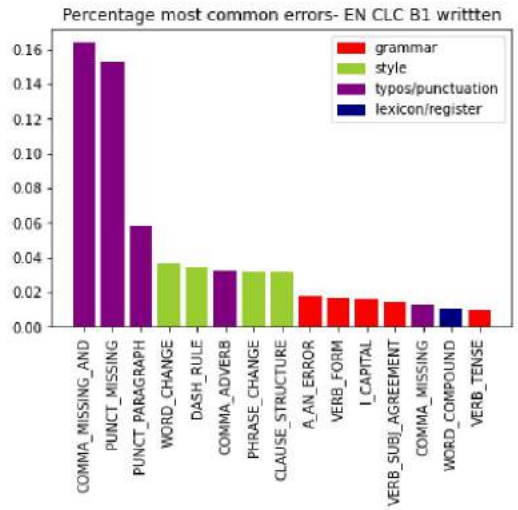
Figure 18: Correlation human (right) vs LT (left) detected errors in CLC- FCE levels

As can be seen from the bar plots shown in Figure 18, the number of human errors, on the left, exceeds the number of errors detected automatically by LanguageTool, on the right, by more than 100. Despite of this, in both cases we can observe that the number of annotated errors tends to be inversely proportional to the level of competence of the learners. The latter can be considered a valid indicator of both the linear progress for the better of language acquisition and the effectiveness of LanguageTool in this correction task, although it appears to be less detailed than the expert evaluators. Nevertheless, the extra advantage of LanguageTool in respect to human evaluators is the consistency in annotating errors and their precise categorisation.

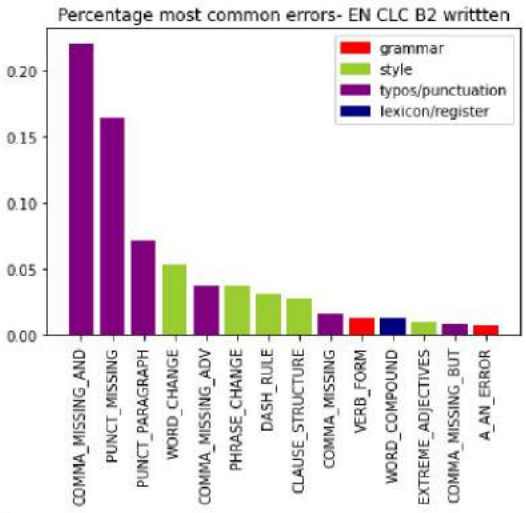
Subsequently, we carried out a more in-depth analysis dedicated to the classification of the automatically detected errors per competence level, for which LanguageTool provides details explaining the potential cause and suggesting a correction. Below are the bar plots for each level.



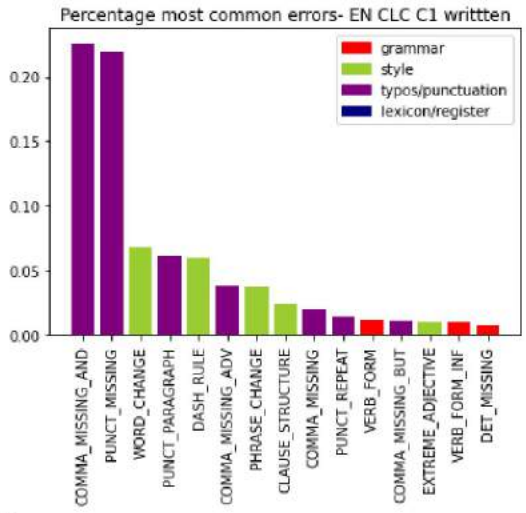
1) A2



2) B1



3) B2



4) C1

Figure 19: Automatically detected errors in CLC- FCE levels of competence

As can be read in the plot displaying the errors in the A2 level (1), the maximum percentage of errors lies around 25%. Among these, the most frequent are punctuation errors related to missing punctuation signs like commas, followed by grammatical, lexical and stylistic errors at around 5%. On the other hand, if we look at the second plot for the B1 level of competence, we notice that the total percentage of errors is lower, about 16%, and that the highest number of errors, apart from punctuation, is related to style, such as sentence structure or searching for synonyms. Due to their higher proficiency, learners display fewer grammatical and lexical errors ranging from 2% or below. This decrease of certain types of errors is even more noticeable in the subsequent levels, i.e. B2 (3) and C1 (4). For these, the highest percentage of errors is still related to punctuation, possibly due to errors in transcription of the exams in digital format or to the mode of the test. On the other hand, the category of stylistic errors features remarkably low percentages, especially with regard to paraphrase suggestions or vocabulary adjustments, while grammatical errors are likewise greatly decreased, becoming almost imperceptible in relation to the total. Especially if we consider what type of errors are involved, they are minor and relate for example to the use of the determiner "an" rather than "a" in front of words starting with vowels, or to the verbal form to be used with the second verb when dealing with particular verbal expressions. These nuances of language are acquired with time and after extensive use of the learned language.

After this attentive analysis, we proceeded to consider the language features related to lexical variety, vocabulary richness and structural complexity of the texts. In order to accomplish this, we used the *textcomplexity* toolkit, which, starting from *CoNLL-U* files, as we did with EFCAMDAT, extracts various relevant numerical metrics, such as MTLD, HD-D, average sentence length and average dependency distance, as well as the percentage of dependents per word. In this case, we have used the arbitrary system described in Table 8 to discriminate between the different levels of competence, so the results could appear somewhat biased.

The total number of tokens extracted via Stanza amounts to 525,106, however, in order to avoid an influence of the text length on the analysis we considered the number of unique lemmas plotted against the levels of competence (see Figure 20.1). Additionally also the HD-D (2) and MTLD (3) are plotted in the figure below.

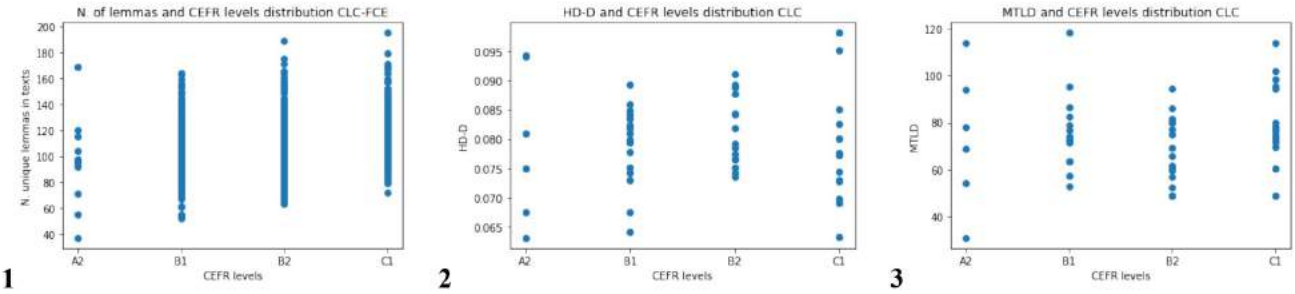


Figure 20: Unique lemmas in text (1), HD-D (2) and MTLD (3) divided per level (CLC- FCE)

As can be noticed from the plots, there is indeed a correlation between the CEFR levels and the number of unique lemmas per text. Apart from a few outliers, in fact, the quantity of lemmas increases as the learners' proficiency increases. Considering, instead, the values obtained with *textcomplexity* (Figure 20.2 & 3*)*, we note that in spite of the small number of texts that the system was able to analyse and some outliers the values increase as the level increases. The trend, however, unlike EFCAMDAT, which contains more numerous samples, is not particularly noticeable. This is partly due to the scarcity of data and partly to a mismatch of the initial labels in the corpus.

The observations made above turn out in line with the results obtained when considering the syntactic complexity of the learners' texts. In fact, when observing the average sentence length (see Figure 21.1), we notice very different values between the levels, with A2, B1 C1 having a longer length than B2. Similarly, some A2 level texts exhibit a higher average sentence length (2) than B2 level ones, while the remaining two levels, namely B1 and C1, display similar values. However, this issue besides being related to the level of competence, concerns the learner's writing style and the assigned task.

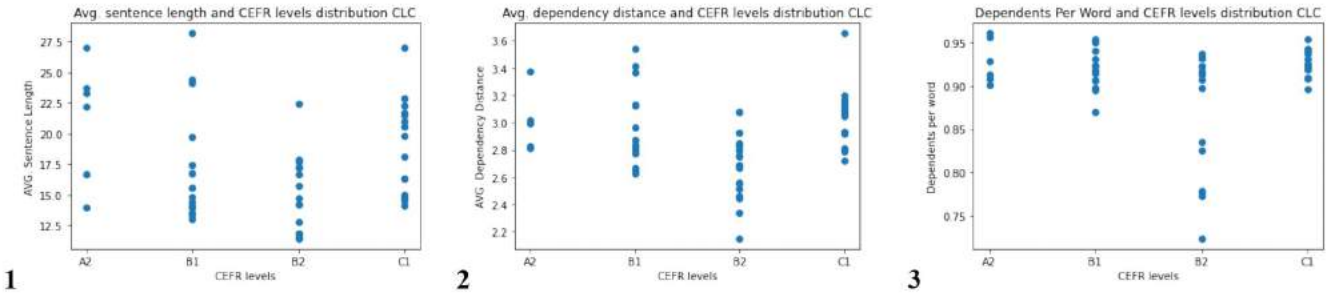


Figure 21: Average sentence length (1), average dependency distance (2) and dependents per word (3) in each level (CLC- FCE)

Finally, considering the percentage of dependents per word (see Figure 21.3), we notice that this is quite heterogenous for the B2 level, while for the others it is concentrated around values between 0.90 and above. For this reason, we cannot consider this metric applicable to these texts.

Overall, the results of this analysis reveal that it is not extremely precise to apply metrics used for literary or native speaker texts to learner texts expecting clear results. Above all, considering the former, in the case of a limited and unbalanced dataset, such as CLC-FCE, the differences between the different levels may possibly appear hardly distinctive.

5.2. Italian corrections

Similarly to what we did for the English language, we proceeded to the language analysis of the MERLIN dataset concerning Italian written examinations. First, we considered the 813 exams available and excluded from this set 7 cases for which there was no CEFR level reported. Following this, we passed these written exams, relating to emails, letters, essays which the different learners had to write, to the automatic checker. By means of LanguageTool we were able to revise the original texts and extract the number of errors found by taking into account the 147 rules recorded in the system. Once these values were obtained, we compared them with the errors identified manually by examiners. After that, we analysed in more detail the errors automatically extracted by LanguageTool for the different proficiency levels and classified them. Finally, we considered the quantifiable linguistic aspects using the same metrics illustrated at the beginning of the chapter.

5.2.1. MERLIN Italian errors

The first step for our analysis, after discarding invalid dataset elements, was to process the texts by means of the automatic spelling and error checker. Before doing so, however, we have normalised the texts to avoid punctuation or typing errors due to incorrect annotation of the original text. This was necessary because the examinations taken in question were originally held in person on paper and then transformed into digital format. Afterwards, as we also had the annotators' corrections available, we decided to compare the number of errors detected by LanguageTool with those determined by humans.

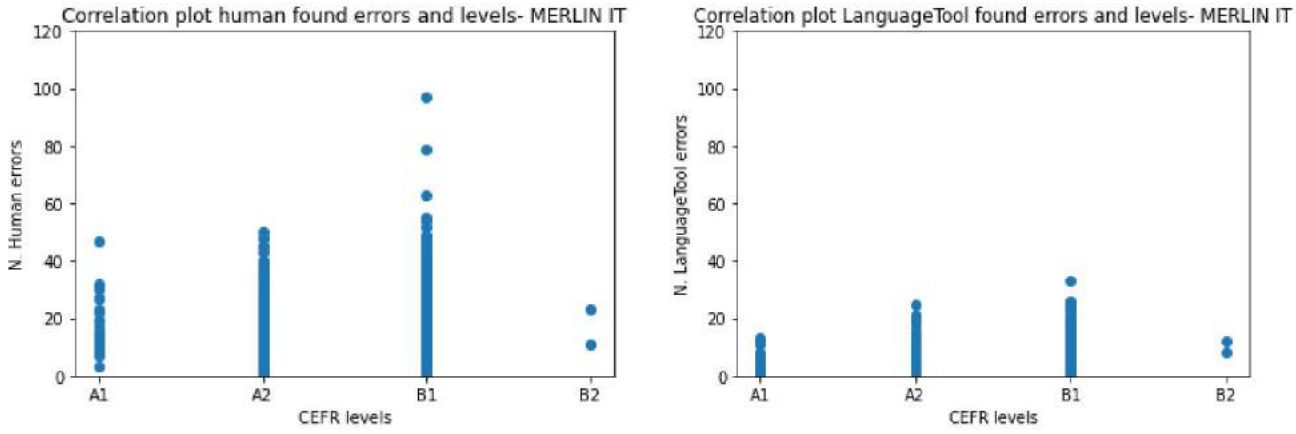


Figure 22: Correlation human (left) vs LT (right) detected errors in MERLIN Italian levels

As can be observed from the scatter plots in Figure 22, the values reported by LanguageTool are about a hundred units lower than those found by human evaluators. Moreover, our expectations were to tendentially find more errors for the lower-level texts and not the contrary, despite their being different in type and nature. Indeed, those at A1 and A2 level, for example, according to both the examiners and LanguageTool made more errors than those at B1 level. Nevertheless, as we will see later, the automatic correction system at times detects not absolute errors, but possible mistakes for which it suggests an explanation and an improvement.

For the first three levels, namely A1, A2 and B1, an upward trend is noticeable, although we would have expected a descending trend concerning the quantities of errors. In this case it would seem that, unlike English, we cannot

employ this metric of the number of errors to discriminate between the different levels of proficiency of Italian language learners. After all, what is relevant is not merely the quantity of detected errors but above all their nature and severity in relation to the expected level of competence.

Following the first part related to the analysis of the number of errors, we moved on to the detailed study of errors and their typologies divided by proficiency level.

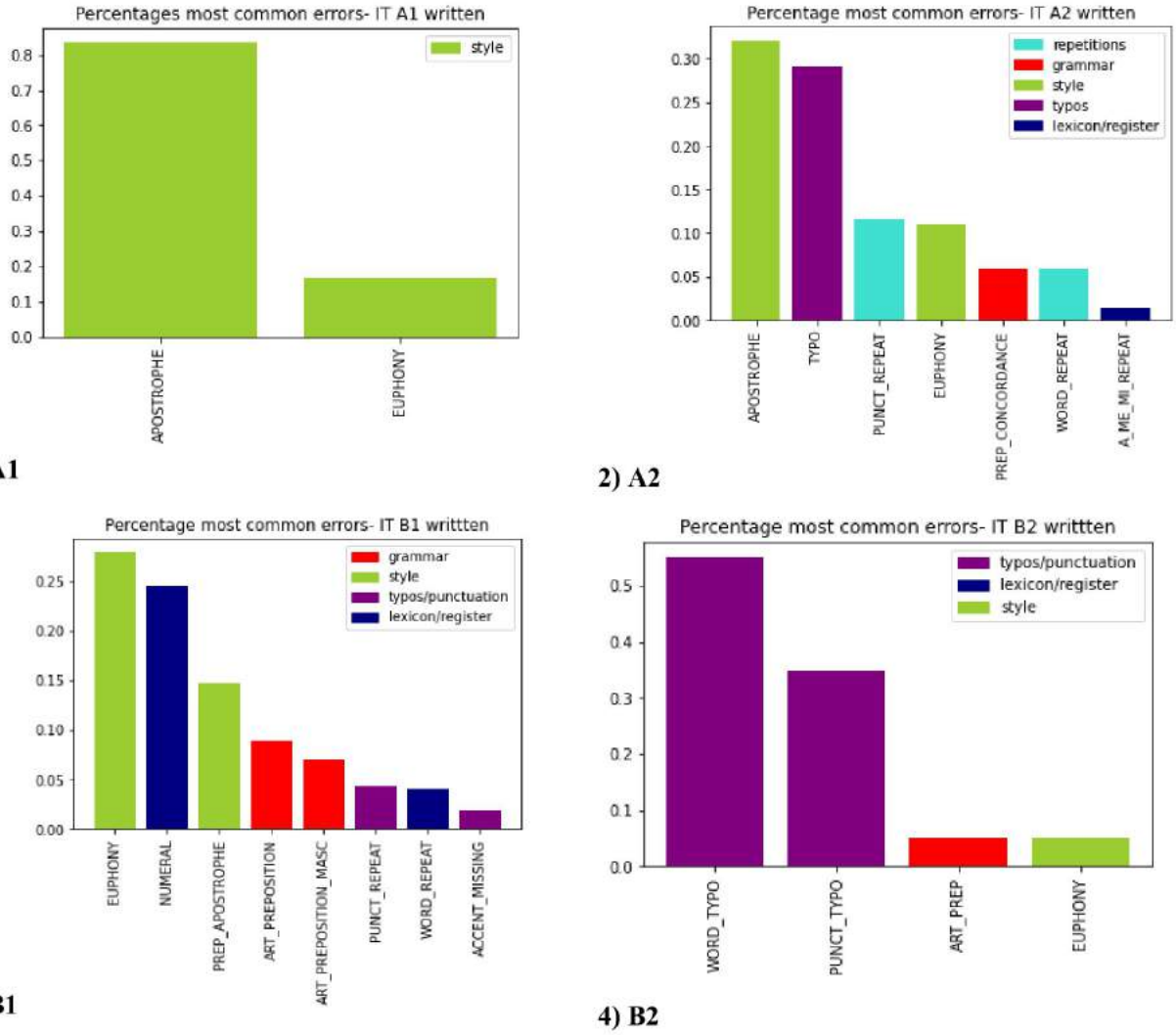


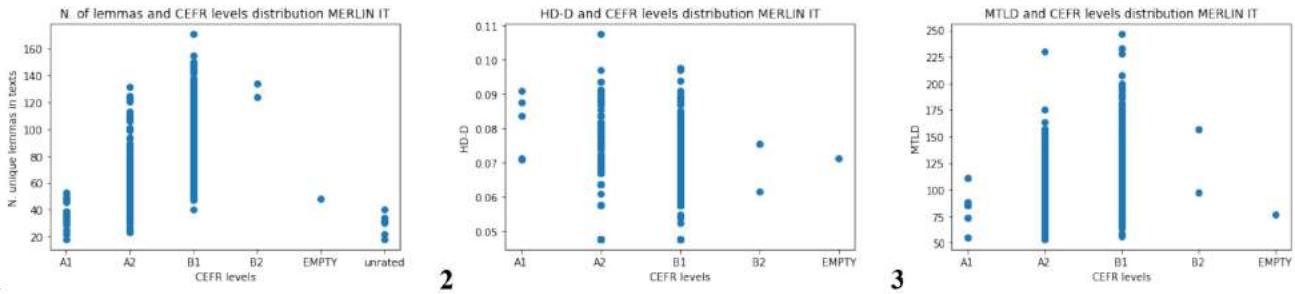
Figure 23: Automatically detected errors in MERLIN Italian for levels of competence

As evidenced in the figure above in the case of the A1 level (Figure 23.1), LanguageTool did not detect a high number of errors related to the areas of grammar, vocabulary or typos, possibly because learners made the best use of their limited knowledge or due to the restricted ruleset available in the checker. Still, we can note that the percentages found lie around 80% and concern two main errors, the lack or incorrect use of apostrophes and the possibility of improving the clarity of the language, for example by avoiding cacophonies. On the other hand, in A2 (2) level examinations the percentage of total errors drops to 30% and these are divided primarily between style errors, again related to the use of the apostrophe and typos of misspelled words, as well as repetitions of words or punctuation, presumably due to annotation errors.

Differently, for the intermediate levels, i.e. B1 and B2 (see Figure 23.3 & 4), the found errors mainly relate to punctuation and style. In particular, in the B1 level, it is common for LanguageTool to report possible improvements regarding euphony, the use of apostrophes and the writing of numbers in place of digits in words. Along with these, minor grammar-related red errors concerning complex aspects of the language, such as the incorrect use of the apostrophe with masculine nouns or the proper placement of the articulated preposition have been detected. On the other hand, typos occur more frequently at B2 level than anything else, confirming a likely more advanced acquisition of the target language by the learners.

5.2.2. Italian extracted linguistic features

Throughout the analysis of language features, we have focused on those language aspects that can be readily quantified, i.e. lexical richness, vocabulary range, and syntactic complexity contained in the learners' written texts. Using *Stanza*, we tokenized, lemmatized and parsed the texts of the language learners. After processing them, we passed the content stored in *CoNLL*-U files to *textcomplexity* and extracted the values for HD-D, MTLD, average sentence length and dependents per word.



1 **2** **3**

Figure 24 : Unique lemmas in text (1), HD-D (2) and MTLD (3) divided per level in MERLIN Italian

From the scatter plots in Figure 24 above, it is possible to observe that we also included those texts without a corresponding CEFR level, in order to attempt to assign them to a level according to this linguistic examination. Given the limited number of texts in each proficiency category, however, it is rather unclear whether or not there is a trend in line with what one would expect. That is, we would normally observe higher proficiency subjects using considerably more words, or to measure lexical variety, more lemmas compared to those with lower proficiency. For instance, this would apparently be the case for the A1 level in comparison with the A2 level, whereas this is hardly noticeable when considering the B1 and B2 levels, especially given the reduced number of texts for this latter class. Additionally, empty and unrated examinations would apparently on the base of this first investigation fall into the lower proficiency class.

On the contrary, considering the HD-D (see Figure 24.2), i.e. the variety of vocabulary, we observe that the absolute highest values are at A2 level, whereas the higher proficiency levels such as B1 and B2 display values below 0.10. Still, these are minor differences and not particularly significant in such a small group of samples. Differently, in the third plot for lexical diversity (3), or MTLD, we see increasing peaks up to the B1 level, as we would expect. However, the B2 level and the empty class, given the few samples, appear not contribute significantly to the analysis.

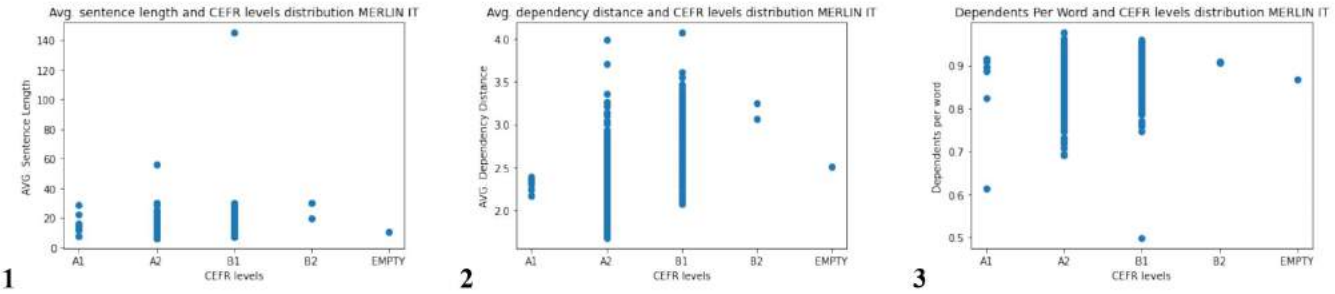


Figure 25 : Average sentence length (1), average dependency distance (2) and dependents per word (3) in each level (MERLIN Italian)

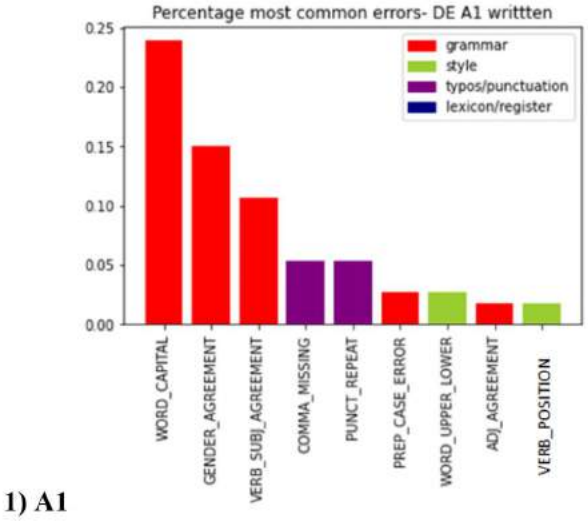
Finally, considering aspects related to the complexity of syntax in texts, we note from the plots in Figure 25 that, indeed, the average sentence length (1) would seem to increase with level. However, given the limited data for the B2 level and in general for the other classes, the results do not appear very precise. On the contrary, in the other two cases referring to the dependencies, first to the distance between them (2) and then to their quantity per word (3), an increasing trend can be observed for levels A1, A2 and B1. Once again, though, the last two classes, namely B2 and empty, seem to be insufficiently informative. Taking into account the restricted number of examinations and their unbalanced distribution in CEFR classes, we can consider this language analysis to be in line with expectations. Definitely, a larger dataset would likely reveal more robust results.

5.3. German corrections

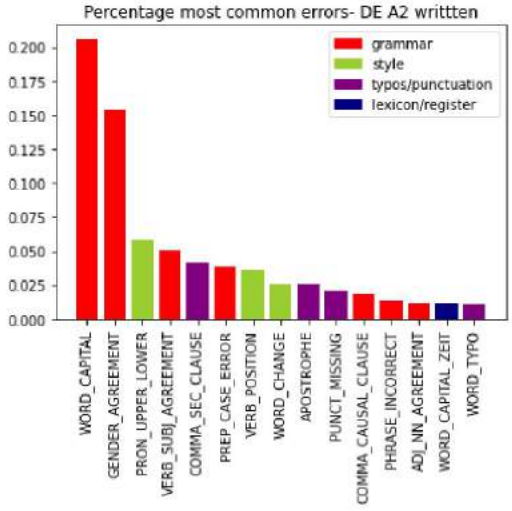
The same procedures as for English and Italian were used to handle the MERLIN German dataset, containing 1,033 examinations distributed over all six CEFR proficiency levels in an unbalanced manner. Thus, we processed the original texts with LanguageTool, which applied the 3,433 pre-established rules to them following tokenization and separation into Parts of Speech. After that, since no human corrections were available in this case, we moved on to categorising the errors automatically identified in each proficiency level. Finally, we considered the linguistic aspects related to vocabulary and discourse complexity.

5.3.1. MERLIN German errors

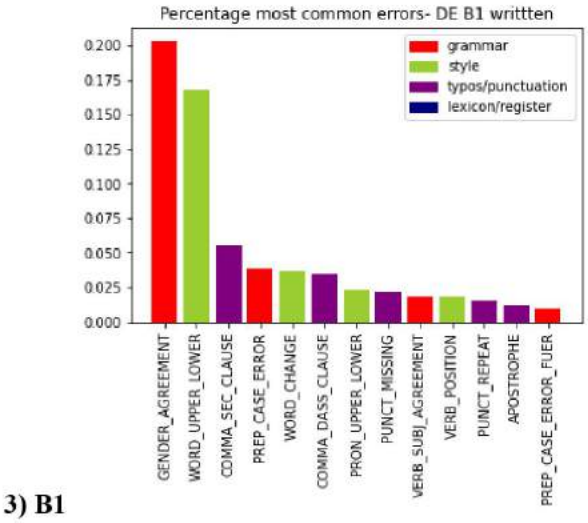
Considering that the German language is mainly made up of norms that differ from the other languages in terms of spelling, as well as lexical and structural rules, the errors found by LanguageTool turned out being various and more specific than in the other languages.



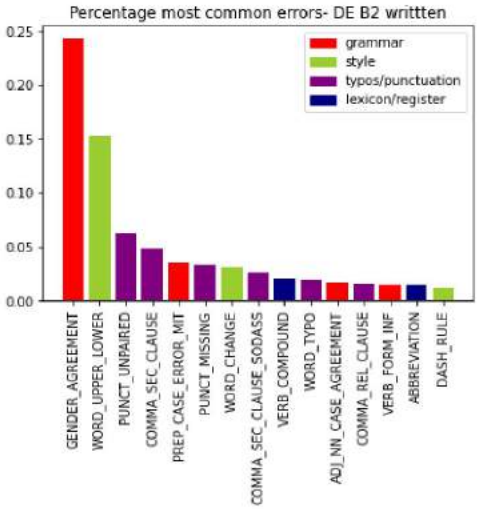
1) A1



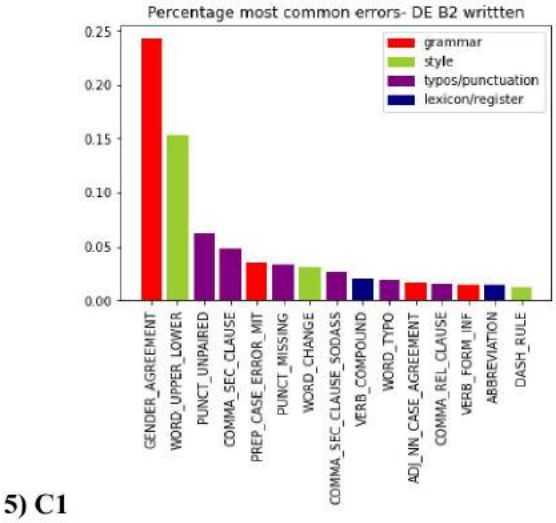
2) A2



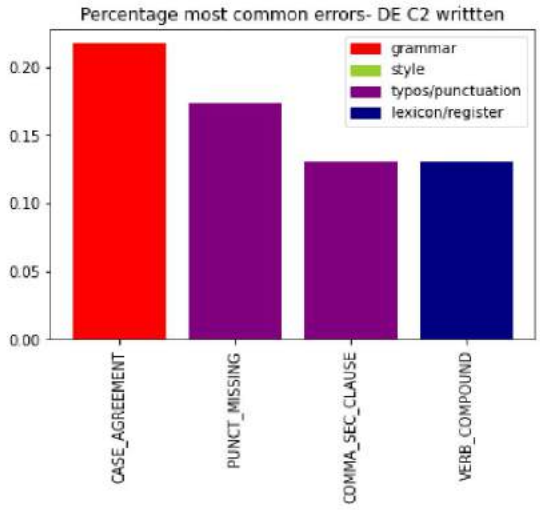
3) B1



4) B2



5) C1



6) C2

Figure 26: Automatically detected errors in MERLIN German levels of competence

For example, as can be noted from the bar plots above relating to the beginners' proficiency levels, that is A1 and A2 (see Figure 24.1 & 2), the largest proportion of errors concerns red errors, that is grammatical errors. Within this category we find cases where words lack capitalization and where gender agreement or subject-verb agreement are violated. Additionally, the A1 level is dominated by errors associated with punctuation, such as the lack of

commas between the main and secondary sentences. In contrast, the second most frequently encountered errors in the A2 level are related to style and concern suggestions rather than errors about the possible introduction of synonyms, changes in verb position or requests to check the used pronouns, like "Sie" and "sie".

Turning to the intermediate proficiency levels, namely B1 and B2 in the third and fourth plots, the most common errors involve both grammar and style, resulting in both cases firstly from incorrect gender agreement between determiners or adjectives and nouns and secondly from suggestions to check the form of capitalised or lowercase words. The remaining errors, present to a lesser extent, concern the use of the comma in secondary sentences, verb forms, typos and compound nouns. Overall, progress in competence is noticeable compared to previous levels, although for German language learners, especially for those who are not L1 speakers within the Germanic language family, certain elements remain somewhat more complex to be internalised and implemented.

Finally, regarding the texts of advanced learners (see Figure 26.5 & 6), we notice first that the percentage of errors detected by LanguageTool is lower than before, between 17% and 20%. Furthermore, the categories of errors are mainly restricted to grammatical and punctuation errors. In the case of the C1 level the most common error is due to lack of gender agreement, while for the C2 it is linked to incorrect case agreement. On the contrary, less common errors concern commas, abbreviations and suggestions in the use of nouns. Thus, we can conclude that these corrections indicate a higher level of competence in the latter group of learners, despite the reduced number of tests available for the analysis.

5.3.2. German extracted linguistic features

For the specific linguistic analysis of the texts, as we did for the other languages, we tokenized, lemmatized and parsed the original texts. We then used this information to measure the lexical richness and density, as well as the syntactic complexity, demonstrated by the German language learners.

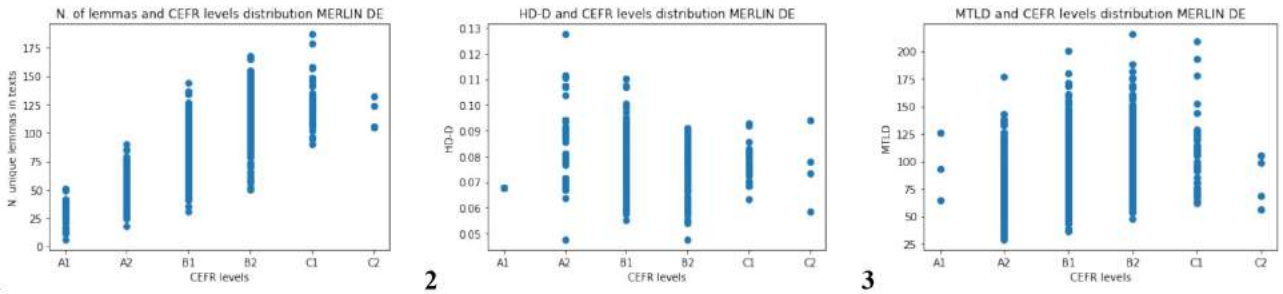


Figure 27: Unique lemmas in text (1), HD-D (2) and MTLD (3) divided per level in MERLIN German

As can be observed from the scatter plots in Figure 27, the number of unique lemmas (1) increases proportionally to the level of competence of the learners. In fact, for example, the number of lemmas in A1 level texts reaches a maximum of about 50, while in C1 level learners it reaches 200. This can, therefore, be considered an index of lexical richness and variety that is acquired after a sustained period of learning, as well as an increasing usage of the target language.

On the contrary, taking into account the vocabulary density of the texts (2), obtained extracting the average TTR of random portions of them instead than of their entire length, which may influence the results across different levels,

we can notice an irregular trend across the six different CEFR levels. This, together with the minimal values distinguishing them, signals that probably this type of measure does not apply to the German texts of this dataset. However, the same cannot be stated for the MTLD measure, as, instead, the plot (see Figure 27.3) mirrors our expectations. Higher levels of competence exhibit higher values, with the exception of the texts belonging to the C2 level, for which, however, we disposed of a very limited sample.

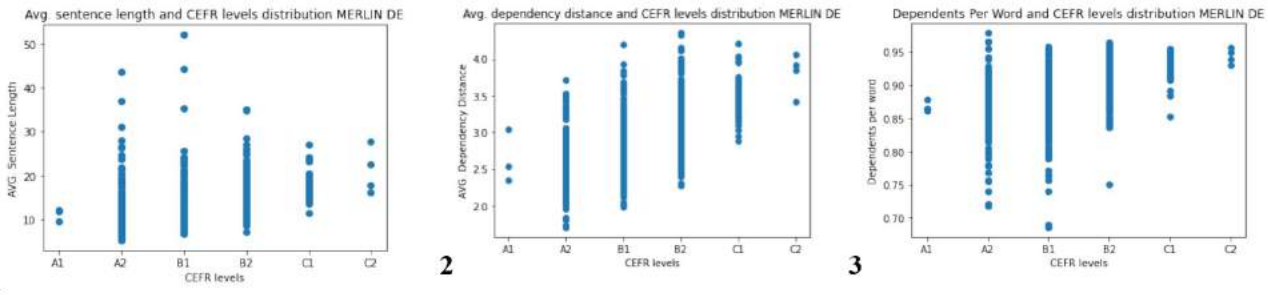


Figure 28: Average sentence length (1), average dependency distance (2) and dependents per word (3) in each level (MERLIN German)

Finally, analysing the syntactic complexity of the texts and calculating the average sentence length, dependency distance and quantity of dependents per word, we notice different outcomes. In the case of the average sentence length (see Figure 28.1), we observe an ascending trend of the values, indicating the number words, up to level B1, while they decrease from B2 to C2. On the contrary, a clear increase can be noticed with the average dependency distance (2), where up to B2 the values appear greater than 4 and maintain themselves somewhat stable for the levels C1 and C2. Similarly, the last plot referring to the dependents per word, exhibits higher intervals for the higher CEFR levels. This can be translated to the fact that the words used by more competent learners display more dependencies and, indeed, higher textual complexity. These results were, therefore, in line with our expectations.

5.4. Conclusions about the errors and linguistic analyses

In general, considering the analyses and results obtained for each dataset of the English, Italian and German languages, we can discern distinct results regarding the errors and linguistic domains. The most relevant outcomes of this analysis are summarised below.

In the case of English datasets we can consider the distribution of error types as a possible macro-aspect to be taken into account when designing an automatic classifier relating to proficiency levels. With a sharper and more distinctive categorisation by type, it could be assumed that in the lower proficiency levels more grammatical and stylistic errors are found, while in the intermediate proficiency levels these concern vocabulary and punctuation, whereas in the higher proficiency levels these relate to minimal possible improvements in style and punctuation. Furthermore, in support of the classifications of the different levels of competence, we found results in line with the linguistic investigations carried out, especially with regard to the number of unique lemmas, MTLD and sentences' length in the different texts. On the other hand, we observed much weaker correlations between HD-D and average dependency distance in the examinations' texts. This probably relates to the varied nature and length of the texts , but also to the authors who are not native speakers.

On the other hand, in the case of MERLIN Italian, we observed some outliers between the different texts, especially for the less represented proficiency levels such as B2 or unrated exams. Considering the number of errors, there seems to be a correlation with the proficiency levels, although the quantities varied as well as for the other datasets between those detected by the human evaluators and those determined by LanguageTool. The latter, however, given the categorisation of errors and their precise measurement allowed us in all three languages to attempt to identify potential patterns to be applied to the classification of the tests. However, given the limited number of tests available and the uneven distribution of texts in the different proficiency levels these results cannot be widely generalised for all Italian learners. Also with regard to the number of lemmas and sentences' length, for the most represented levels, i.e. A1, A2 and B1, we found the expected results, whereby the most competent were those who used more linguistic elements in their written texts. On the contrary, even in this case the HD-D and the average dependency distance have not been particularly revealing.

Finally, as far as the German language is concerned, we can state that the analysis of the error typologies allowed us to extract useful information with respect to the linguistic areas in which learners have more difficulties for the diverse levels, especially concerning the grammar norms related to gender and case agreement. Even when considering the values of quantified unique lemmas and word dependents, we obtained similar results to English and Italian, whereby the quantities of them increased as proficiency improved. Yet again, this was not very clear for the least represented categories, such as C2 with less than ten examinations.

Overall, we can conclude by stating that this analysis prior to the use of the automated neural architecture for measuring proficiency in texts has brought to light some important and some problematic aspects, including the necessity of precise quantification and categorisation of errors and language elements content, as well as the presence of underrepresented groups of learners in the different datasets. Undoubtedly, having methodical and precise measurements, such as those of LanguageTool, renders classification more straightforward and clearer, as the principles are objective and not biased or alterable. Yet, it must be kept in mind that language learning and acquisition can be conditioned by a set of factors including input, teaching, methods used, amount and frequency of use. These are also combined with individual differences, such as prior experience of language learning and subjective neurocognitive properties, such as memory. Adult learners, one should not forget, find it cognitively more demanding to learn the elements of a language other than L1 precisely because there is a language system already established in those learners (Ullman 2005). For them, it is mainly a matter of memorising and internalising abstract rules and structures, which in the case of inflected forms and idiosyncratic features are more demanding. The methods we used basically start from the linguistic rules and their violations by the learners, so they well represent the basic principles for a proficiency assessment in adult language learners.

CHAPTER SIX: EXPERIMENTS FOR THE AUTOMATIC ASSESSMENT OF LANGUAGE EXAMS

Our project intends to work with the various datasets and a basic Transformer model, namely *bert-base-uncased* (see § 4.2.2.) in order to classify the students' examinations within the different categories corresponding to the six CEFR language proficiency levels. However, in order to achieve this objective, we have conducted several experiments, both with the original and corrected students' texts, modifying the parameters and the architecture of the model, adapting it to the data with the aim of obtaining the best possible results.
Basically, we planned to use as model input for both the training and the testing stages the students' original texts and their versions manually corrected by human experts, if available, or the automatically corrected ones generated by LanguageTool. The true labels we passed to the model correspond to those contained in the corpora after human correction or those obtained by mapping the assigned exam scores to distinct CEFR levels. Moreover, to make the model properly function, we had to transform the levels into ordinal qualitative variables expressed first in nominal form and then in discrete numerical values (e.g. A1=0, B2= 3, C1=5). We evaluate the proposed models in terms of the following metrics:

- **Precision**: the ratio of correctly predicted observations for a class considering the total corrected predicted observations;
- **Recall**: the ratio of correctly predicted observations for a class in respect to the total number of elements in that given class;
- **F1-score**: the weighted average of precision and recall, considering both incorrect positive and negative predictions;
- **Accuracy**: the ratio of the total number of correct predictions over the entire sum of considered elements.

The above applies both to the English language content, which will be discussed in more detail in the following two sections, and to the Italian and German language content, which will be covered in more detail subsequently.

6.1. English First Cambridge Open Language Database

The first experiments we conducted consisted of the creation of a BERT-based model using the learners' data from the EFCAMDAT corpus. Given the number of exams available, we considered them sufficient for the complete training and testing of a neural architecture. Therefore, we considered 100,000 exams from the previously created partitions (§ 4.3.1.) as the training set, and the remaining 1,447 as the test set to evaluate the results of the model. As a first attempt, we employed the data described above to train a model that based exclusively on the students' original texts could classify them in the different CEFR levels. Secondly, we transferred the same training parameters into a model that additionally received as inputs the human corrections provided in the original dataset. Lastly, we experimented passing as inputs to a new model, trained according to the same procedures of the previous ones, the original students' answers and their automatically corrected versions provided by LanguageTool.

6.1.1. EFCAMDAT model trained with original students' texts

More in detail, the architecture of the first EFCAMDAT experiment, the components of which have been previously described in section 4.2.2., received the original students' answers to the assignments as inputs. Together with these, we also passed the model the labels corresponding to the CEFR levels which had been assigned to each text. Following the tokenization of the examinations and their transformation into arrays, we established an initial maximum sequence length of 450 elements. The configuration details also include Adam as optimizer, a Stochastic Gradient Descent method based on adaptive estimation projected to be efficient while requiring reduced amounts of memory, especially when dealing with tasks that contain large data or numerous parameters to be considered (Kingma et al. 2014). Additionally, we selected the *sparse categorical cross entropy loss* as parameter to compute the loss between the true labels and the predicted labels since we are dealing with more than two label classes provided as integers. The selected metric to estimate prediction was the accuracy. Moreover, we trained the model for 16 epochs using a batch size equal to 16 and a validation split of 0.1.

Once the model was trained, we proceeded to the evaluation using the test set consisting of 1,447 files. The five logits returned by the final layer of the model represent the CEFR levels within which the texts are classified. By applying the argmax function we obtain the index of the predicted class for each learner text. The following Figures, namely the classification report and the confusion matrix obtained with the Scikit-learn library in Python, display the results on the EFCAMDAT test set.

	precision	recall	F1-score	support
0	0.98	0.99	0.99	697
1	0.98	0.95	0.96	444
2	0.94	0.98	0.96	213
3	0.90	0.84	0.87	73
4	0.90	0.90	0.90	20
Accuracy			0.97	1447
Macro-average	0.94	0.93	0.93	1447
Micro-average	0.97	0.97	0.97	1447

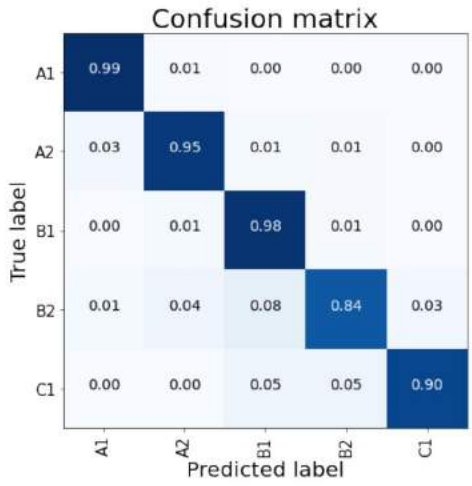


Figure 29: Classification report and confusion matrix on EFCAMDAT test set (original texts only)

As can be seen from Figure 29, the classification is very accurate for all classes, both in terms of precision/recall and F1-score, as well as overall accuracy. The best predictions are encountered for the classes A1, A2 and B1, which were also the most numerous. Nevertheless, also B2 and C1 received accurate predictions ranging from 84% to 90% of accuracy.

Although the training set contained an unbalanced number of items for each class, or language proficiency level, the model trained using exams from the EFCAMDAT corpus managed to classify 96,8% of them correctly. Such results are visibly clear in the confusion matrix above.

6.1.2. EFCAMDAT model trained with human examiners' corrections

Following this initial attempt described in the previous sub-section, we decided to use the corrections of the original texts manually annotated by human examiners to train another model. Therefore, the latter received as input each time two texts, original and corrected one, as visible in the architecture on the right in Figure 10. From these, the encoded arrays are passed to the classifier, the logits of which correspond to the CEFR levels of competence from A1 to C1. The parameters according to which we trained this other model are the same as the ones used in the previous experiment, namely the same batch size, learning rate and maximum text length. Only the number of epochs differs, being in this case 20. Our expectations were to possibly detect an improvement in the classification of students' texts according to the different proficiency levels. Indeed, from our understanding, the model could have benefited from the corrections provided in order to estimate the distinct competences embedded in the textual inputs.

The obtained results of this experiment are embedded in the following classification report and confusion matrix.

	precision	recall	F1-score	support
0	0.98	0.99	0.99	697
1	0.97	0.95	0.96	444
2	0.90	0.97	0.93	213
3	0.92	0.78	0.84	73
4	0.95	0.90	0.92	20
Accuracy			0.96	1447
Macro-average	0.94	0.92	0.93	1447
Micro-average	0.96	0.96	0.96	1447

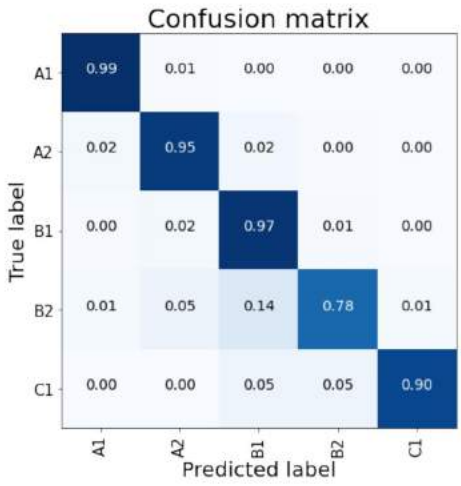


Figure 30: Classification report and confusion matrix on EFCAMDAT test set (human corrections)

The figure above displays an accuracy which is extremely close to that of the previous model, trained with only the learners' original texts. However, a closer observation of the confusion matrix reveals a slight deterioration in the predictions for class B2 and B1 compared to the former experiment. This may suggest that although the model has been supplied with corrections, these do not represent an important additional resource for the resolution of its task. In the following subsection we compare these results with those obtained by passing automatically generated corrections to the architecture.

6.1.3. EFCAMDAT model trained with original students' texts and LanguageTool corrections

Following the attempt previously described using the human corrections of the exams, we conducted a parallel one this time applying the corrections of the original texts generated automatically by means of LanguageTool to train another model. The latter, indeed, got the original students' written assignments and their automatic corrected

versions as inputs. Their encoded arrays were then received by the classifier of the architecture, which, as before, computed the best predictions for each proficiency level class. The training parameters are 30 epochs, learning rate of 2e-5, maximum text length of 450 words and 16 as batch size, equal to the ones of the two previous experiments. The following figure displays the results on the test set by means of the classification report and the confusion matrix.

	precision	recall	F1-score	support
0	0.99	1.00	0.99	697
1	0.98	0.96	0.97	444
2	0.92	0.97	0.94	213
3	0.95	0.85	0.90	73
4	0.95	0.90	0.92	20
Accuracy			0.97	1447
Macro-average	0.96	0.93	0.95	1447
Micro-average	0.97	0.97	0.97	1447

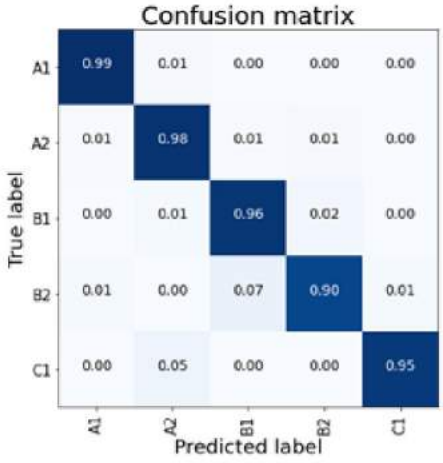


Figure 31: Classification report and confusion matrix on EFCAMDAT test set (automatic corrections)

A detailed examination of the values in the confusion matrix and those of the classes in the classification report reveals slight improvements compared to the model initially trained with the original texts alone. These concern for example class A1, with 99% accuracy, A2 with 96% and B2 with 85%. In all these cases the model has improved by some minimal values, mirrored by the overall accuracy, which corresponds precisely to 97,2%.

The outcome of this experiment suggests that the BERT architecture is likely to be able to classify the texts of non-native learners reasonably well on the basis of proficiency using only the original texts themselves, without any additional information regarding errors or scores given. Nevertheless, if the latter are provided, in the case of methodological and precise individuated features, the results are expected to be better than with manually extracted and highly variable corrections.

6.2. Cambridge Learner Corpus for the First Certificate in English exam

For a further experiment related to the English language, we used the CLC- FCE dataset, which contains 2,469 examinations, considerably fewer than the previous EFCAMDAT dataset. We decided to train respective models using mainly the original leaners' texts, their manual annotated versions and the automatically generated corrections. Therefore, we employed a data partition according to which 2,017 learner examinations constituted the training set, whereas the remaining 194 constituted the test set. We had to exclude 10 original examinations since the information about them were incomplete because they were not provided with an assigned score. We used the same mapping system described in Table 8, considering the decimal scores assigned for the English First exam. Moreover, we provided a maximum text length equal to 380 words and trained the model with a learning rate equal to 2e-4 for a total number of 200 epochs.

Given the imbalanced distribution of examinations and their respective weights for different classes, before the beginning of the training processes we decided to target the issue. To improve class imbalance, since the difference in class frequencies in our case could have affected the overall predictability of the model, we resorted to the *class_weight* parameter available in the Sklearn utils package (King & Zen 2001). This enabled us to automatically compute more balanced weights to assign to each class considering the total number of items in the selected set and their disparity in each class.

The following subsections illustrate the experiments conducted by using first only the original students' texts, then the versions manually revised by language examiners, and finally the versions automatically corrected with LanguageTool alongside the original texts.

6.2.1. CLC-FCE model trained with original students' texts

As a first experiment with this dataset, namely CLC- FCE, we decided to try to classify proficiency levels by providing the model only the students' original texts without any correction. Similar to what we have carried out with the EFCAMDAT corpus, we wanted to examine whether our architecture would be able to perform a multi-class classification using only the original student texts and still achieve significant results. Below are the training curves of the model.

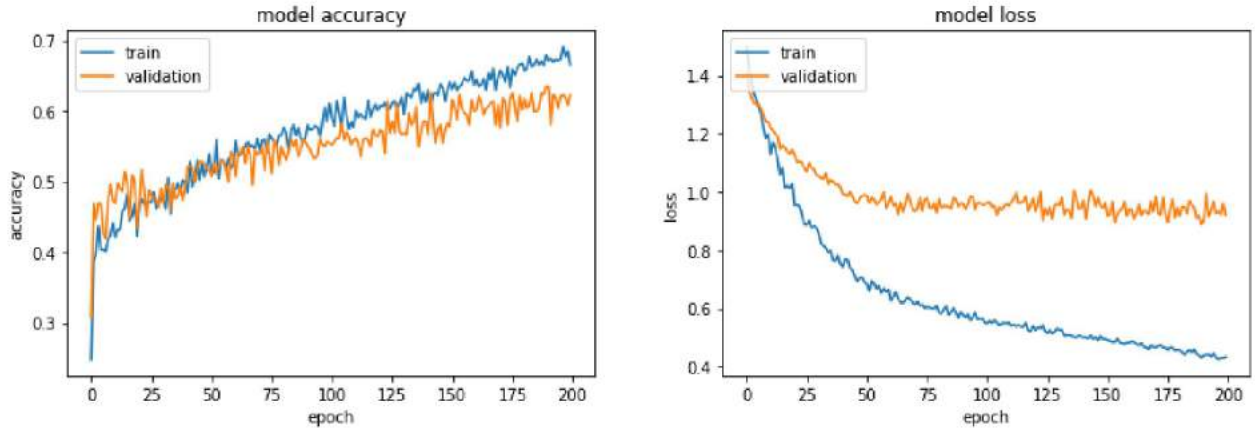


Figure 32: CLC- FCE model training and validation accuracy and loss curves (original texts only)

The accuracy and loss curves appear rather regular, although in the case of the validation set the loss does not decrease as we would expect. Most probably this is due to its under-representation of the dataset classes, since this partition is constituted by only 20% of the training set and corresponds to barely 201 examinations.

	precision	recall	F1-score	support
1	0.00	0.00	0.00	0
2	0.67	0.42	0.51	53
3	0.68	0.65	0.67	101
4	0.47	0.72	0.57	40
Accuracy			0.60	194
Macro-average	0.45	0.45	0.44	194
Micro-average	0.63	0.60	0.60	194

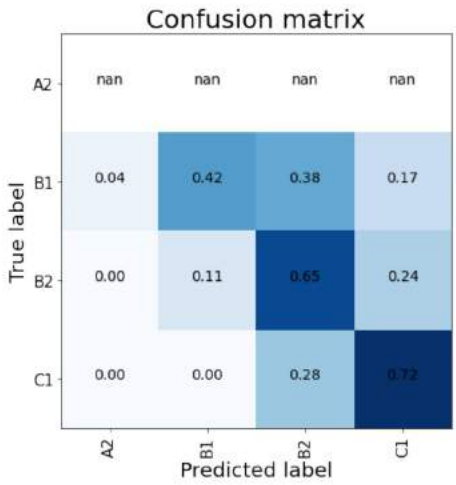


Figure 33: Classification report and confusion matrix on CLC- FCE test results on CLC- FCE model (original texts only)

Attentively observing the classification report and the confusion matrix above, we notice that since the weights of the classes have been re-balanced, for the levels for which texts were available, the values obtained on the F1-score are quite equilibrated. The overall accuracy is **60%** on the four classes in the dataset, namely A2, B1, B2 and C1, obtained after we have arbitrarily matched the assigned scores to the CEFR levels we consider in our project. Moreover, as can be noticed, since the CLC-FCE corpus was created specifically from First Certificate of English examinations, there is no support for the A2 level in this experiment, while the B1 and B2 classes are the most populated ones.

This initial result appears to be discrete, although we must compare the values obtained with those of the subsequent experiments with corrections. In fact, we expect them to contribute to an improvement in the model's classification performance.

6.2.2. CLC- FCE model trained with human examiners' corrections

In addition to the students' original texts, the CLC- FCE corpus also contained a version of these corrected by expert examiners. In order to evaluate the functioning of our automatic correction and classification system, we decided to test a model that received as input the original texts and their human-corrected versions. We employed the exact identical parameters of the model described earlier, that means the same training and evaluation set, the same number of epochs and learning rate, changing only the corrections that were additionally passed to the BERT model.

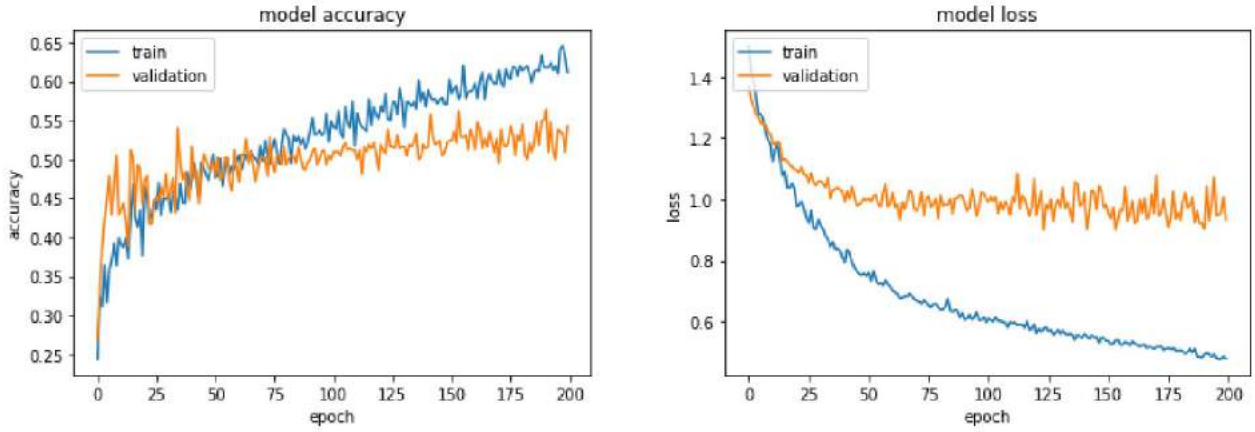


Figure 34: CLC- FCE model training and validation accuracy and loss curves (human corrections)

As shown by the accuracy curve of the training in Figure 34, this lies at approximately 65% . Differently, on the test set, if we observe the figure below, the obtained overall results appear close to those achieved by the model trained with the original students' examinations only. However, after the previous experiment, we proceeded to remove class A2 from those considered for evaluation, since no exams were included for this category in the test set. For this reason, only three distinct CEFR classes appear here.

	precision	recall	F1-score	support
2	0.61	0.51	0.56	53
3	0.65	0.65	0.65	101
4	0.50	0.60	0.50	40
Accuracy			0.60	194
Macro-average	0.59	0.59	0.58	194
Micro-average	0.61	0.60	0.60	194

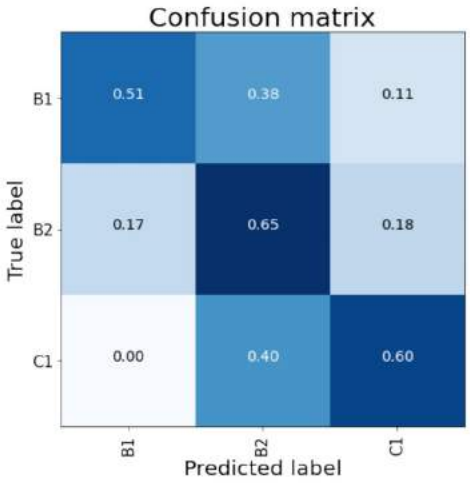


Figure 35: Classification report and confusion matrix on CLC- FCE test results on CLC- FCE model (human corrections)

As we can notice both from the matrix and the values enclosed in the classification report above, the results are in line with those obtained with the previous model (see Figure 33). Indeed, even the total accuracy percentage, i.e. 60%, happens to be the same. The only difference is that the best predicted class in this case was not C1 but B2.

6.2.3. CLC-FCE model trained with original students' texts and LanguageTool corrections

As a final experiment with an architecture trained exclusively with the content of the CLC-FCE dataset, we made use of the automatic corrections provided by LanguageTool on the students' original texts. This means that the model received as initial input on the one hand the students' texts with errors in them, and on the other hand texts automatically generated by the checker with the corrections of the latter. These two, after being encoded, were fed as arrays into the multi-class classifier as in the previous experiment with the human corrected versions.

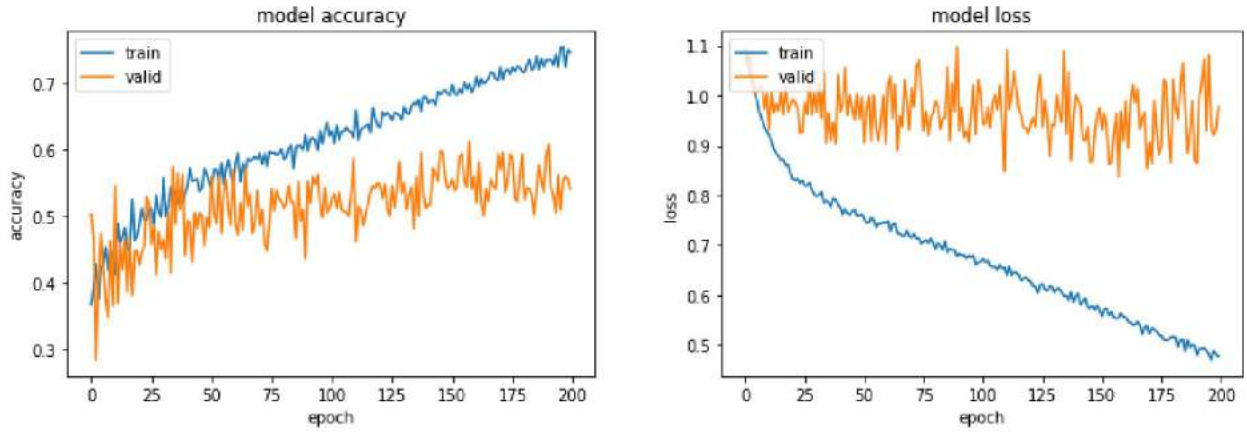


Figure 36: CLC- FCE model training and validation accuracy and loss curves (automatic corrections)

From the training curves above we can notice an accuracy of over 70% on the training set, although the results on the validation set appear low, most likely given by its under-representation of the classes in the dataset since it is made of 20% of the exams contained in the training set.

	precision	recall	F1-score	support
2	0.73	0.42	0.53	53
3	0.65	0.76	0.70	101
4	0.47	0.53	0.49	40
Accuracy			0.62	194
Macro-average	0.62	0.57	0.57	194
Micro-average	0.63	0.62	0.61	194

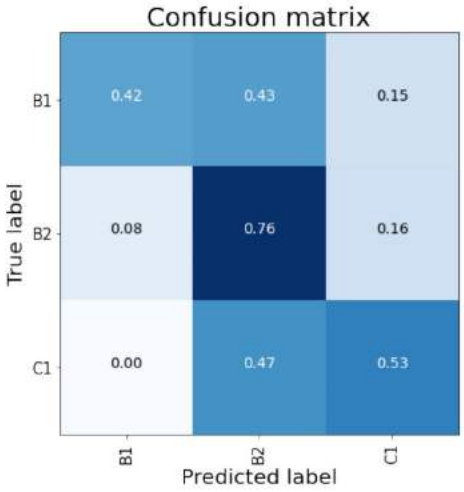


Figure 37: Classification report and confusion matrix on CLC- FCE test results on CLC- FCE model (automatic corrections)

The classification report in Figure 37 displays a balanced distribution of accuracy in the F1-score values, while the overall accuracy is equal to **62%**. Considering that these are exams that we have arbitrary remapped to CEFR levels, the results appear actually consistent. Figure 37 above additionally depicts the classification results embedded in the confusion matrix. The latter by means of the darker blue shades in the central diagonal indicates that the B2 level of competence received the best predictions, although the values of the adjacent classes, B1 and C1, seem not highly dissimilar from the former.

Considering the results of the previous experiments, this model proves to be the best of all. Most probably also owing to the automatic corrections of LanguageTool, the model manages to improve the predictions for the different proficiency classes.

6.2.4. Testing the EFCAMDAT model on the CLC- FCE test set

Since we had already a trained model for English that proved to be significantly accurate, namely EFCAMDAT, we decided to carry out a last experiment using the former and evaluating its performance on the CLC- FCE test set. Hence, we made use of the previous 194 randomly extracted learners' texts, to test the EFCAMDAT model trained on different data but in the same source language.

Before training the model, we remapped the assigned exams scores from 1.2 to 5.3 to the CEFR levels of competence. We followed the principles described in Table 8 (§ 4.3.2). Obviously, this procedure represents a constrained arbitrary mapping of the students' results for a mainly B2 level exam by distributing them along four total CEFR levels.

The reason for this experiment is to determine whether we could actually use the model already trained for the same language with a different dataset and obtain discrete classification results, despite the labels not matching those of the dataset used for the training.

	precision	recall	F1-score	Support
0	0.00	0.00	0.00	2
1	0.04	0.33	0.07	3
2	0.29	0.46	0.35	48
3	0.74	0.22	0.34	141
4	0.0	0.00	0.00	0
Accuracy			0.28	194
Macro-average	0.21	0.20	0.15	194
Micro-average	0.61	0.28	0.33	194

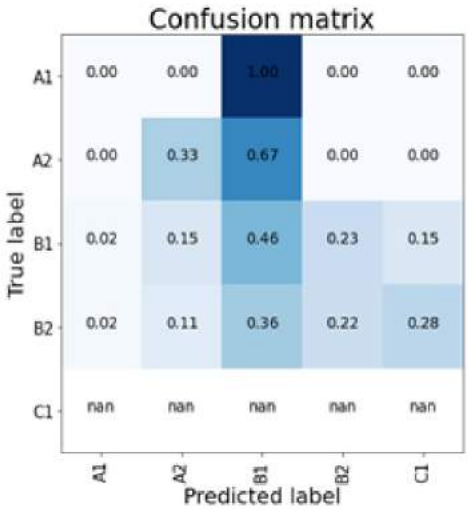


Figure 38: Classification report and confusion matrix on CLC- FCE test results on EFCAMDAT model

The classification report (see Figure 38) shows the distribution of classes according to this new mapping system. We can point out that since the main purpose of the examinations contained in this corpus was to measure B2 level skills, class three exams appear to be the most numerous while class four, i.e. C1, present in the original model's training set is no longer represented in this test set. For this reason, not a number (nan) values appear in the confusion matrix. The F1-scores for class B1 and class B2 are, however, quite close, although the former is in the clear minority compared to the latter. More information about the results of the models is provided later in the final section (§ 6.5.).

6.3. MERLIN Italian

In this subsection, we describe in detail the experiments we performed with the Italian language corpus. It consists of a total of 813 texts distributed in 29 texts for the A1-level, 381 texts for the A2-level and 394 texts for the B1-level (excluding unrated texts). On the basis of this data, we proceeded to execute various experiments, first by simply using the original incorrect learners' examinations, then by employing these texts with the corrections manually applied, and finally by means of the corrected versions automatically generated by LanguageTool.

6.3.1. MERLIN Italian model trained with original exams

As an initial experiment with the Italian language corpus, we decided to attempt to build a model to classify students' proficiency levels based solely on their original written examinations, without their corrections. For this purpose we trained and evaluated a model applying the *k-fold cross-validation* procedure. This technique is usually employed to avoid overfitting the model both when the amount of training data is restricted and when the parameters to be considered by the model are manifold. In our case, we also applied cross-validation to better monitor the learning-validation process of the model. Therefore, we created three random partitions of circa 260 files each (§ 4.3.3.1.) and alternatively used them to train and test the model. We set a learning rate equal to 2e-5, 50 epochs and a validation set corresponding to 20% of the training set.

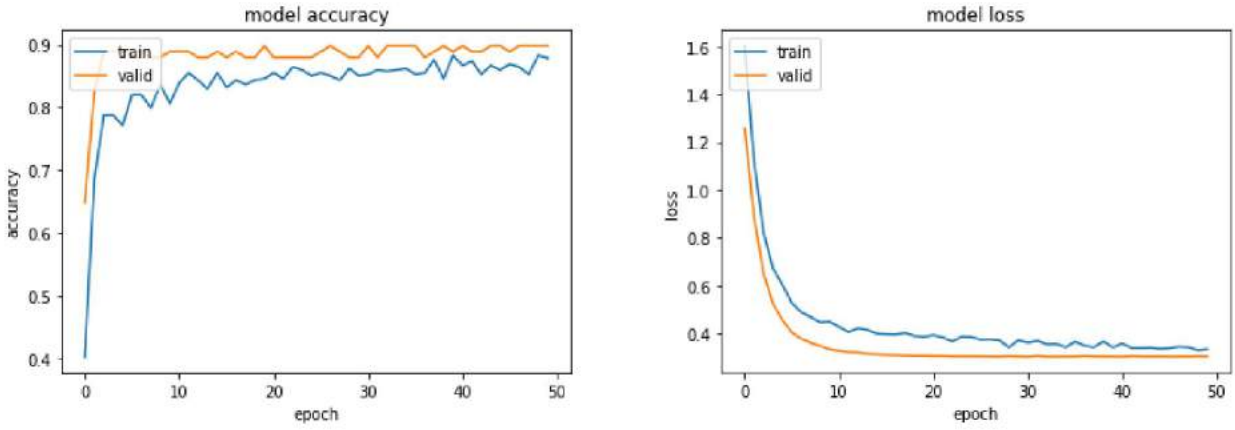


Figure 39: MERLIN Italian example of training & validation loss and accuracy curves (only original tests)

Above (Figure 39) a curve from the 3-fold training process displays an accuracy on the training set over 85% and over 90% on the validation set.

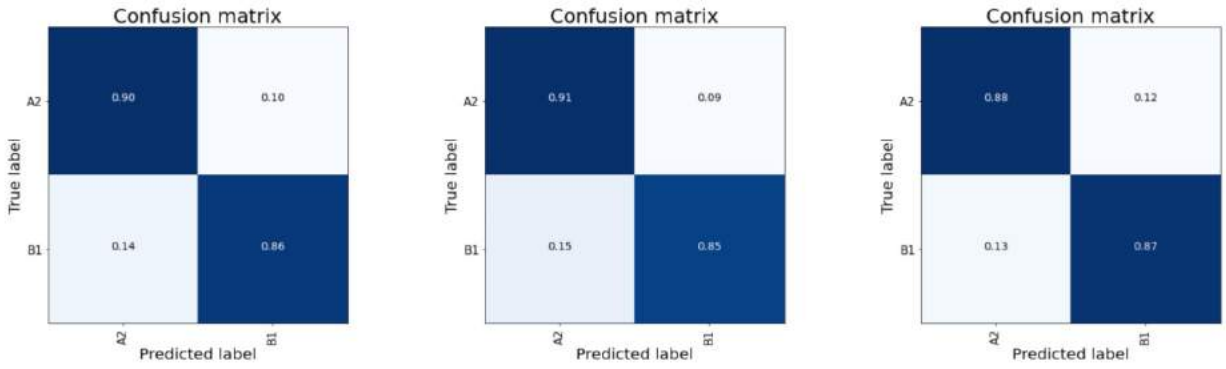


Figure 40: MERLIN Italian confusion matrices (cross-validation only original tests)

The resulting accuracy on all of the three test sets corresponds to a mean value of **87,6%**. From the three confusion matrices above (see Figure 40) we can notice, first of all, that given the unbalanced number of exams for each level of the dataset (see Table 4 for more details), we decided to consider only the levels with the highest number of exams, namely A2 and B1, for the evaluation of the model. Furthermore, we can observe that the results obtained for both of them are close and high, between 85% and 91% of accuracy, with a slight upper improvement for A2 examinations. Overall, despite the fact that this was carried on a limited amount of data, these initial results appear valid and accurate.

6.3.2. MERLIN Italian model trained with human examiners' corrections

In the case of the Italian tests, we were also provided with the correct version of each text produced by one annotator. We, therefore, tried to use the human corrections as inputs for the model in the training together with the original students' examinations. The parameters we used for this experiment are the same as the ones of the previous case.

The curves in Figure 41 display the accuracy and loss values of one of the three folds during the training process. The first is equal to approximately 88% on the training and validation set, while the second gradually decreases up to less than 0.4 in both sets.

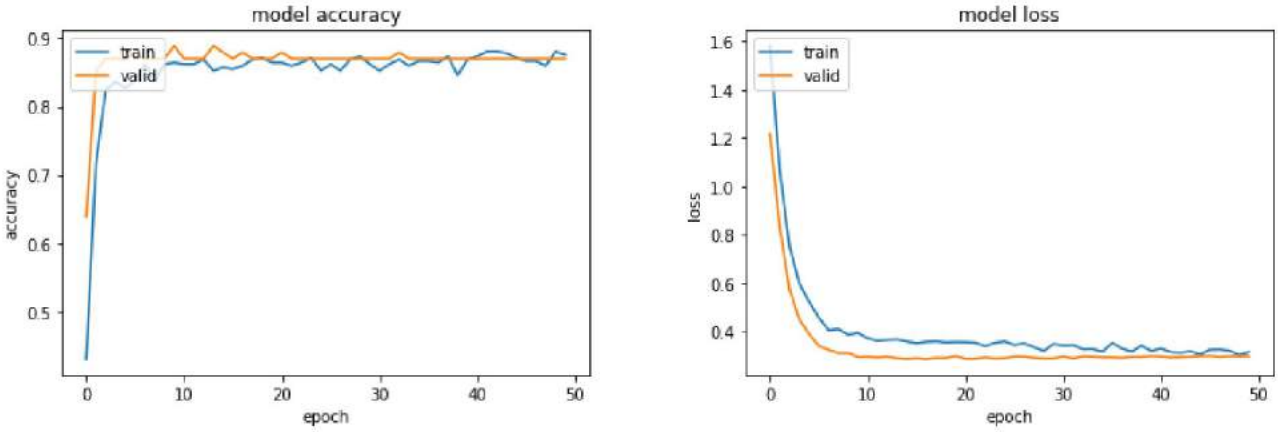


Figure 41: MERLIN Italian model training and validation accuracy and loss curves (human corrections)

The results obtained from the evaluation (see Figure 42) indicate a mean accuracy of **87,6%**. This outcome appears in line with the one previously obtained using only the original texts.

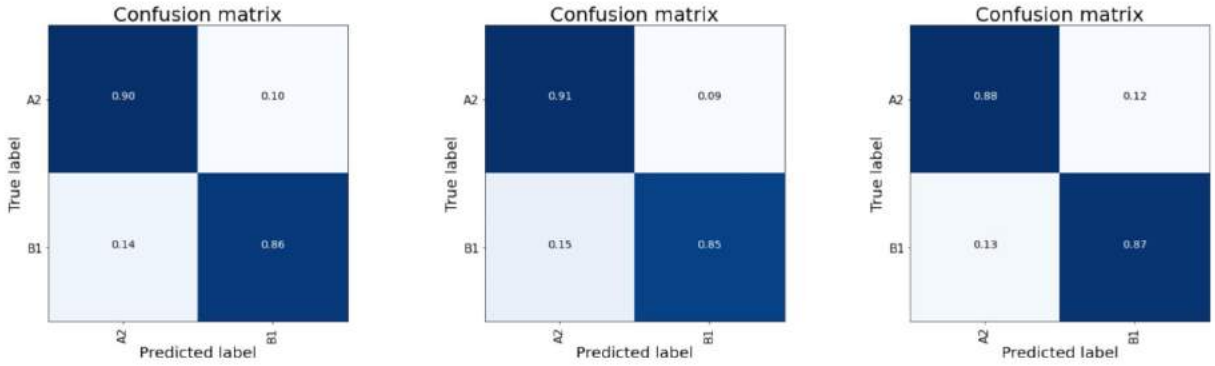


Figure 42: MERLIN Italian confusion matrices (cross-validation human corrections)

Thus, apparently whether we just pass on the uncorrected student texts or add human corrections to them, hardly anything changes to our architecture in terms of classification accuracy. In the next sections, we consider whether this remains true in the case of the automatically generated corrections by LanguageTool.

6.3.3. MERLIN Italian model trained with LanguageTool automatic corrections

Unlike the previous attempts made with cross-validation, as regards these latest experiments made with the automatic LanguageTool corrections of the students' texts, we decided to first make a trial with two partitions, without cross-validation, and then to compare the results, once again with the latter strategy.

In the first experiment, we mainly used two partitions, 609 files for the training set and 102 files for the test set. We excluded the files with a length inferior to 20 words and trained the model for 200 epochs with a learning rate equal to 2e-5. This time, our training and test set also contained an empty examination, to which no level was assigned, and some A1 level examinations. Before the training process, we rebalanced the weights of each class to avoid deteriorations in the performance using the Sklearn's *class_weight* parameter. After this initial phase, we obtained an accuracy of circa 80 % on the training set.

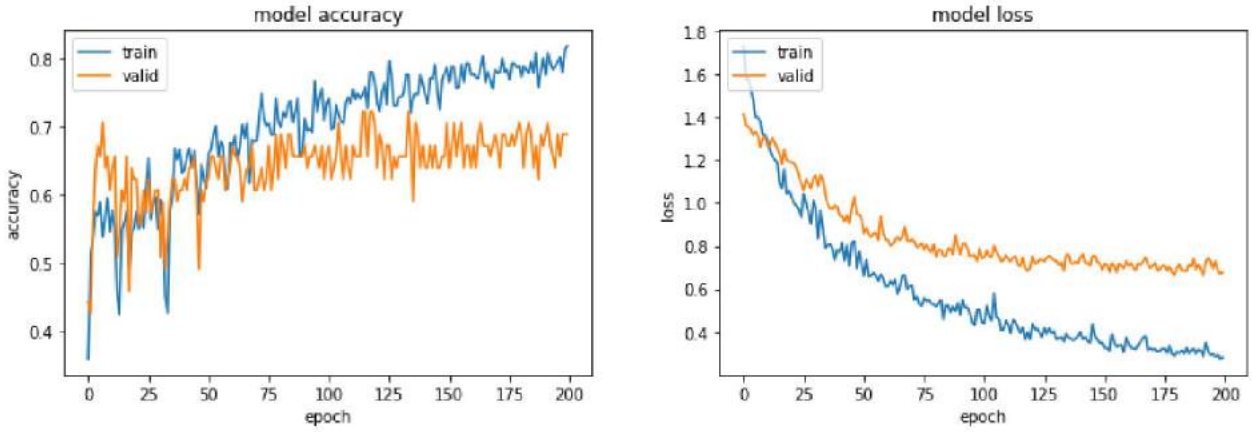


Figure 43: MERLIN Italian model training and validation accuracy and loss curves (balanced classes)

From the curves in Figure 43 we can also observe that the trend in the loss reflects our expectations, meaning that it gradually decreases, even in the case of the limited validation set. Equally, the accuracy progressively improves, reaching significant peaks on the validation set, too.

	precision	recall	F1-score	support
EMPTY	0.00	0.00	0.00	1
0	0.33	0.25	0.29	4
1	0.85	0.69	0.76	48
2	0.76	0.86	0.81	49
3	0.00	0.00	0.00	0
Accuracy			0.75	102
Macro-average	0.39	0.36	0.37	102
Micro-average	0.78	0.75	0.76	102

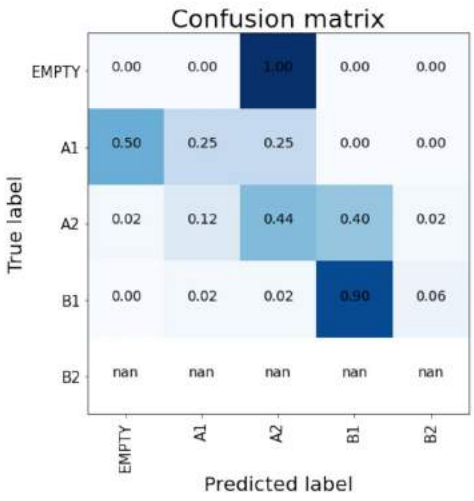


Figure 44: Classification report and confusion matrix MERLIN Italian model (balanced classes)

The results attested by both the confusion matrix and the classification report (Figure 44) reveal that only the class "empty" with a single sample and the class B2 with no support in the test set were not predicted, while all the others had predictions that largely varied according to their number in the training set, despite having rebalanced the weights before. The overall accuracy of **75%** appears to be less than the one obtained from the previous experiments. However, due to the unbalanced and reduced nature of the dataset, it cannot be considered a fully valid measure for the evaluation of the model. In an attempt to improve the performance of the model on this data, we conducted the additional experiment illustrated in the following section.

6.3.4. MERLIN Italian model trained with LanguageTool automatic corrections using cross-validation

Considering the problems encountered previously due to the restricted data available and their imbalanced nature, we opted for the *k-fold cross-validation* procedure used also for the previous two initial models. Therefore, we employed the same three partitions used for the experiments in sections 6.3.1. and 6.3.2. to train a model with the original learners' texts and their automatic corrections. We adopted the same training parameters of the previous models and obtained the results displayed in the following curves on the training and validation sets.

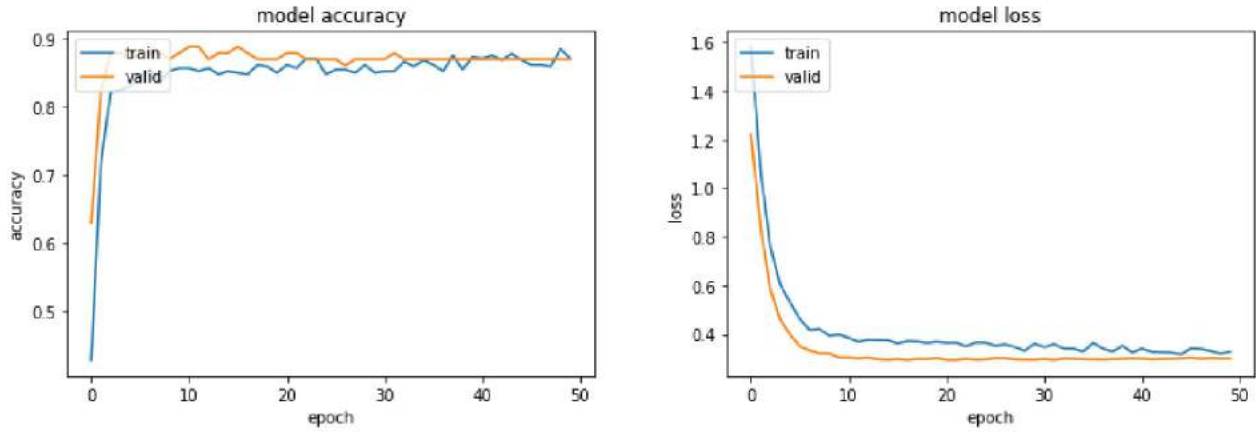


Figure 45: MERLIN Italian example of training & validation loss and accuracy curves (automatic corrections with cross-validation)

After we alternately used two of the three folds to train the model, we evaluated its performance on the one left out. At the end we considered the average accuracy values of the test set partitions. The confusion matrix of each fold is represented below.

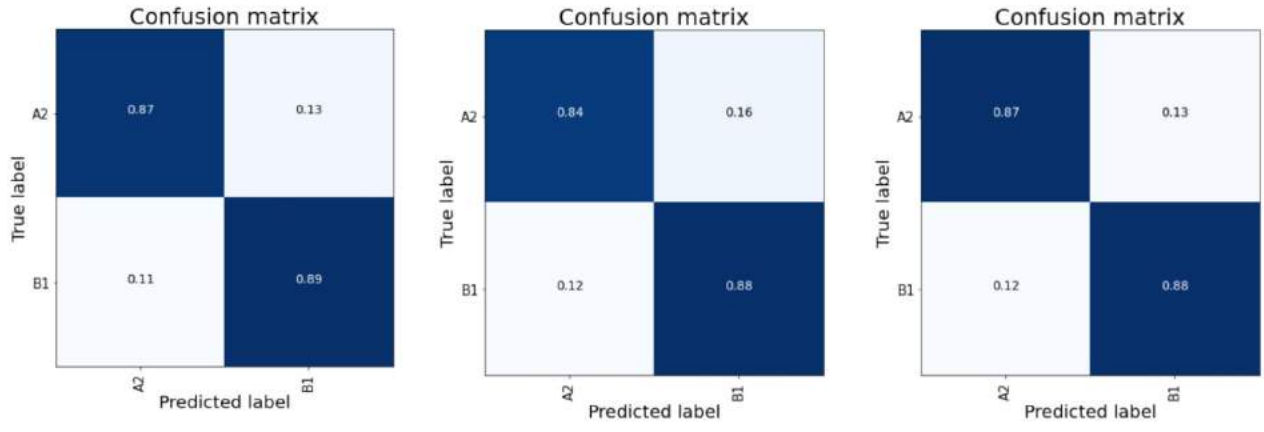


Figure 46: MERLIN Italian confusion matrices and example of classification report (automatic corrections with cross-validation)

As can be observed, we once again only selected the most numerous represented classes in order to avoid false representations of the model performances given the limited dataset. Therefore, Figure 46 displays the accuracy

scores obtained for the classification of A2 and B1 exams, which are between 84% and 89%, balanced in the two cases. The mean accuracy corresponds precisely to **87%** overall the three folds. Compared to the previous model with automatic corrections, this result appears better, although the same cannot be said in comparison with the other two initial models.

6.3.5. MERLIN Italian outcomes

Given the limited material available for the Italian language, but also the availability of human corrections, we conducted the diverse experiments illustrated in detail in the previous sections. The outcomes were partly in line with our expectations and partly conditioned by the limited resources and the imbalance of the number of exams per level.

The following table summarises the results of the various experiments carried out with the MERLIN Italian dataset.

MERLIN ITALIAN EXPERIMENTS			
Cross-validation with original texts only	**Cross-validation with original texts and LT corrections**	**Dual training and test set (balanced weights)**	**Cross-validation with original texts and LT corrections**
87,6%	87,6%	75%	87%

Table 9: MERLIN Italian experiments results (accuracy)

The differences among the various models and experiments do not appear to be striking, indeed, they are very similar in terms of training and test results, especially observing the ones for which cross-validation was used. One of the numerous reasons for this may be related to the limited amount of data used. Probably with a much larger and more diverse dataset the outcomes may have displayed greater diversity.

6.4. MERLIN German

Similarly to what we did for the Italian examinations from MERLIN, we used the 1,033 exams available for the German language, distributed in a total of six levels of competence, to train a model and evaluate it. However, in order to obtain the best model possible out of such limited amount of data, we experimented with different data splits and parameters to be passed to the original BERT-based architectures (see Figure 10). Hence, as before, we have trained a system taking the original students' texts as inputs first, and another taking the previous added to their automatically obtained corrected versions.

6.4.1. MERLIN German model trained with original exams

As a first experiment related to the German language, we decided to attempt to train a model using only the learners' original texts, without providing any additional inputs regarding errors and corrections. Starting from the

57 exams available for the A1 level of competence, the 306 available for the A2, the 331 available for the B1, the 294 available for the B2, the 42 and 4 available for the C1 and the C2 levels, we created three partitions to perform a similar operation to what we did with the Italian dataset. We decided to remove the levels for which we had less than 100 exams to mitigate the effect of the data imbalance on the model's performance. For the training ad test we resorted to cross-validation, using three folds of 340 exams circa. . Thereafter, we selected a learning rate equal to 2e-5, a maximum text length corresponding to 380 words, a batch size of 16 texts and 120 as number of epochs. The training curves for accuracy and loss on the last cross-validation partition can be seen below.

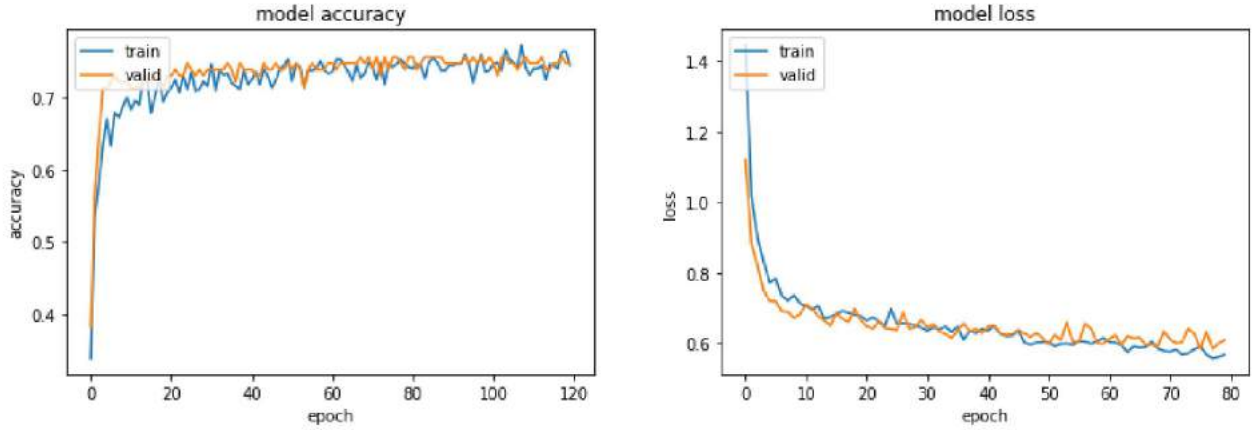


Figure 47: Accuracy and loss training curves for MERLIN German model (original texts only)

The loss and accuracy curves obtained during the training process with each of the three different combined partitions show an average accuracy of around 70% on the training and validation sets (see Figure 47).

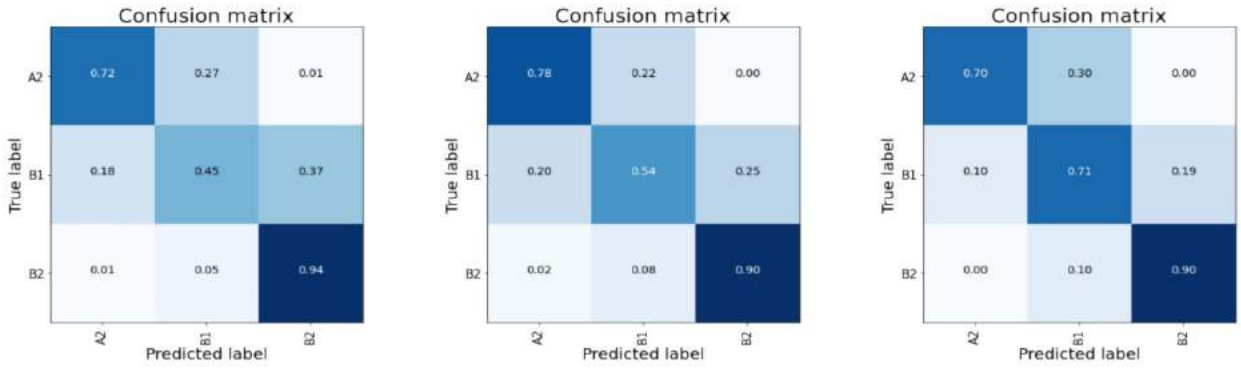


Figure 48: MERLIN German confusion matrices (original texts only)

Even considering the confusion matrices above we notice that the average accuracy results for the CEFR levels A2, B1 and B2 are close to 70%, more precisely their average value corresponds to **72,6%**. Carefully observing the accuracy scores in the matrices, we note that the best predicted level is B2, while the worst results are found for the B1 class. However, since this represents an initial experiment with only the original texts and limited examinations per level, the results appear to be reasonable.

6.4.2. MERLIN German model trained with LanguageTool automatic corrections using cross-validation

For the second category of experiments relating to MERLIN German, we decided to add as inputs for the model the automatically generated corrections of the learners' texts provided by LanguageTool. In order to obtain comparable results with the previous experiment, we decided to adopt the same 3-fold partitions and the same training parameters. Therefore, we trained the model for 120 epochs and obtained the results shown in the following figures.

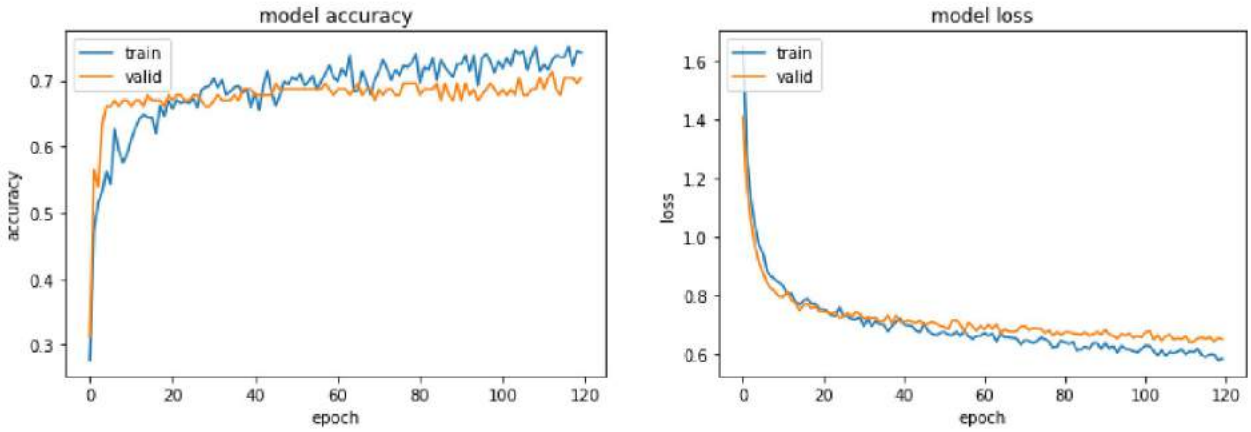


Figure 49: MERLIN German example of training & validation loss and accuracy curves (automatic corrections with cross-validation)

The loss and accuracy curves obtained during the training process with each of the three different combined partitions show an average accuracy of 77% on the training set and 70% on the validation set (see Figure 49). Although the latter accounted for only 20% of the entire training set, it proved to be a good representation of the examination sample of the different competence levels.

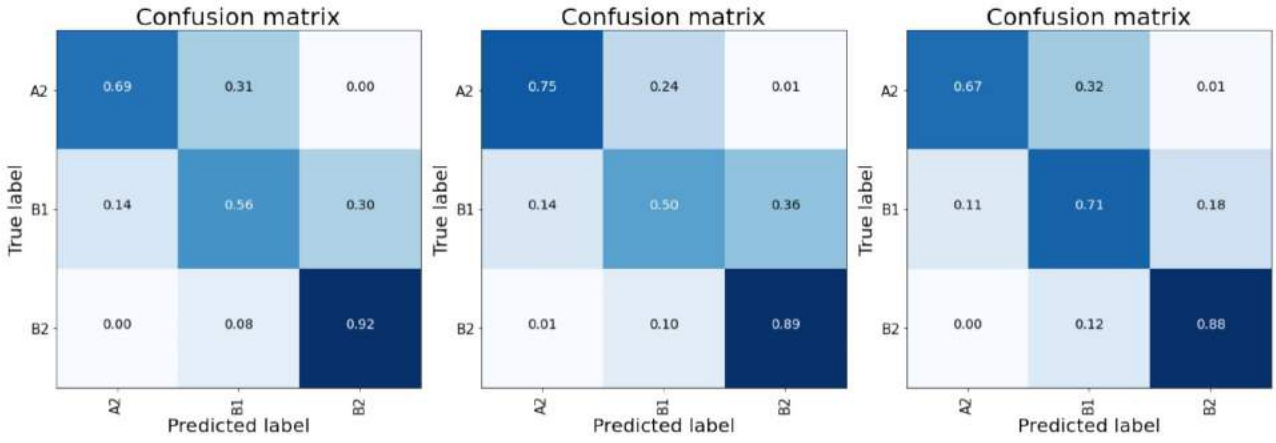


Figure 50: MERLIN German confusion matrices and example of classification report (automatic corrections with cross-validation)

Examining the confusion matrices resulting from the performance evaluation of the model on a test set of about 240 examinations, we can notice that the best predicted class is also in this case the B2 level of competence. However,

the accuracy values for the B1 and A2 categories also appear in a noticeable accuracy range of 50-75%, especially the latter. The mean accuracy result amounts to approximately **72,3%**. The difference to the previous model trained without corrections appears to be extremely narrow. In order to investigate a possible wider deviation, despite the reduced dataset, we decided to perform one last experiment, which is illustrated in the following section.

6.4.3. MERLIN German model trained with LanguageTool automatic corrections (dual partitions)

Starting from the original 1,033 exams available in the MERLIN German dataset, we created two partitions, one for the training set and one for the test set. The first one included 775 written exams, of which, however, we removed the texts with less than 20 words since they could have affected the model's performance given their particularly short length. Subsequently, we selected a learning rate equal to 2e-5 and a maximum text length corresponding to 380 words, as in the case of the previous experiments with cross-validation. Additionally, before the beginning of the training process, for which we established a duration of 60 epochs, we joined the few exams available for class 4 and 5, C1 and C2, under the C1 label. The reason for this choice was to confirm or overturn the choice of removing the latter classes in previous experiments, as well as to test the overall performance of the model on all the levels of the original dataset.

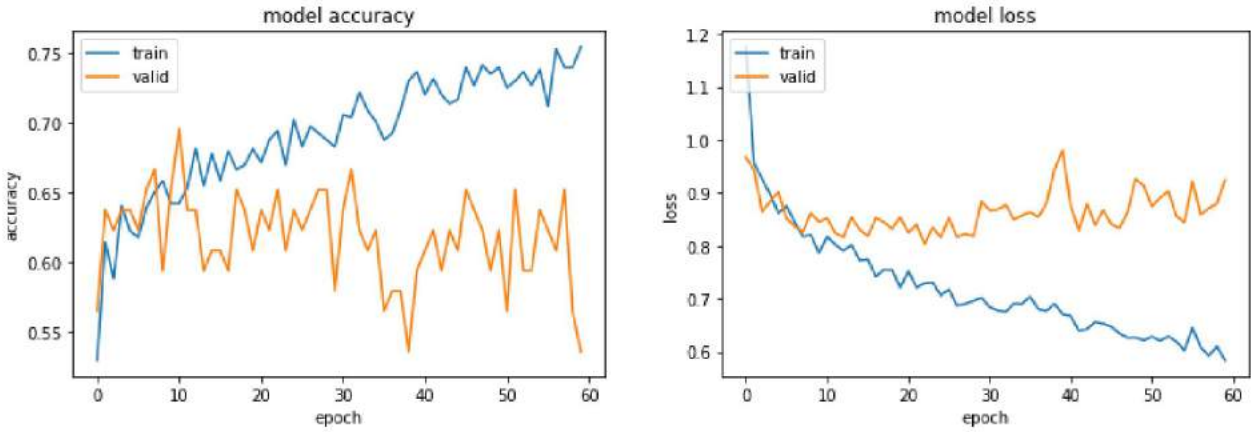


Figure 51: Accuracy and loss training curves for MERLIN German (dual partitions)

Figure 51 above displays the model accuracy and loss curves related to the training. The accuracy on the training set in blue can be seen to reach a maximum value of 75%, while the one on the validation set, represented by 20% of the total training set, reaches a maximum of 70%.

	precision	recall	F1-score	support
0	1.00	0.62	0.77	8
1	0.70	0.76	0.73	41
2	0.68	0.60	0.64	43
3	0.74	0.74	0.74	34
4	0.25	0.67	0.36	3
Accuracy			0.69	129
Macro-average	0.67	0.68	0.65	129
Micro-average	0.71	0.69	0.70	129

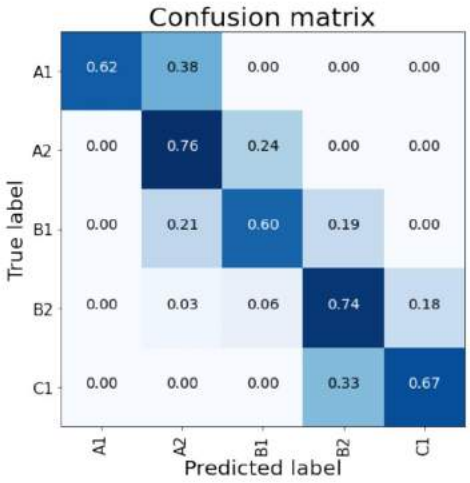


Figure 52: classification report and confusion matrix of MERLIN German base-model

As can be observed from the values contained in Figure 52 in relation to the results of the model on the test set, the overall accuracy is **69%** and the distribution of the F1-scores for each class appears to be balanced, except for the 4th class or C1 level, where the obtained values are the lowest (36%). The best predictions, instead, were found for the level A2, with an assigned 0.76 F1-score.

Despite the fact that the weights of each class were not rebalanced and some of them presented lower numbers of support, the results on each class appear to be distributed around 62% and 76%, fairly remarkable given that these are as many as five classes or levels of competence between which the model had to make predictions.

6.4.4. MERLIN German outcomes

Although the German examination data were taken from the same dataset as the Italian ones, their actual organisation and available documentation were different. First of all, they were larger in number and belonged to more classes. This led to the use of different strategies during the experiments. For this reason, the tests we carried were not exactly parallel to the others. For example, in this case no manually corrected versions were provided, so we could not carry out that type of experiment (§ 6.3.2.). In addition, the results on the German language appeared more balanced per class than in the case of the Italian language, so there was no need for a forced rebalancing of the weights per category.

The results of the final three different types of experiments performed first without and then with cross-validation are reported in the summary table below.

MERLIN GERMAN EXPERIMENTS		
Cross-validation with original texts only	**Cross-validation with original texts and LT corrections**	**Dual training and test set**
72,6%	72,3%	69%

Table 10: MERLIN German experiments results (accuracy)

As can be observed from the table above, the best results on this German dataset were obtained using the cross-validation strategy, although there are actually minor differences between the case where automatic corrections were used and the case without. On the other hand, the last attempt with only two partitions for training and test proved itself valid for as many as five proficiency levels, while less accurate than the previous two on less classes classification.

6.5. Final written experiments' summary

From the experiments carried out for each language described in the previous sections, we now present our drown conclusions. We firstly discuss the two models for the English language, trained with the EFCAMDAT and CLC-FCE datasets correspondingly, then we report on the results on the Italian MERLIN dataset and finally those on the respective German MERLIN dataset. The overall achieved results can be seen in the table below (Table 11).

Database	Accuracy with original texts only	Accuracy with original texts & human corrections	Accuracy with original texts & LanguageTool corrections
EFCAMDAT	96,8%	96,6%	97,2%
CLC- FCE	60%	61%	62%
MERLIN-ITALIAN	87,6%	87,6%	87%
MERLIN-GERMAN	72,6%	-	72,3%

Table 11: Summary of English, Italian and German models experiments

Regarding the models trained with the EFCAMDAT dataset, containing the largest number of exams available for all six CEFR levels, we found that the models trained with the inputs of the original texts and LanguageTool automatic corrections are remarkably accurate in our multi-class classification task. In fact, they present the highest accuracy found throughout the experiments, namely approximately 96,8% and 97,2%. Differently, the model trained using the human corrections performed slightly worse, achieving an accuracy of 96,6%. In contrast, the model trained from the CLC- FCE corpus, presumably given both the limited data, the imbalanced distribution of the different classes and the forced remapping of the scores assigned to the CEFR levels, demonstrated a much lower accuracy of 62%. Much similar degree of accuracy was observed with the experiments using human corrections, i.e. 61%, and the original texts in the classification, i.e. 60%. However, among the three attempts, the most successful was the first one with the automatic LanguageTool corrections.

As far as the Italian language is concerned, although the available dataset contained the lowest number of examinations among all the corpora considered in this project, the results obtained are the second most accurate overall. In fact, when we tested the language-based model on the original texts accompanied by the automatic corrections, the accuracy value found was 87%. On the other hand, in the experiments carried out with the human

corrections and the original texts only, we obtained the extremely close results, i.e. 87.6% of overall accuracy. Nevertheless, given the limited number of examinations considered, we cannot claim that there are significant differences between the various experiments, so a larger and more balanced dataset would be necessary.

In the case of German, with the baseline model trained as for the other languages using the students' original texts and LanguageTool corrections we found an accuracy of 72% using cross-validation. The reason for this is that after having attempted to employ a dual data partition for both Italian and German, we noticed that the reduced availability of tests in the datasets and the rather unbalanced distribution in the different proficiency classes influenced the model outcomes (see Table 9 and 10). This is another reason why we resorted to rebalancing the class weights for the different experiments mentioned above. In addition, we also experimented with creating a model that used only the learners' original examinations in the training and classified them according to different levels of competence. In this case, we did not notice much difference in the results, obtaining the same accuracy as before, i.e. 72%. Once again, this could most likely be due to the low data availability and, therefore, further datasets and experiments could contribute to notice stronger differences.

Overall, we can finally state that the models trained with the original texts and LanguageTool all perform well, although with larger or smaller differences for the distinct datasets of the three languages, depending on the number of available data and class partitions. Indeed, the best model obtained is the one trained with the largest amount of data, i.e. EFCAMDAT. Moreover, the employment of human corrections as opposed to automatic ones cannot be considered more advantageous so far given the hardly differentiated results found. Human detected errors are more numerous, but less detailed, not classified into meaningful categories and inter-subjectively variable, at times inconsistent. For this reason, the use of an automated language checker, such as LanguageTool, for text correction could provide an improvement on a neural model in determining language learners' proficiency rather than using human expert corrections. Nevertheless, our employed BERT-architecture has been evidently capable of performing accurate multi-class classification both when receiving the original learners' texts as sole inputs, and when being provided with additional automatic corrections.

CHAPTER SEVEN: THE CASE STUDY OF ENGLISH ORAL EXAMS OF THE FREE UNIVERISITY OF BOZEN-BOLZANO

This section is dedicated to a case study of B2 English language exams extracted from the system of the Language Centre of the Free University of Bozen-Bolzano. Since this is a relatively limited dataset, we could not follow the exact same procedures adopted with the written texts for English, Italian and German. However, in order to maintain consistency with the adopted techniques we decided to work with the audio transcriptions, partly also for privacy reasons. As a first step, we carried out data processing, which means that we transcribed the data with an ASR system, corrected these automatic transcriptions with the help of an annotator and organised each part 1 and part 2 of the oral examinations in a *tsv* file also containing their assigned scores. In addition, we proceeded to analyse factors such as WER and MER to assess both the transcription system and the learners' speech itself with numerical measures.

After this initial part, we used LanguageTool for the automatic detection of errors not related to punctuation but to the use of the language in terms of grammar, style and register. Then, we began the experiments. In order to do so, we used two architectures similar to those initially applied to the English language, trained with the EFCAMDAT and CLC-FCE datasets. Yet, given the different structuring of the data, monologues and not emails or essays, as well as the different assignment of scores and not more CEFR levels, we modified the models by making one that would predict whether the learners had passed or failed the test and one that would assign scores similarly to how the examiners had done. Finally, given the overall results of the experiments, we considered some linguistic features that could serve as additions to automatic assessment systems to evaluate learners' spoken language levels.

7.1. Data description

In addition to the open-source data available for written examinations in English, Italian and German for different levels, this project considered also data related to oral English tests. These concern oral examinations conducted by the Language Centre of the trilingual Free University of Bozen-Bolzano. The latter represents a unique university in the Italian system, standing out likewise at European level for the possibility of simultaneously studying in Italian, German and English throughout distinct academic domains. From the first year of their studies, students have to learn a range of subjects in the three languages, and it is therefore necessary for them to demonstrate sufficient levels of competence in the respective. To improve and test their language skills, the university offers special language courses and exams every semester administered by the Language Centre. From the latter, we were provided with the 60 oral examinations used for this case study aimed at the attempt of the automatic assessment of oral English proficiency at B2 level.

More in particular, our dataset consists of 120 audio clips produced by 60 students who took the exam for the CEFR B2 level of competence in February 2021. Each of them had to deal with the following three modules:

1. module: listening, reading comprehension and language structures (time available: 80 minutes)
2. module: written production of 280 words (time available 50 minutes)
3. module: oral production split in two monologues of 2 minutes and 30 seconds of duration.

The tests are entirely computer-based, and each module must be passed to gain access to subsequent modules. Oral productions are recorded on the students' personal computer and automatically stop once the available time expires. On the screen the student can see the assignment and a timer. There are several assignments available. In our case, they were mainly related to short narrations, discussions and descriptions about general life-style opinions, holidays and topical issues.

Each test was scored by experts from the Language Centre of the University of Bozen-Bolzano. A score of 0.6 or above was required to pass the exam. Lower scores signify failure to pass it. The majority of the examinations contained in our dataset, namely those of 45 students, received a sufficient score to assess the B2 level of competence. The remaining 15, instead, got assigned scores under 0.6, meaning that the demonstrated proficiency was insufficient.

In addition to this score between 0 and 1, evaluators also rated communicative effectiveness, lexis, grammar and fluency and pronunciation. These parameters could receive a score from 0 to 5 (see Appendix B Figure 3).

7.2. Data processing

In order to work with the oral exam data, we transcribed the content of all 120 productions using an Automatic Speech Recognition System available within the Kaldi toolkit for speech analysis and processing (Povey et al. 2011). The latter is a flexible set of systems written in C++ based on finite-state transducers. The data used for the training come from the LibriSpeech ASR corpus consisting of approximately 1000 hours of 16kHz English Speech based on material for reading audio books (Panayotov et al., 2015). Since we were dealing with non-native speakers with different L1s, and recordings made on each student's device in different sound environments, which may not have been entirely free of background noise or interference, at times the automatically recognized words did not accurately correspond to the learners' utterances. In an effort to get the most accurate transcriptions possible, an annotator manually corrected the automatically generated transcriptions. The obtained texts were revised keeping the references to hesitations or disfluencies, mainly transliterations of the vocalizations [ə(:)] and [ə(:)m], corrected from "and" ,"am", or at times ignored by the automatic system, to "uhm" and "uh". All the errors made by the students were preserved and just incorrect words recognised by the Kaldi system have been edited.

After this process, the data were organized in a *tsv* document containing the anonymised code corresponding to each student, gender, transcribed oral production number one and oral production number two, assigned pass/fail scores and points for sub-competences. Alongside these, since none contained punctuation and we also wanted to follow a similar procedure to the one previously applied to written texts, the texts' transcriptions were automatically corrected using LanguageTool.

7.2.1. Evaluation of automatic transcriptions and students' speech

Before the experiments for the automatic assessment of the examinations began, we analysed the effectiveness of the ASR system and to some extent also the comprehensibility of the students' speech. For this purpose, we adopted the Word Error Rate (WER) and Match Error Rate as evaluation metrics (Morris et al. 2004). We extracted them using Python's *jiwer* tool which computes the minimum-edit distance between the ground-truth transcription and the hypothetically correct transcription produced by the annotator or automatically by LanguageTool. Considering

the minimum-edit distance allows us to quantify the amount of dissimilarity between the strings or words of the two texts, evaluating in this manner both the quality of the automatic speech recognition system (MER) against the annotator's transcription, and in part also the fluency of the English non-native students (WER) against the LanguageTool corrected exam versions.

Figure 53 displays graphical representations of the match error rate values obtained by taking into account the proportion of I/O word matches and their probability of incorrectness based on the following formula:

$$MER = (S+D+I)/(N=H+S+D+I) = 1-H/N \quad [5]$$

S= Number of substitutions (erroneous words)

D= Number of deletions (word omissions)

I= Number of inserted words

H= Number of correct correspondences

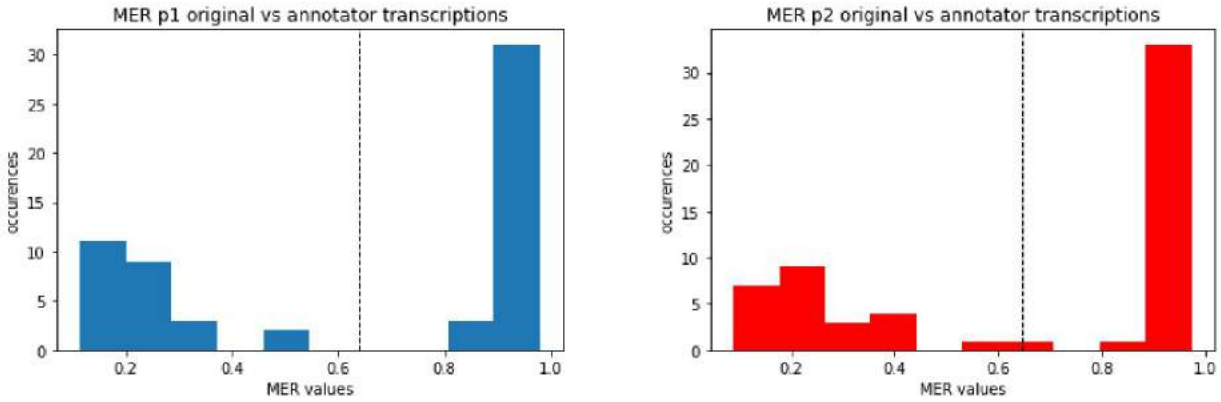


Figure 53: MER between ASR output and corrected transcriptions for Part 1 & Part 2 of oral exams

The two histograms display that the average match error rate between the automatic transcripts and the annotator-corrected transcripts lies approximately over 65% for both part 1 and part 2 of the oral examination, as signalled by the dotted line indicating the mean values. The benchmark WER for ASR systems on native English speakers lies under 20% (Szymanski et al. 2020), while for non-native English speakers it ranges from 15% up to around 35% (Wang et al. 2020). In our case we obtain worse results because the automatic recognition system has not been adapted or modified to the learners' speech and it was originally designed for read speech from native speakers.

To cross-check the obtained error rate, we also resorted to the standard word error rate, which measures the proportion of incorrect words to the total number of words processed according to the following formula:

$$WER = S/(N=H+S) = 1-H/N \quad [6]$$

S= Number of substitutions (erroneous words)

D= Number of deletions (word omissions)

---

[5] Adapted from Morris et al. (2004), pp. 4.

[6] Adapted from Morris et al. (2004), pp.1.

I= Number of inserted words

H= Number of correct correspondences

Although this is a widely used evaluation metric for ASR, WER suffers from some inconveniences, e.g. if S=D=0 then the counted insertions for each word are duplicated, resulting in a WER of 200%. Furthermore, in presence of noise WER values can reach values higher than 100% giving more importance to I, insertions, than to D, deletions (Errattahi et al. 2018). This last point can be noticed in Figure 54 representing the WER for the texts of part 1 and 2 of the oral examination obtained by comparing the original text produced by the ASR system and its version automatically corrected by LanguageTool. In fact, the WER values represented on the x-axis exceed 1, indicating more than 130%.

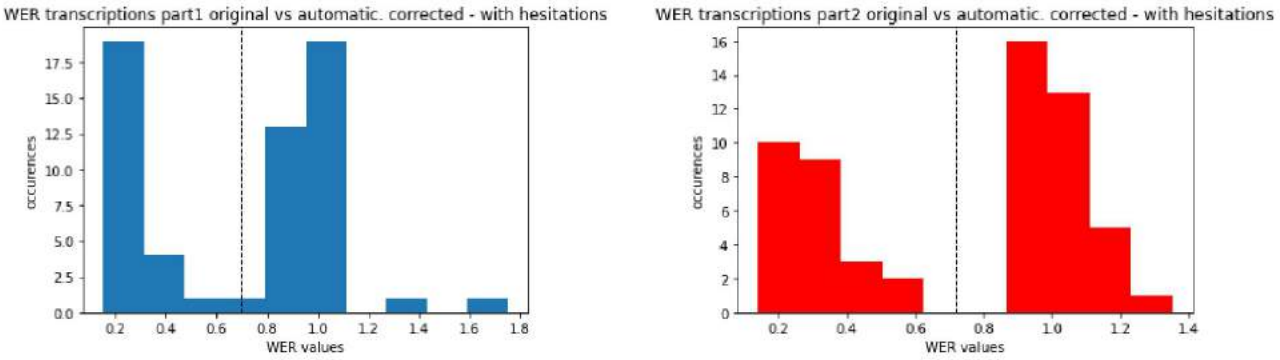


Figure 54: WER between ASR output and LT corrected transcriptions for Part 1 & Part 2 of oral exams

To check these results against the human corrected versions we did the same thing this time by passing the correct transcriptions by the annotator and the versions automatically generated by the ASR system. In Figure 55 below, we can actually notice that the average WER found in this case is, indeed, close to what we observed before. Thus, considering the automatic transcription and the manually corrected one, there are actually significant discrepancies between the two.

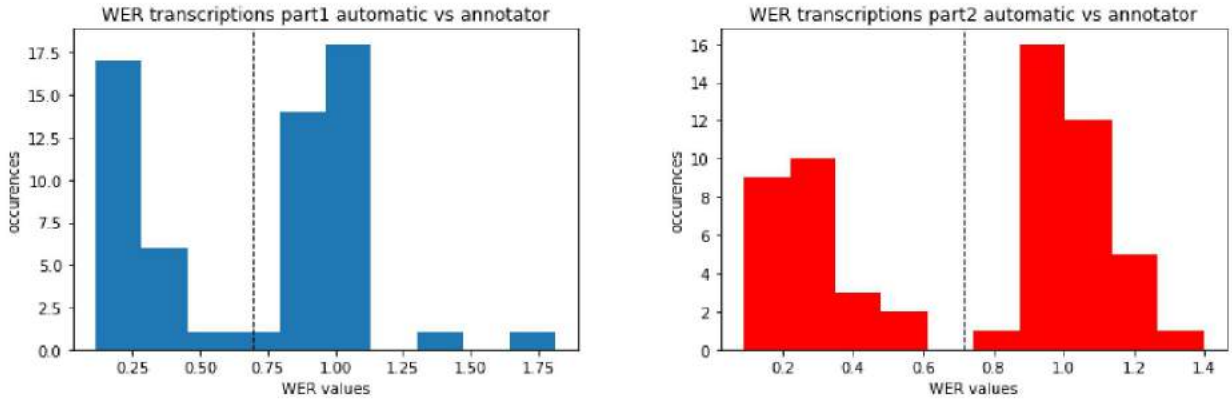


Figure 55: WER between ASR output and manually corrected transcriptions for Part 1 & Part 2 of oral exams

With regard to the students' appropriately fluent and native-close pronunciation, we created a scatter plot by correlating the WER values found in the two cases described above and the fluency scores assigned by the expert evaluators of the University Language Centre. In Figure 56 we can compare the obtained results.

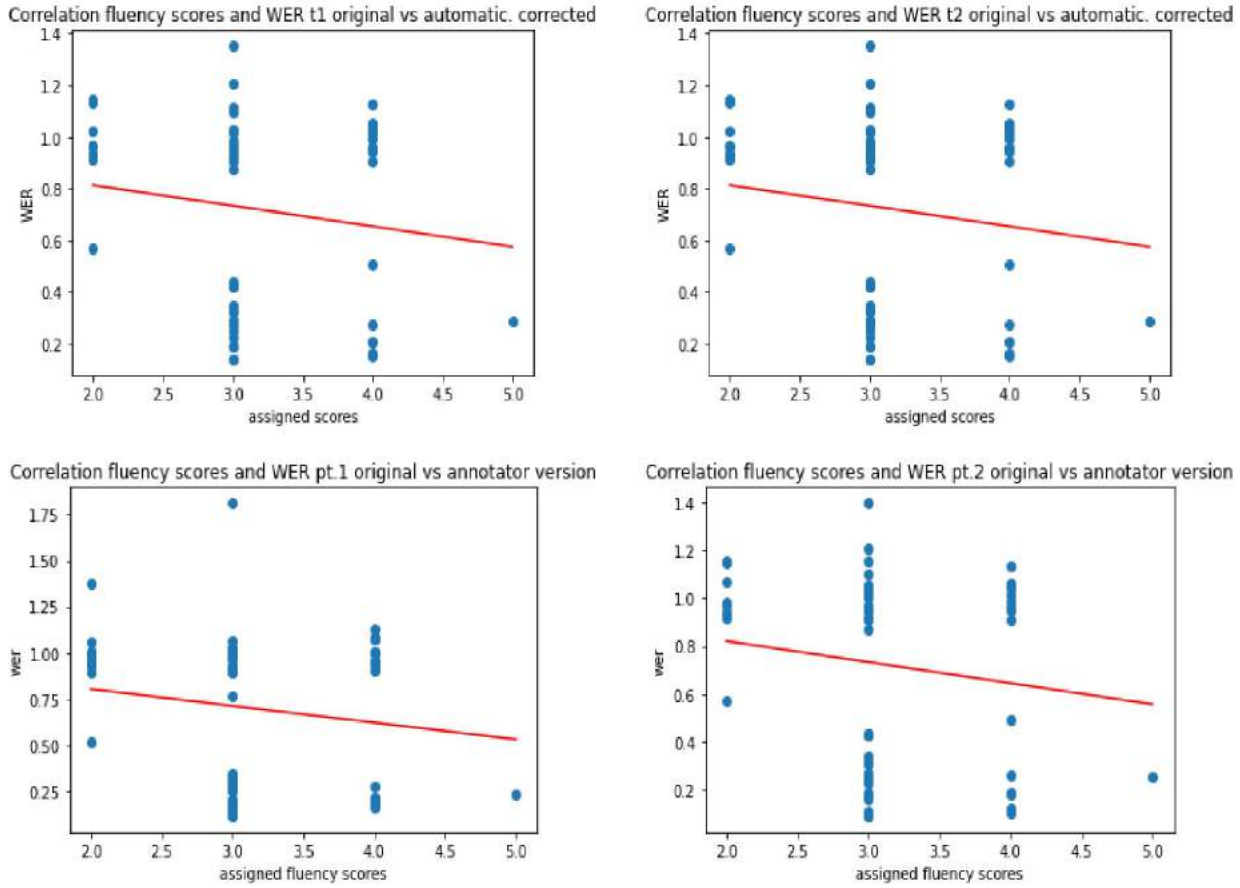


Figure 56: Correlation between ASR - corrected transcriptions and students' fluency assigned scores

The plots in the first row relate to the WER between the ASR transcriptions and the automatically corrected versions using LanguageTool and their correlation with the assigned fluency score. Instead, the plots in the second row represent the correlation between the WER of the ASR transcriptions and the manual corrected versions with respect to the fluency scores. In both cases we observe that those examinations which received a score below 4 for fluency in pronunciation tend to present a higher WER than those with scores between 4 and 5. The maximum WER values increase in the case of automatic and manually corrected transcriptions with the inverse correlation between the two values becoming marginally more pronounced. Overall, we can notice a trend in the above scatter plots, however, given the limited amount of data, the obtained results cannot be considered definitive.

From this analysis we can, therefore, confirm the requirement for human correction of the ASR outputs in this case, whereby the English speeches did not come from native speakers and were not recorded with high quality audio equipment. On the other hand, it is understandable that the unadapted system exhibited inaccuracies, which were eventually levelled out by the annotator in order to use the best possible version in our experiments. The alternative solution for these types of issues would be a specifically designed ASR system for the recognition of non-native English speakers.

7.3. Errors and linguistic features analysis

Prior to the models' automatic predictions of the exams' levels and scores, we needed to acquire an automatic correction of the transcribed learners' monologues. For this purpose, we once again resorted to LanguageTool,

although this proof-reader is mainly used for the correction of texts of a written rather than oral nature. In this case, however, its function was to point out errors of grammatical, style, register and lexical type, rather than relating to punctuation. Thus, we could obtain a version to compare with the original and to pass on to the base model as we had already done for the other languages.

After acquiring the automatically corrected versions of the exams, in the same way as we did for the written texts in Chapter 5, we have also considered the errors identified by LanguageTool, this time comparing the original ASR transcriptions and those corrected by the annotator. In fact, we expected that the percentages related to the number of errors would result higher in the former than in the latter since the raw transcripts contained word recognition errors.

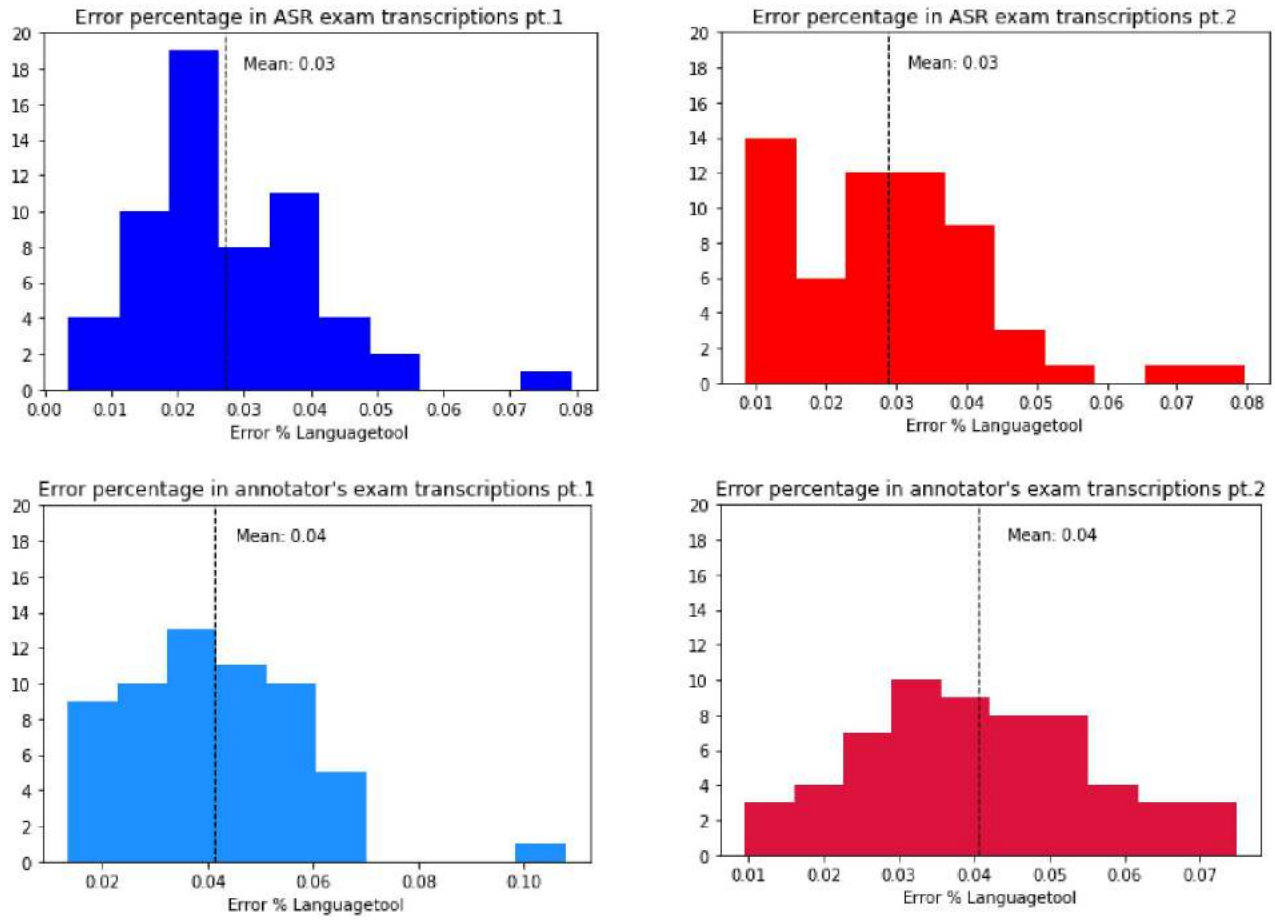


Figure 57: Errors percentages in ASR (top) and manual transcriptions (bottom) of students' oral exams

Despite the fact that the automatic exams transcriptions were incorrect, the average error percentage detected by LanguageTool was one hundredth lower, i.e. 0.03%, while the average error percentage for the texts corrected by the annotator amounted to 0.04%. This is due to the fact that the recogniser is partly capable of handling incorrect spoken inputs, as it maps them to existing words in the target language that possibly respect the most common language constraints. Indeed, given also that the difference between the ASR and manual transcriptions is relatively minimal, it is reasonable to assume that LanguageTool represents a valid tool for corrections, even if the typology of the texts supplied to the system differs from the usual essays, emails and letters involved in the assessment of written language skills.

More specifically, the tool was able to identify the quantified and categorised errors represented in the following bar plot.

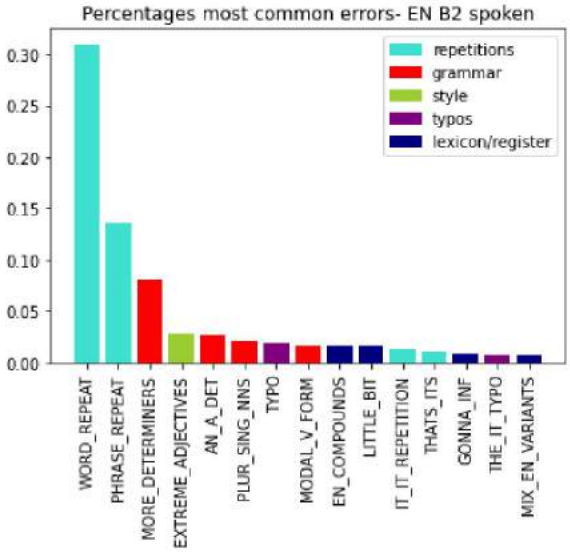


Figure 58: LanguageTool detected errors in oral English exams

Based on Figure 58 above, we can observe that the highest number of errors automatically detected corresponds to repetitions of words and phrases, typical of the oral modality, and often present in the speeches of language learners (Saito 2017), especially in unnatural situations that can cause stress, such as tests, or in which there is a limited amount of time to plan the discourse online. In addition to these, to a lesser extent there are grammatical errors, such as the use of multiple determiners or incorrect modal verb forms, represented by the red bars, and errors of register and lexicon relating to the use of informal language forms or a mixture of American and British terms. Moreover, self-corrections were also reported, corresponding to the cases represented by "more determiners", "IT_IT_repetitions", "that's - it's" and "the - it" typos in relation to LanguageTool matching norms. Overall, however, the findings are in line with what would be expected of B2 level English language learners (see Figure 15.4 and Figure 19.3).

Moreover, we extracted the percentage of errors detected by LanguageTool and correlated it with the scores assigned by the human examiners. The reason for this was to attempt to reveal a link between the efficiency of LanguageTool and human ratings of students' proficiency from transcripts.

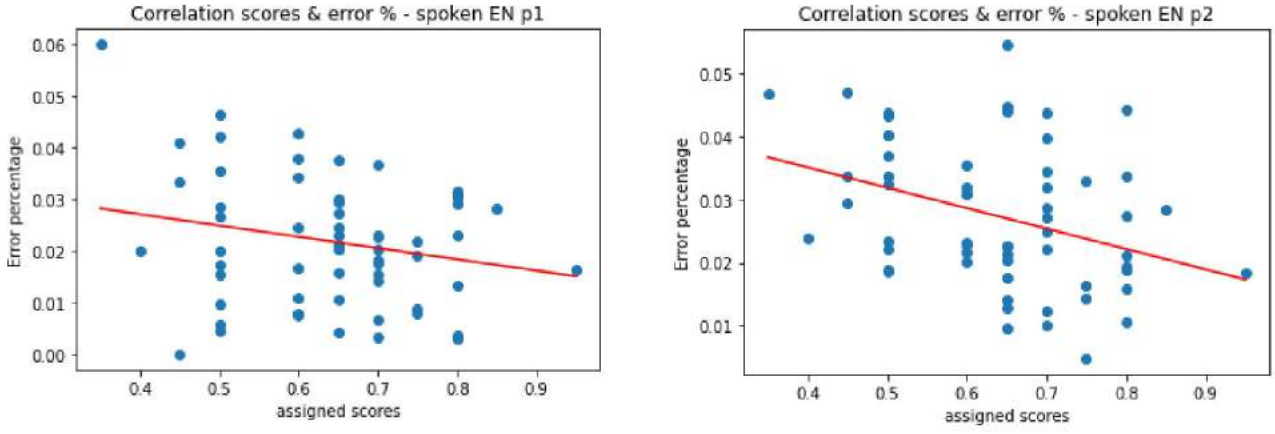


Figure 59: LanguageTool percentages of detected errors in correlation with human assigned scores for oral exams

As can be observed from the scatter plots in the figure above, there is indeed an inverse correlation between the automatically detected errors and the scores assigned to the English language learners. As the score increases, the percentage of errors seems to decrease, following the indications of the red line. This is even more evident for part 2 of the oral examinations than for part 1. Although the nature of these transcriptions deviated from the standard texts that the automatic tool usually corrects, we can assume that LanguageTool performed reasonably well in this task.

Additionally, we decided to consider the linguistic aspects of the oral examinations on the basis of the transcriptions to ensure that the scores assigned by the evaluators could be traced back to the characteristics of the learners' oral productions extracted in a more precise and objective manner. For this reason, we computed, slightly differing from the operations done for the other languages, the number of tokens and unique lemmas contained in each monologue. The results obtained from these considerations are described in detail below.

First, we tokenized and lemmatized the transcripts automatically obtained by means of the ASR system and corrected by the annotator. To do this, we could uniquely use Stanza, since the other tool to which we resorted for the analysis of the written examinations, namely *textcomplexity*, exclusively worked on texts of that nature and not oral ones. Our expectations were to find as many unique tokens and lemmas as possible in the students with the highest received scores and vice versa for the others. Indeed, as the plots in the figure below show, this would seem to be the case for both part 1 and part 2 of the oral examination.

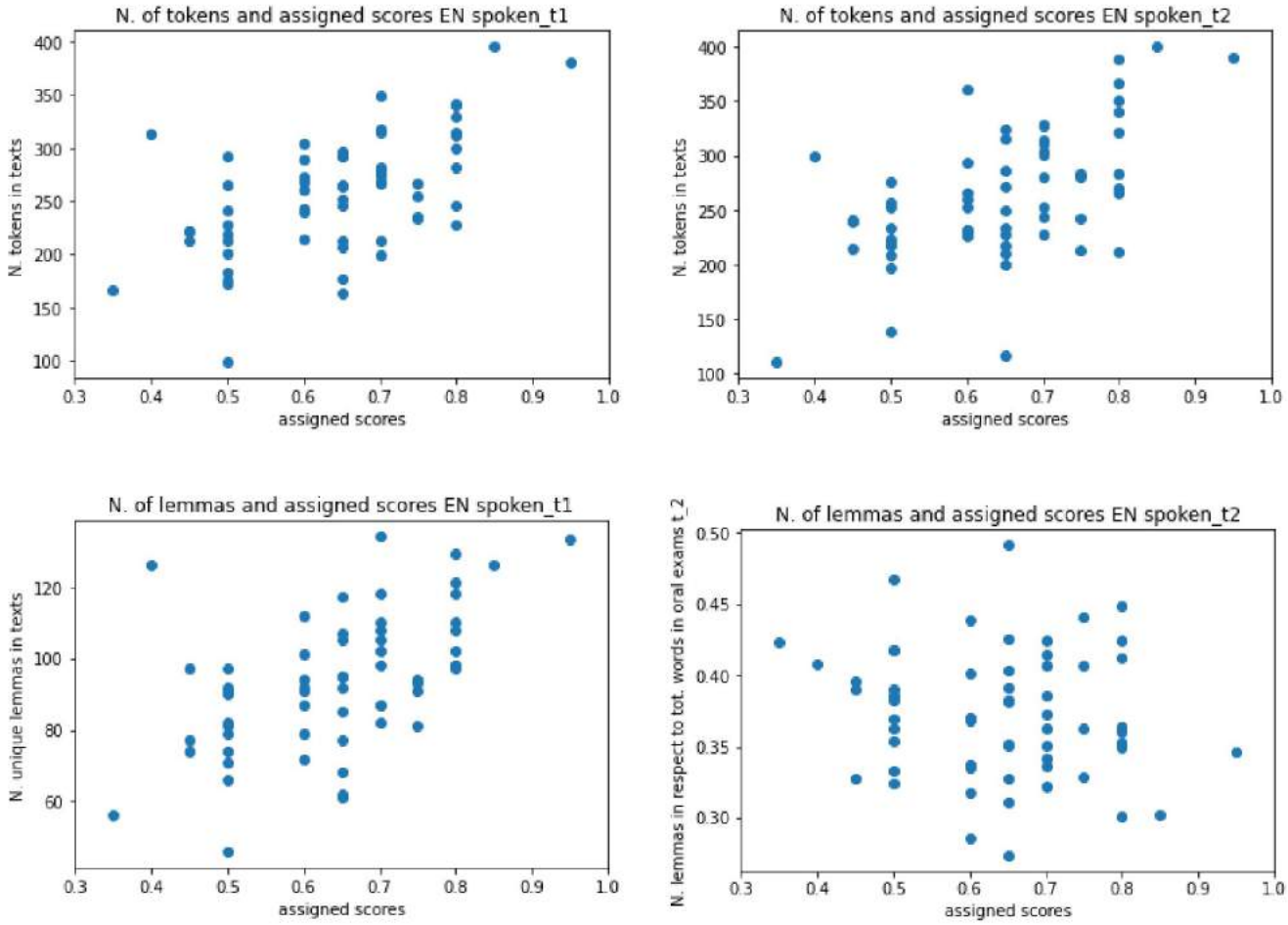


Figure 60: Number of tokens and lemmas in oral examinations part 1 (left) and part 2 (right)

As a matter of fact, we can notice an ascending trend displaying a direct correlation between the scores assigned by the Language Centre experts and the number of tokens and lemmas found in each transcription. However, there are also some outliers, for example the exams which got a score of 0.4. This may indicate either an imprecise human evaluation or the incompatibility of this typology of metric for oral examinations. Given the limited number of exams, also, we may expect this sample not being really representative of the oral test sessions carried out at the Free University of Bozen-Bolzano.

Given the apparently strong association between the number of uttered words and students' proficiency, we decided to make an attempt using the latter as an index for the classification of competence levels. Therefore, we considered the balanced accuracy we may have obtained if we had used this feature, namely the number of words pronounced in the students' monologues, as indicators of their competence in an evaluation task.

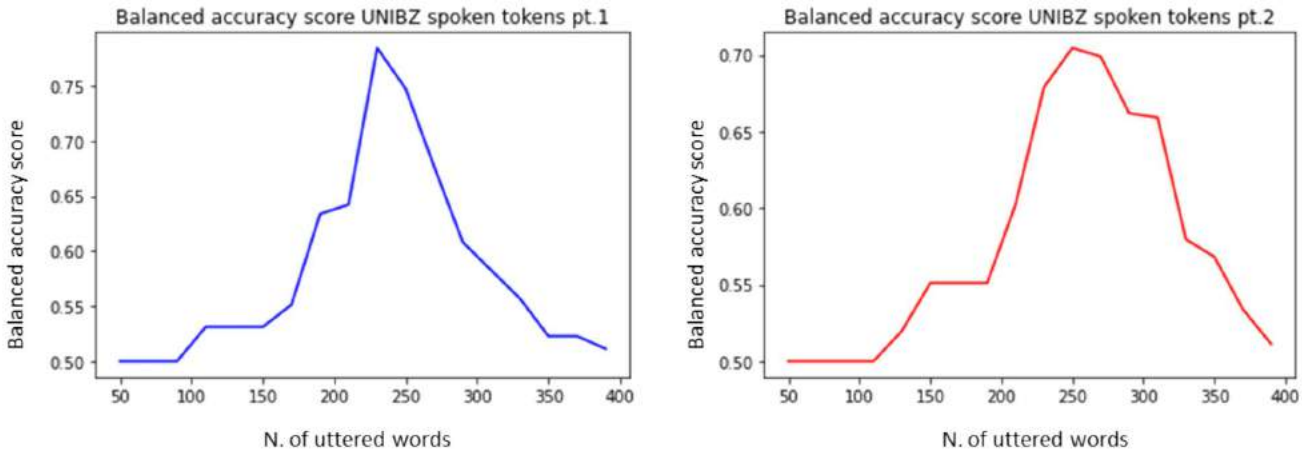


Figure 61: Balanced accuracy assigned scores depending on the number of uttered words

The curves above, hence, display that the number of words on the x-axis may be good predictors for students' proficiency using the evaluation system provided by the examiners, the scores of which are disposed along the y-axis. This information could be, therefore, added to the ones considered, although it is highly interpersonally variable and may not constitute an optimal evaluation metric for language proficiency in oral examinations. This represents one of the reasons why we preferred to proceed in a similar way to what we did for the written exams, providing the architectures the entire transcribed monologues of the students and their detailed automatic corrections.

7.4. Experiments with pre-trained English models on oral exams

Due to the restricted amount of data available for spoken English language proficiency, we eventually discarded the idea of building a neural model exclusively based on 60 examinations. Instead, given that the data concerned the level B2, for which we had valid transcripts, we decided to conduct experiments of diverse nature. Since the test data belonged to a single class, B2, and had different numerical scores assigned to them, we experimented with both predicting pass or fail and assigning discrete exam scores. Below are details about the procedures and results obtained with the original model based on the EFCAMDAT dataset and the one based on the CLC- FCE dataset. We employed them in their initial format, as well as making changes in their architecture to predict instead of the class the exam scores to be assigned.

7.4.1. EFCAMDAT model applied to oral examinations

Initially we used the most accurate model of English written examinations, namely EFCAMDAT, to conduct three different experiments with it. In the first case we employed the original model to predict different CEFR levels without passing the levels assigned to the exams by human evaluators. In the second and third case, we modified the model to obtain negative or positive scores instead of classes and assign either the maximum or the average of them to the texts in order to predict pass or fail. Since these are a first and second part of the oral exam taken by each of the 60 students, the experiments were performed twice on each set of examinations.

7.4.1.1. CEFR levels predictions without assigned levels

The first set of experiments performed on the oral data from the Free University of Bozen-Bolzano were made with the model trained on the written data of EFCAMDAT, which had reached an accuracy of 97% on written examinations. We aimed at analysing which of the CEFR levels the oral exams would be assigned to by passing the model the original transcribed versions without the learner's hesitations and the versions automatically corrected by LanguageTool.

Our aim with this initial test was to check how the pre-trained model would automatically classify the passed and failed tests, which according to the examination format should more or less belong to the B1 and B2 CEFR levels. Therefore, we took all 60 examinations as test set and performed the model evaluation.

The majority of the students' monologues were classified as expected between level 2 and level 3, meaning B1 and B2, both in the cases where we considered the assigned score (see Figure 62) and the passed/failed exams (see Figure 63).

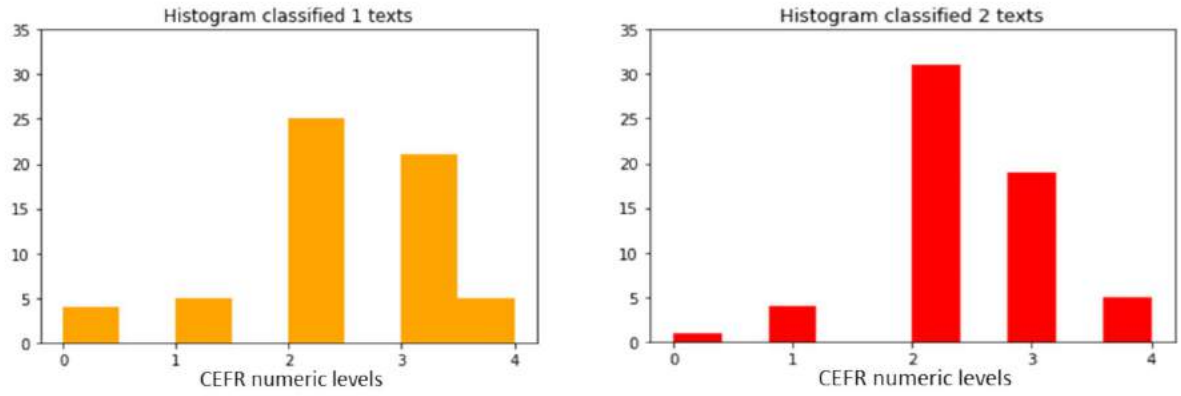


Figure 62: EFCAMDAT automatic exams classification of part 1 (left) & part 2 (right) of oral exams

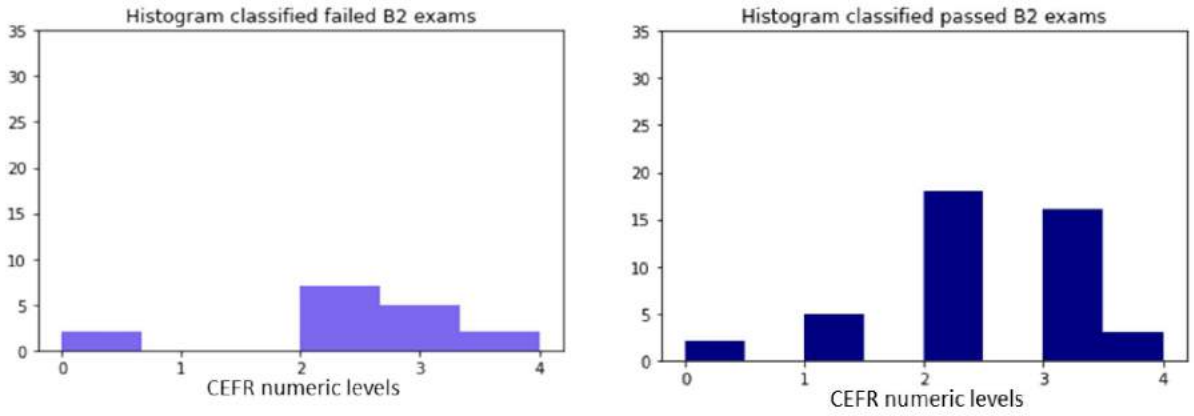


Figure 63: EFCAMDAT automatic exams classification of failed (left) and passed (right) oral exams without levels

More in detail, if we observe the classification of the passed exams in Figure 63, we notice that indeed most of them appertain to the intermediate classes, while relatively fewer exams were assigned to the A1, A2 and C1 levels. Furthermore, paying attention to the classification of exams in part 1 and part 2 in Figure 62, produced by the same 60 students, we can notice a peak of more than 30 exams classified as B1 and 23 classified as B2.

Unfortunately, as we do not hold oral examinations belonging to the other CEFR levels, it is not possible to compare this classification with other proficiency classes. Considering that the model actually received only the original texts of level B2 solely and their corrections as input, the results appear to be discrete.

7.4.2. Conversion of EFCAMDAT into an inferential model

In a second set of experiments to better visualize the automatic classification of texts executed by the model, which was trained, however, on data related to written and not oral tests like these, we modified its architecture. The changes primarily concerned the final layer, which previously outputted a number between 0 and 4 for the CEFR competence levels from A2 to C1. In particular, we switched the activation function of the dense layer from *softmax* to *linear activation*. When making the predictions, we also removed the *argmax* function. The latter would have returned the maximum values along the axis of predictions mapped to the existing CEFR levels, but we needed a single float, since our aim was to only extract the exact score assigned to the fourth class corresponding to the B2 level. As we were using data from B2 English exams to test the model, we considered this to be an appropriate solution to adopt, partly because also their scoring by experts was binary, passed or failed, and not distributed across multiple competence levels.

We proceeded to extract the scores provided by the model, either positive or negative, to the oral exams' transcriptions and compared these results with those actually provided by the language experts. We decided to make two distinct experiments:

1. Experiment: we first passed separately one part of the oral exam and the other, we considered the respective predictions and then averaged them[7].

2. Experiment: we concatenated the transcripts of the two parts, along with the versions automatically corrected by LanguageTool, into a longer single text on which we performed the predictions with the model.

More details on the modalities of these two attempts and the obtained results are available in the following sub-sections.

7.4.2.1. Pass & Fail predictions with the oral exams means

For experiment number one with the inference model, we first made predictions on the automatic transcriptions of the first part of the oral exam, and the version automatically corrected by LanguageTool. Then we repeated the same procedure on the second part of the exam. Given that the EFCAMDAT model was trained on written data assessed at five different levels, namely A1, A2, B1, B2, and C1, we restricted the considered predictions to the B2 or fourth class. We then calculated the mean over them for the assignment number 1 and number 2 of each of the 60 students who took the test. Below is a graphical representation of the new pipeline of the model with the modified architecture.

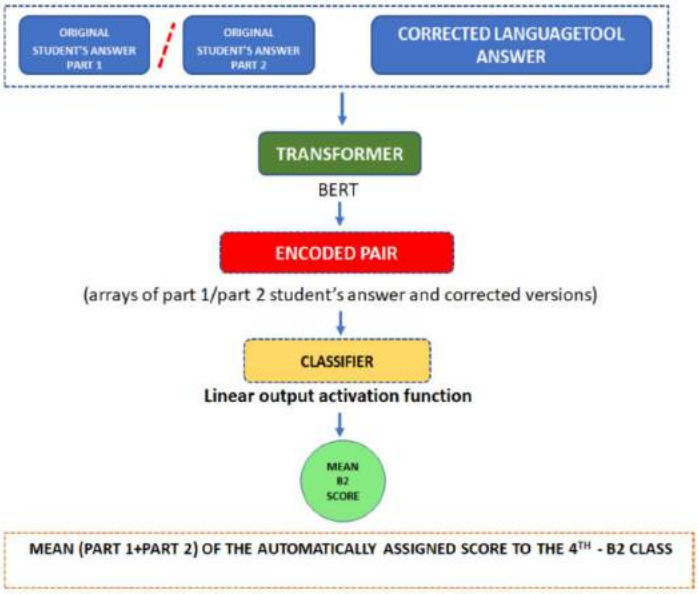


Figure 64: Modified EFCAMDAT model architecture for exams scoring[8].

[7] The reason for this last point is that although the students delivered two different monologues, they were assigned one overall score.

[8] This new model receives as input in turn part 1 and part 2 of the oral examination, together with the corrected version coming from LanguageTool. The pair is encoded, and the classifier receives the resulting arrays concatenated into one. Using the linear activation function, the output is the average result obtained for the fourth class of the five, A1, A2, B1, B2 & C1, namely the one related to the B2 level we are interested in.

We divided the predictions made for the students who passed the exam, scoring more than or 0.6 points, and those made for the rejected students, who scored less than 0.6 points. This separation was adopted to check if the model could have classified the students' exams binarily by assigning them scores less than 0 if not of B2 level, and vice versa for those with higher proficiency.

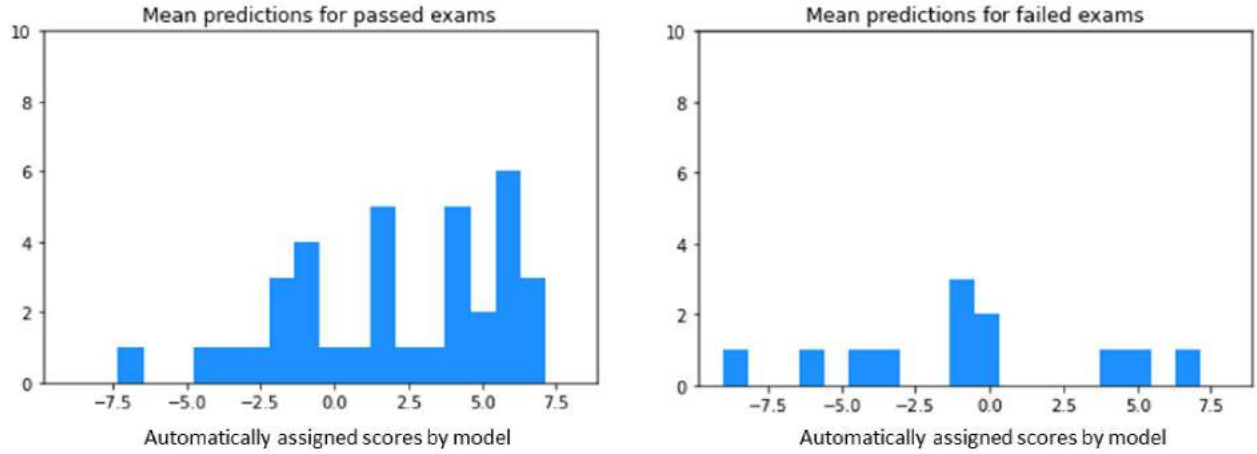


Figure 65: EFCAMDAT inference predicted scores for passed (left) & failed (right) mean oral exams parts

As can be observed from the histograms in Figure 65, in the case of passed oral examinations, most of them obtained a positive score, and were therefore classified correctly. On the contrary, in the case of failed oral exams, most of them received a negative score, although a limited number were nevertheless considered positive. This score value was particularly high in the case of the texts from the transcriptions of part 1 of the oral exams.

Since our dataset is unbalanced, 45 passed exams and only 15 failed exams, we decided to use the *balanced accuracy score* as a metric of evaluation for the model's performance. This metric proves to be particularly suitable for binary classification tasks when the distribution of elements per class is unequal. Balanced accuracy results from the sum of *sensitivity*, or true positive rate, and *specificity*, or true negative rate, divided by two, meaning the number of classes. Differently, the typically used accuracy score appears to be more suitable for multilabel classification tasks. It corresponds to the sum of true positive predictions and true negative predictions divided by the sum of positive and negative predictions. However, accuracy does not account for imbalance datasets and can be, therefore, misleading.

Below we compare the results obtained with different thresholds, between -8 and 8, on oral examination data by considering balanced and unbalanced accuracy score curves.

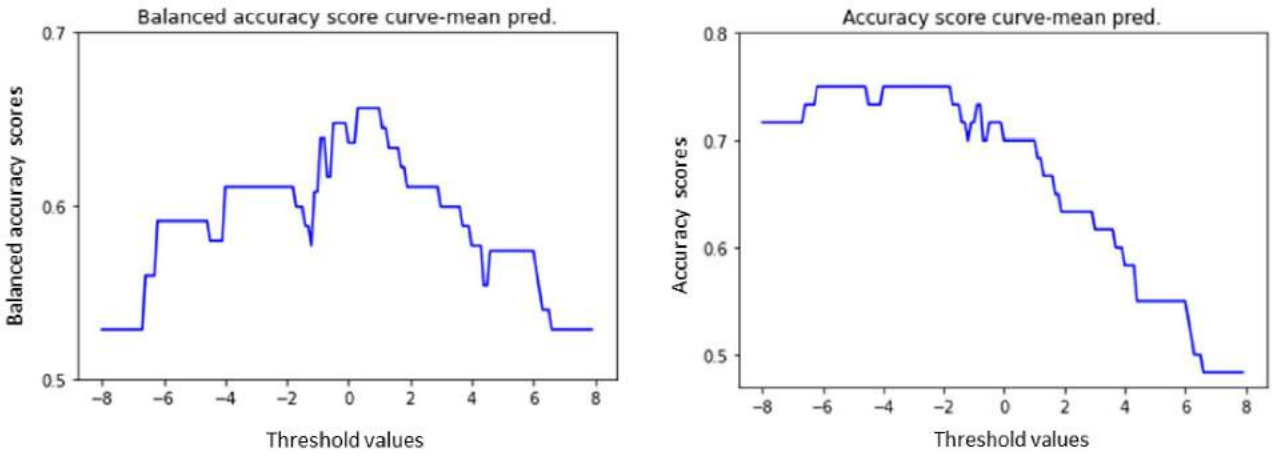


Figure 66: Accuracy (right) and balanced accuracy (left) curves for the EFCAMDAT scorer model

As can be observed from the curves in Figure 66, the balanced accuracy curve takes better account of the inferiority of data for failed exams, i.e. with values less than 0, with respect to data for passed exams with values greater than 0. In particular, we obtain the highest accuracy score, equal to **65%**, with a threshold between 0.30 and 1. In contrast, with the standard accuracy it is not possible to discern a clear division between failed and passed examinations, although the obtained value is equal 65% in this case as well.

7.4.2.2. Pass & Fail predictions with concatenated oral exams

For the other experiment with the EFCAMDAT inference model, instead of passing the part 1 and part 2 transcriptions separately to the model and then averaging the results, we concatenated the two together into one (see Figure 67). In this way we got a single prediction per student that can be traced back to the passed and failed scores assigned by the University Test Centre language experts. As with the previous experiment, the model receives as input the automatic transcription of the learner's monologues and their corrected version generated by LanguageTool, this time concatenated.

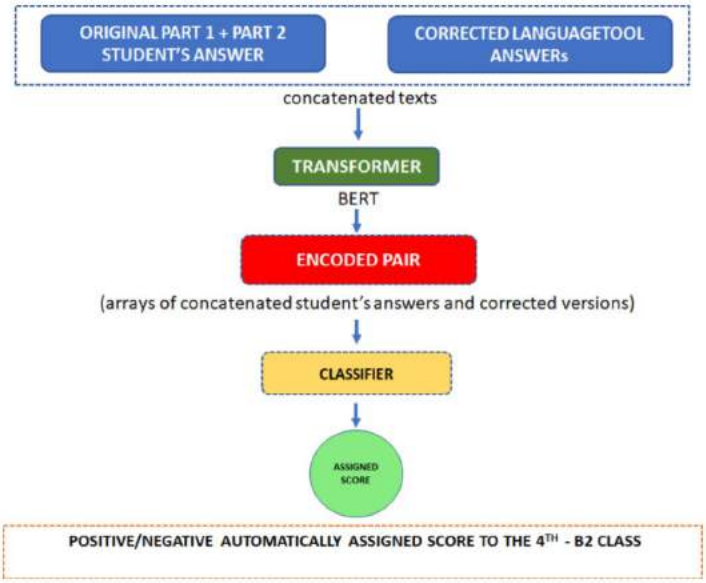


Figure 67: EFCAMDAT modified architecture for predictions on concatenated part 1 and part 2 of oral exams

The obtained outputs are the automatically assigned scores for the exams that are considered sufficient and those that are not. The results are represented in the following histograms.

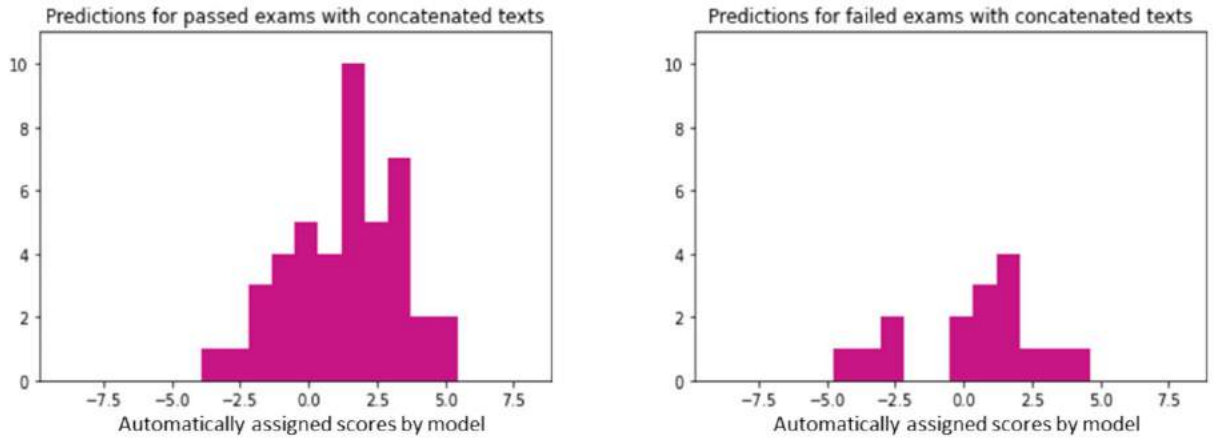


Figure 68: EFCAMDAT inference predicted scores for passed (left) & failed (right) concatenated oral exams parts

Again, as in the previous experiment, most of the passed exams received a score greater than 0, while vice versa applies to failed exams. However, in both cases there are still some incorrect assignments. This means that exams that were actually rated as sufficient for the B2 level were at times assigned scores lower than 0 by the system, thus designating them as insufficient. However, it must be considered that the model was trained not for binary classification on a single class, but for multi-class classification on five. In addition, the originally used data were written texts such as emails and letters, far from the oral productions of this case-study. For these reasons we did not expect a spotless performance on this data.

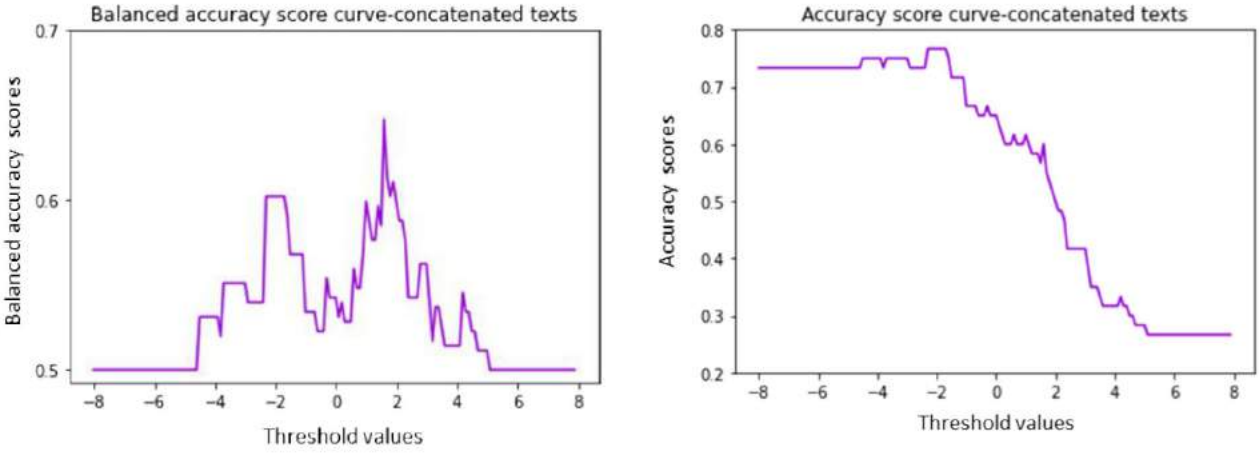


Figure 69: Accuracy (right) and balanced accuracy (left) curves for EFCAMDAT inference model on concatenated exams parts

To evaluate the results of the model we again resorted to balanced accuracy instead of standard accuracy. In this manner we receive a clearer picture of the distinction between passed and failed examinations by applying different thresholds. In particular, the best results were obtained with a threshold between -2.3 and -1.7 achieving **60%** accuracy. Although the standard accuracy plot shows higher accuracy values, maximum 75%, it is hard to identify a threshold that clearly divides passed and failed examinations. This is due to the fact that the binary nature of the classification task is not accounted for by this evaluation metric, nor is the unequal distribution of the data.

7.4.3. EFCAMDAT experiments summary

With regard to the different experiments performed with the pre-trained EFCAMDAT model for written examinations over five distinct proficiency levels, we achieved diverse results on our oral data. In the first one in which we attempted to automatically classify exam transcripts without passing any labels, we obtained a discrete outcome, considering that most of the oral exams were classified between B1 and C1 level, with the B2 target placed in the middle between these classes.

In other experiments conducted with the EFCAMDAT inference model to predict positive or negative scores along different thresholds, we noticed better results by averaging the scores assigned to the respective parts of the examination. In fact, the latter experiment achieved a balanced accuracy of 65%, compared to the one carried out with the concatenated transcriptions of part 1 and part 2 where the maximum value was 60%.

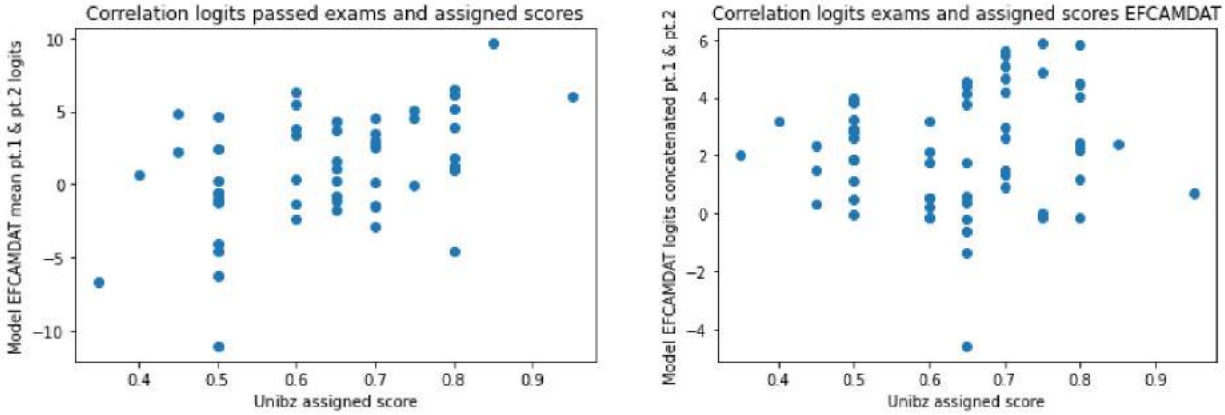


Figure 70: EFCAMDAT inference model correlation between mean (left) & concatenated (right) transcriptions and human assigned scores

Figure 70 above summarises the scores assigned by the EFCAMDAT inference model by first passing single part 1 and part 2 transcriptions and corrections as inputs, and then averaging the results obtained for class B2, second by passing the concatenated texts and corrections, and obtaining a single assigned score. The scatter plots reveal a conceivable correlation between these and the totals assigned by the evaluators of the language centre of the University of Bozen-Bolzano. In fact, an upward trend can be noted, especially in the case of the results obtained with the mean of the scores for each part, between the negative and the lower scores and, vice versa, with the higher and positive scores respectively assigned by humans and the model.

7.4.4. CLC- FCE model applied to oral examinations

After having carried out a number of experiments using the first English language model, namely EFCAMDAT, we decided to undertake further experiments based on the other English model, namely the one created with the CLC-FCE dataset. Having been conceived mainly with data of intermediate English learners, we expected it to be more in line with the class to which the exams of the Free University of Bozen-Bolzano exclusively belong, i.e. B2. First, we replicated the class prediction experiment without passing CEFR level labels but using only the transcripts of the original oral exams and their corrections made by LanguageTool. Secondly, we modified the architecture to

transform it into an inferential model capable of predicting the scores for the texts in order to decide whether they were sufficient or insufficient for the B2 level, as human examiners did. Thus, we once used an architecture that individually considered part 1 and part 2 of the oral examination and averaged the assigned score, and another time we concatenated the texts together so as to obtain a single result. More details on each of these experiments can be found in the next subsections.

7.4.4.1. CEFR levels predictions without assigned levels

For the following experiments we decided to use the CLC-FCE base model. The former had obtained an accuracy of 61% on the test set, where three of the original five classes of the dataset were included. As we did with the BERT-based model trained on the EFCAMDAT data, in the first experiment we aimed to predict the CEFR classes of oral exams by passing as input only the original exams transcripts corrected by the annotator and those obtained after processing them with LanguageTool.

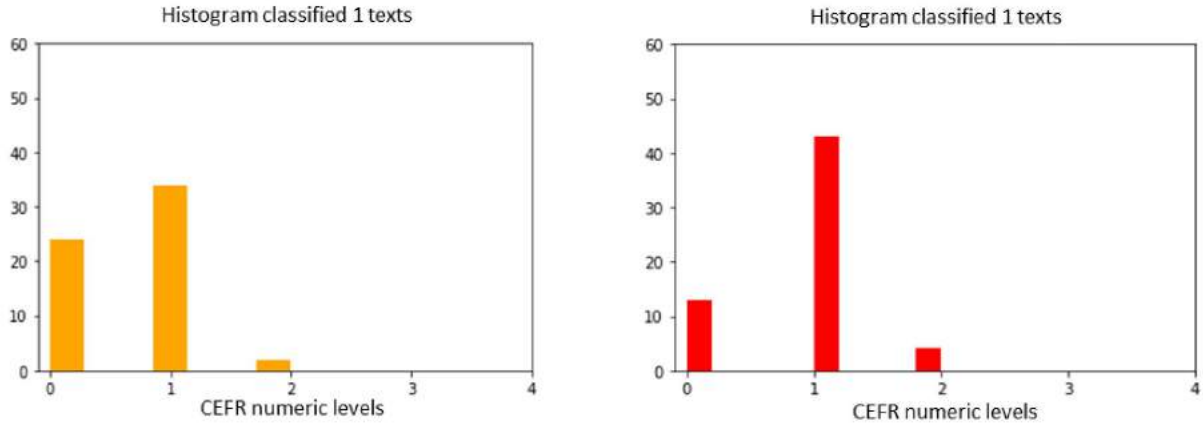


Figure 71: CLC-FCE automatic exams classification of part 1 (left) and part 2 (right) oral exams without levels

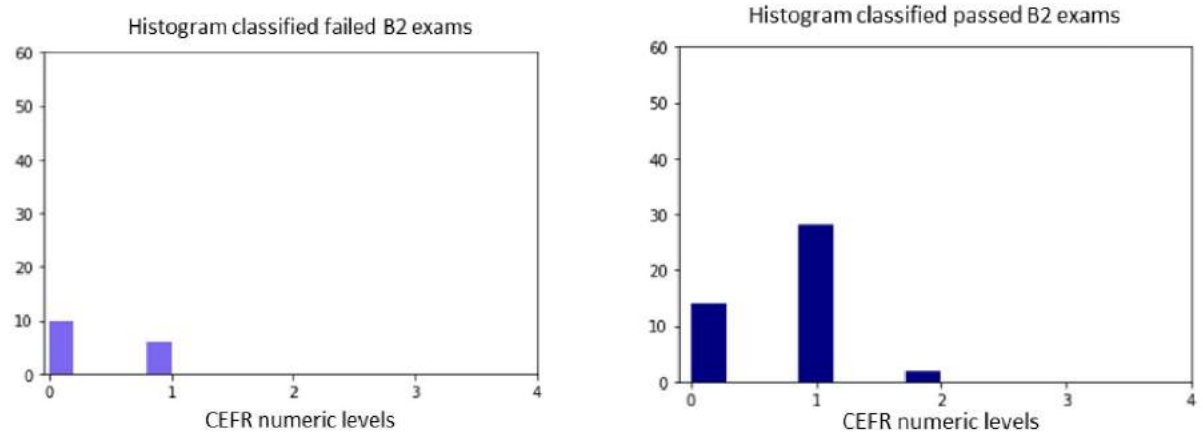


Figure 72: CLC-FCE automatic exams classification of failed (left) and passed (right) oral exams without levels

As can be observed from the above bar plots (Figure 71 & 72), the predictions of this model are distributed along the first three classes, i.e. level A1, A2, B1. None of them, however, belongs to the class of the majority of the exams contained in the dataset from the Free University of Bolzano, meaning the B2 level. Nevertheless, from the

predictions made for the failed exams in respect to those of the passed exams (see Figure 72), we notice that indeed the model was able to distinguish higher level exams from lower ones. The latter were predicted within class 0 (A1) and 1 (A2), while the former included also class 2 (B1). Overall, given the limited amount of available data, we are not able to spot such a clear distinction as with the EFCAMDAT model (see Figure 62 and 63).

7.4.5. Conversion of CLC-FCE into an inferential model

Following the other experiments performed with the EFCAMDAT model trained on English written data, we modified also the CLC- FCE architecture to receive not the CEFR numeric classes as outputs but positive and negative scores assigned to each exam. Therefore, we proceeded to modify the architecture changing the activation function of the Dense layer to linear activation and removing the *argmax*. In this way, we selected the models' predictions with index 3, corresponding to the B2 level, and considered once the mean of the results obtained first on part 1 and then on part 2, and once their concatenated versions. Again, the model received as inputs the original text version and the LanguageTool corrected ones.

7.4.5.1. Pass & Fail predictions with the oral exams means

In the first inferential experiment, meaning the one expected to assign scores to part 1 and part 2 of the oral examinations respectively, we assume to be able to distinguish a threshold between passed and failed exams. This means that according to our predictions, the inferential model will be capable of differentiating them for lower and higher proficiency.

Below are the results obtained for passed and failed exams from the CLC-FCE inferential model. The plots represent the mean average of the scores assigned to each part.

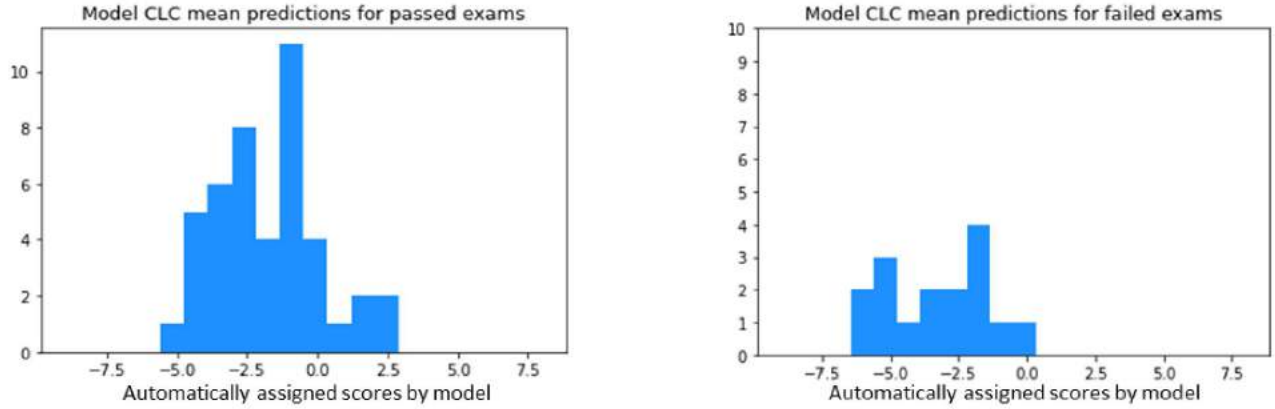


Figure 73: CLC-FCE inference predicted scores for passed (left) & failed (right) mean oral exams parts

Considering the plots in Figure 73, we notice that the average prediction for part 1 and part 2 of the oral examinations in the case of passed examinations receives scores ranging from -5 to 2.5. In contrast, failed examinations exhibit negative scores between 0 and -7.5 as we would expect since the learners did not demonstrate, also according to the human experts, a sufficient level of competence to pass the test.

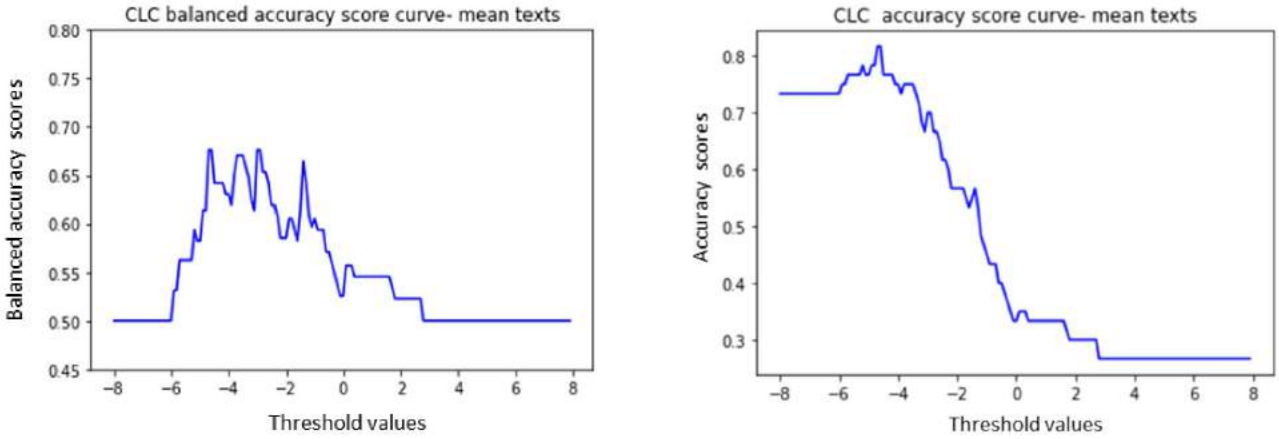


Figure 74: Accuracy (right) and balanced accuracy (left) curves for the CLC-FCE scorer model

The outcomes anticipated by the histograms in Figure 74 are indeed confirmed by the balanced accuracy curve above. An apparent threshold is visible at value 0, while that is not as clear in the standard accuracy curve, which does not account for the imbalance of the two classes. The best thresholds, meaning -3,5 and 3,7 confirm an accuracy of **67%**. The latter results better even in respect to the original results obtained on the written examinations CLC-FCE test set.

7.4.5.2. Pass & Fail predictions with concatenated oral exams

For the second round of inferential experiments with the CLC-FCE modified model, returning automatically assigned scores and not the CEFR levels of competence, we concatenated the texts of the two examination parts together into one. In this way we got a single prediction per student that can be mapped to a complex score assigned by the language experts of the Free University of Bozen-Bolzano. As with the previous experiment, the model received as input the automatic transcription of the oral exams and its corrected version generated by LanguageTool.

The outputs are the automatically assigned scores for the exams that are considered passed and those that considered failed. The results are displayed in the following histograms.

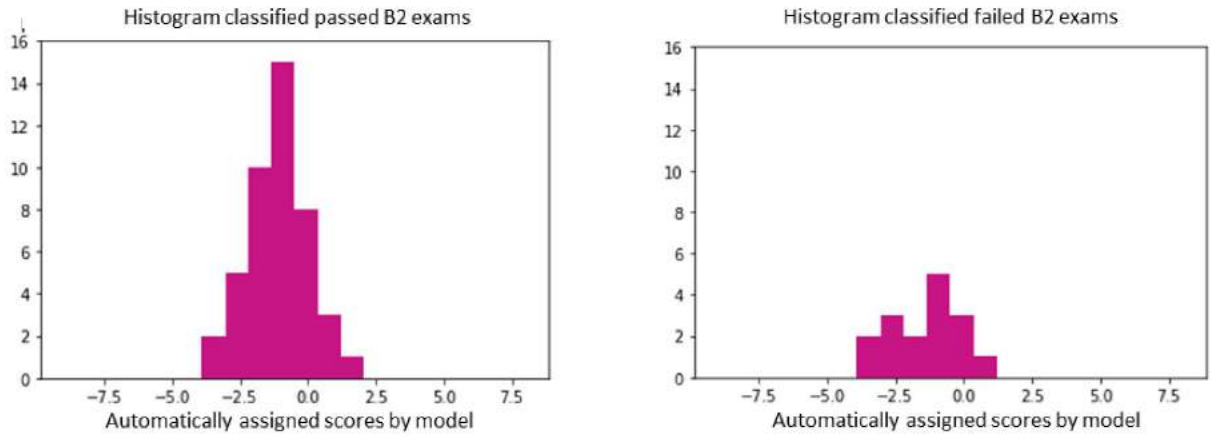


Figure 75: CLC-FCE inference predicted scores for passed (left) & failed (right) concatenated oral exams parts

From the figures of the scores assigned for passed and failed examinations, no significant distinction can be observed. Among the passed examinations, some received negative scores between -4 and 0, while others received positive scores between 0 and 2. However, we would have expected negative scores to be assigned to failed exams exclusively. On the contrary, looking at the second histogram, we see that this model actually assigned more than 80% of failed examinations a negative score. This indicates that in this case the architecture could, indeed, distinguish the two classes.

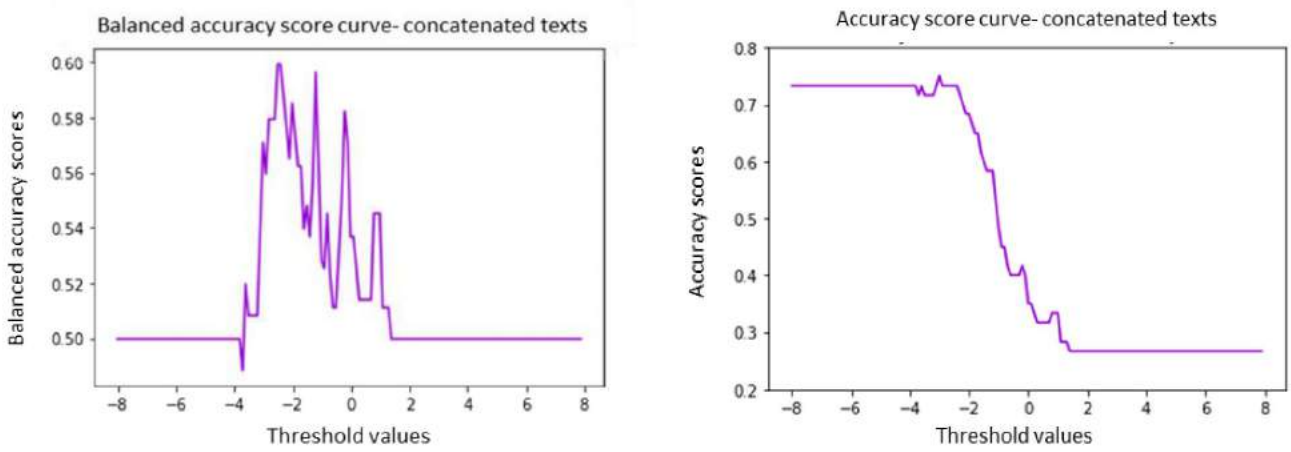


Figure 76: Accuracy (right) and balanced accuracy (left) curves for CLC- FCE inference model on concatenated of oral exams parts

The previously observed results are confirmed by the balanced accuracy curve with different thresholds between -8 and 8. In fact, the best value is obtained with the threshold -2 separating sufficient and insufficient exams and corresponds to **60%**.

7.4.6. CLC-FCE experiments summary

Considering the different experiments performed with the English CLC-FCE model, trained in a similar way to EFCAMDAT but with less accurate results given the unbalance and the reduced size of the dataset, we obtained outcomes in line with our expectations.

In the first experiment, in which we used the original model to classify transcriptions of oral examinations without passing any labels, we found incorrect results whereby the texts were classified as belonging to levels below B1, whereas the expected result was the opposite.

However, in the other experiments, carried out by transforming the model into inferential to obtain exam scores, we surprisingly obtained values that were quite similar and even better than the ones from the EFCAMDAT model. This last statement proves to be true if we consider the balanced accuracy obtained in the rating of part 1 and part 2 of the oral examinations, for which we subsequently calculated the average. In this case, the percentage obtained is 67%, whereas with the EFCAMDAT model it was 65%. In the case of concatenated exams, however, the accuracy achieved is the same, namely 60%.

Finally, in the following figure we compare in detail the results obtained by the CLC-FCE model in assigning scores first to the two texts individually and then to the texts concatenated into a single one.

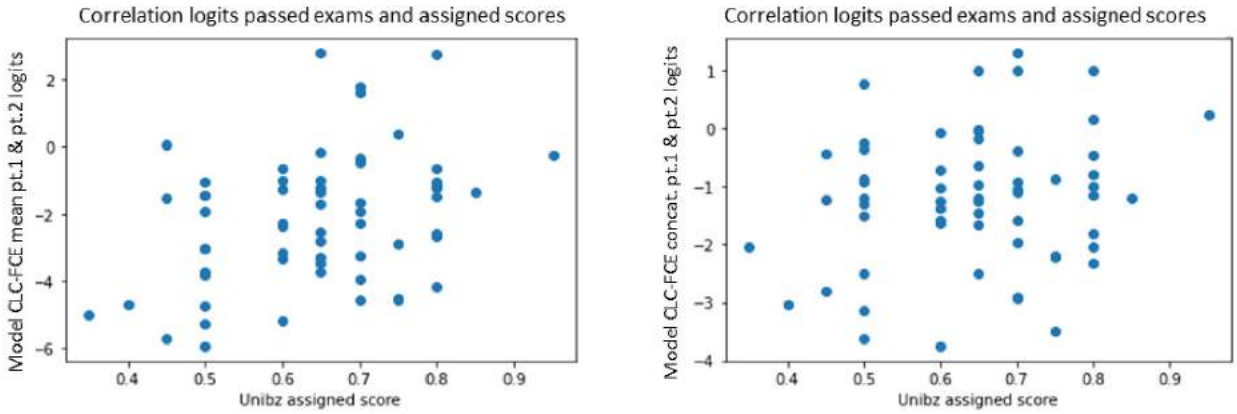


Figure 77: CLC- FCE inference model correlation between mean (left) & concatenated (right) transcriptions and human assigned scores

As can be seen from the plots above, comparing the scores assigned by the evaluators of the Language Centre of the Free University of Bozen-Bolzano with those automatically assigned by the model, we notice an ascending trend. As a result, the exams that received better human scores generally also obtained positive scores from the model.

7.7. Results with the EFCAMDAT and CLC-FCE models on oral examinations

On the basis of the experiments described in this section up to this point, for which we had to use the models already trained for the English language, we can draw the following conclusions. First of all, the oral examinations prove to have a different structure and characteristics than the written texts with which the two models have been trained. This undoubtedly constitutes a factor affecting their performance. In addition, the limited availability of data, moreover unbalanced, does not favour the models' optimal functioning.

As regards the results respectively obtained, summarised in Table 12 below, we derive that the best model of the two considering balanced accuracy while classifying the leaners' exams according to CEFR classes and scoring the transcripts of part 1 and part 2 of the oral examination is CLC-FCE, which exhibits 62.5% and 67% accuracy values. This had already been anticipated in the previous summary sections, where the results of each model were examined in detail (§ 7.4.2.3. & 7.4.4.3.)

Database	Balanced accuracy assigning scores to part 1 & part 2 (mean)	Balanced accuracy assigning scores on concatenated transcripts
EFCAMDAT	65%	60%
CLC- FCE	67%	60%

Table 12: Summary of English EFCAMDAT and CLC-FCE models experiments on oral examinations

If, on the other hand, we consider how the two models automatically classified texts by passing only the original transcripts of students' audios without hesitations and LanguageTool corrections, we can claim that the EFCAMDAT model turns out to make better predictions (see Figures 62 and 63) than CLC-FCE (see Figures 71 and 72). However, due to the limited number of examinations contained in this oral dataset, more distinctive and better results could be obtained if it were larger and more balanced.

7.8. Possible additional speech features

Given the nature of the data from the Free University of Bozen-Bolzano, meaning their orality, we decided to conduct further analyses regarding quantifiable factors that could determine students' competence and the received scores from the language experts. In particular, we considered the number of hesitations and the duration of pauses contained in each audio, together with the speaking rate present excluding the long pauses of the original recordings. The results obtained from these analyses are described in detail below.

First of all, starting from the written transcriptions of the learners' monologues corrected by the human annotator, we analysed the number of hesitations and the pauses length in correlation with the assigned scores related to their proficiency. We would expect more fluent students to generally display a lower number of hesitations in their speech and shorter pauses.

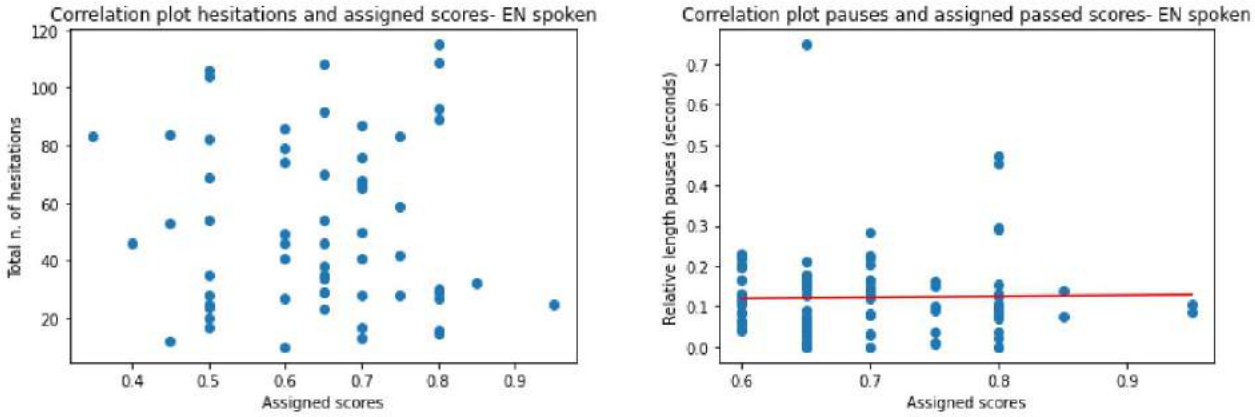


Figure 78: Number of hesitations and pauses' length correlated to assigned scores in oral examinations

The above plots display that, overall, students with higher scores tended to make less hesitations while speaking than the students who got lower scores. Nevertheless, on our limited dataset we can also spot some outliers, for example among the students who received 0.8 from the evaluators. This can be also found when observing the plot related to the length of pauses in seconds, where the correlation between these and the assigned positive scores is inverse, but yet relatively low.

Then, we considered the students' speaking rate, namely the total number of words uttered per minute. Our results are displayed in the plots below.

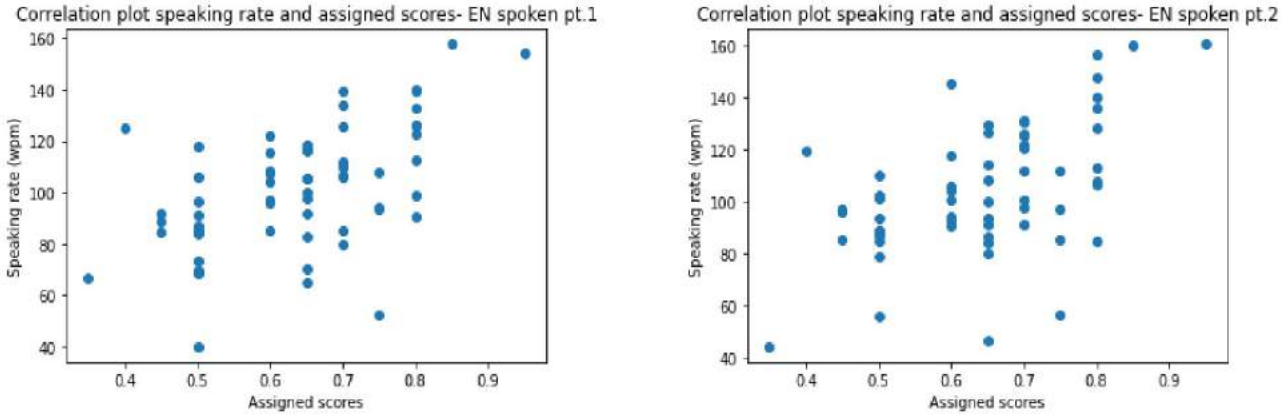


Figure 79: Average speaking rate in correlation with assigned scores in oral examinations

From the above figure we observe that also in this case there is a rather visible correlation between the assigned scores and the average speaking rate calculated considering the delivered speeches while excluding long pauses. Students with higher received scores appear closer to the average speaking rate of English native speakers, corresponding to values ranging between 120 and 150 wpm (Baese-Berk & Morrill 2015). This factor is probably highly associated with the impression evaluators' get from a student apparently confident and competent. However, this may also mislead the judgements, while pronouncing a higher number of words and speaking faster do not unequivocally translate into better fluency and proficiency in a language. An automatic evaluating system may indeed avoid being influenced by such a bias, making it more robust and efficient than human evaluators.

Despite these analysed aspects being possibly useful to human examiners when evaluating learners' speeches, they are without any doubt subject to inter-subject variability, apart than context dependent. For example in situations of stress and discomfort, people, especially second language learners, tend to be more dubious or instead faster at uttering sentences. On the contrary, in an informal situation where the speaker is relaxed, for example recalling a story or an event, they may be slower and make longer pauses or hesitations. Therefore, these numerically quantifiable metrics can be objective and provide resourceful additional information for the evaluation of language competences, but at the same time need to be contemplated paying attention to the whole learner's performance.

CHAPTER EIGHT: DISCUSSION OF RESULTS AND CONCLUSION

8.1. Discussion of overall results

In this thesis project, we applied automatic tools and BERT-based models for the assessment of adult language learners' competences in English, Italian and German. On the basis of the analyses and experiments conducted on the diverse considered datasets, we can draw distinct conclusions regarding the derived linguistic structures, the automatically extracted errors and the employed neural architectures.

First, as far as linguistic features such as number of unique lemmas, HD-D and MTLD in written texts are concerned, we can state that based on the observed results they are valid indicators of learners' competence. However, they undergo more or less marked variations depending on the task, the text length and some individual characteristics of the learner such as style or L1. In contrast, average sentence length, average dependency distance and dependents per word proved to be applicable parameters in our case to a limited extent due to the type of texts and their variability.

Secondly, for the errors automatically detected and corrected by LanguageTool, we noticed that comparing them with human corrections, when available, despite being less numerous, they were highly accurate and methodical. Moreover, they largely reflected the expected scale of learning in foreign languages, according to which at lower levels of proficiency errors are generally found concerning grammatical norms not yet fully memorised, while at higher levels these are mostly mistakes related to style or register.

Thirdly, as regards the different experiments carried out with neural architectures, we can confirm that all of them successfully performed the classification task. In particular, the models that received as input the original texts and the automatic corrections proved to effectively capture the latter and exploit them for a more efficient classification of the exams according to proficiency levels. In addition, the minor differences found in the cases where human and automatic corrections were used indicate that these are not extremely necessary to obtain a proper assessment of learners' competences with a Transformer model like ours.

Finally, experiments with oral examinations have proved the possibility of evaluating them automatically starting from pre-trained models. Nevertheless, for more precision, either more data would be needed to fine-tune previous models or to build an independent architecture based entirely on oral examinations. Furthermore, by comparing the results achieved by the inferential models with those assigned by the human experts, we identified linguistic and speech-related features that corresponded to the resulting classification. As a matter of fact, they could be used as additional information to enhance proficiency assessment in speech-based exams.

8.2. Limitations

In the course of this project, we became aware of some limitations due to the amount of data available, their imbalance and also the lack of evaluations objectively comparable with CEFR levels.

For example, in the case of the Italian and German language datasets, the limited number of texts, less or around one thousand, required the necessary use of cross-validation techniques to train and evaluate the neural model. In

spite of this, however, the achieved results were much lower than those obtained with the larger English language dataset, EFCAMDAT. In addition, the presence of widely varying quantities of examinations between one proficiency level and another also brought the need of rebalancing the weights for the training and test processes. Yet, the obtained outcomes were still affected by such variance.

Finally, having to remap the scores assigned by human evaluators to the CEFR levels arbitrarily, we noticed rather poor results in the automatic classification of such examinations, both with the CLC-FCE dataset and with that of the Free University of Bozen-Bolzano. The reasons for this could be traced back to a not clearly quantifiable evaluation measure to be passed on to an automated model. In addition, regarding the oral examinations of the University of Bozen-Bolzano, since their availability was limited and their distribution imbalanced, we were not able neither to fine-tune or independently train an architecture to classify nor rate the students' monologues.

The above-mentioned problems constituted limitations in this project which could be overcome in the event of future data becoming available in larger numbers and in a more balanced manner, as well as with shared methodical assessments concerning the levels of competence illustrated in the Common European Framework. If this were the case, better and more comparable results would be obtained between the three languages. For the oral exams section, on the other hand, ad hoc tailored methods could be devised.

8.3. Future research

The research work carried out so far clearly constitutes the preliminary stages of a project that could be expanded and enriched with more data and further analyses. The next possible steps could be as summarised below:

- adding further principled parameters for better written exams evaluations,
- training architectures for the Italian and German language using a more significant number of exams, annotated and evaluated according to similar principles, possibly more equally distributed across the different levels of proficiency indicated in the CEFR scale;
- collecting more spoken data related to the English language and/or the Italian and German ones in order to be able to train a model uniquely on them;
- identifying speech features to support written-trained models in oral exams classification and scoring;
- attempting the creation of an end-to-end system that based directly on the students' audios is able to classify them.
- developing a multilingual automatic assessment system for language competences that allows to directly evaluate the proficiency of multilingual learners in a cross-linguistic way, possibly exploiting transfer learning.

Moreover, in the future, this research work could be continued by comparing oral and written examination data from the same and different learners, in order to explore not only automatic assessment systems for each modality and language, but also to address the open questions of whether foreign language learning is different or homogeneous across written and spoken modalities within the same subjects and groups of learners.

8.4. Conclusion

In conclusion, we can state that despite the preliminary nature of this study, the results obtained are quite favourable for the written exams and promising for the oral exams. Undoubtedly, there is a large scope for improvement in both parts, especially for what concerns oral examinations. Furthermore, the experiments carried out on Italian and German constitute a step forward compared to the limited information available on the evaluation and automatic correction of texts in foreign languages other than English. We believe in the importance of research in this area and the continuation of the initiated multidisciplinary project.

References

Baese-Berk, M. & Morrill, T. H. (2015). Speaking rate consistency in native and non-native speakers of English, The Journal of the Acoustical Society of America, 138 (3).

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D. & Weintraub, M. (1990). Automatic evaluation and training in English pronunciation. In: Proceedings of the ICSLP-90: 1990 International Conference on Spoken Language Processing, Kobe, Japan.

Bialystok, E. & Sharwood Smith, M. (1985). Interlanguage is not a state of mind: An evaluation of the construct for second-language acquisition, Applied linguistics, 6 (2), pp. 101-117.

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A. & Vettori, C. (2014, May). The MERLIN corpus: Learner language and the CEFR. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation LREC, Iceland: European Language Resources Association, pp. 1281-1288.

Burstein, E. (2003). The E-rater scoring engine: Automated essay scoring with natural language processing. In: Shermis, M.D. and Burstein, J. (eds.), Automated essay scoring: a cross-disciplinary perspective, Erlbaum: Mahwah, pp. 113-122.

Canale, M. & Swain, M. (1980), Theoretical Bases for Communicative Approaches to Second Language Teaching and Testing, Applied Linguistics, vol. 1, n. 1, pp. 1-47.

Chapelle, C. A. & Voss, E. (2008). Utilizing technology in language assessment. In: Hornberger, N.H. (ed.), Encyclopedia of Language and Education, Boston: Springer, pp. 123-134.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge: M.I.T.

Cohen, A. D. (2016). Teaching and learning second language pragmatics. In: Hinkel, E. (ed.), Handbook of Research in Second Language Teaching and Learning: Volume III, 428-451.

Corder, S. P. (1967). The Significance of Learner's Errors, International Review of Applied Linguistics in Language Teaching, vol. V, n. 4, pp. 161-170.

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment, Cambridge, Cambridge University Press.

Cucchiarini, C., Strik, H. & Boves, L. (2002). Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech, Journal of the Acoustical Society of America, vol. 6, n. 111, pp. 2862-2873.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota.

Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review, Procedia Computer Science, 128, pp. 32-37.

EUROSTAT. (2020, September). Foreign language learning statistics. Eurostat Statistics Explained, [Accessed on the 1st of May 2021]. https://ec.europa.eu/eurostat/statistics-explained/index.php/Foreign_language_learning_statistics.

Geertzen, J., Alexopoulou, A. & Korhonen, A. (2014). Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In: Miller, R. T. (ed.), Selected Proceedings of the 2012 Second Language Research Forum, Somerville, MA: Cascadilla Proceedings Project, pp. 240-254.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory, Neural computation, 9(8), pp. 1735-1780.

Huang, Y., Geertzen, J., Baker, R., Korhonen, A., Alexopoulou, T., & First, E. E. (2017). The EF Cambridge Open Language Database (EFCAMDAT): Information for Users.

Hymes, D. (1972). On Communicative Competence. In: Pride J. and Holmes, J. (eds.) Sociolinguistics: Selected Readings, Harmondsworth: Penguin, pp. 269-293.

James, C. (2005). Contrastive analysis and the language learner. In: Allerton, D., Tschichold, C. and Wieser, J. (eds.), Linguistics, Language Teaching and Language Learning, Basel: Schwabe, pp. 1-20.

Jang, E. E. (2017). Cognitive aspects of language assessment, Language Testing and Assessment, pp.163-177.

Juffs, A., Han, N. R. & Naismith, B. (2020). The University of Pittsburgh English Language Corpus (PELIC) [Data set]. http://doi.org/10.5281/zenodo.3991977.

King, G., & Zeng, L. (2001). Logistic regression in rare events data, Political analysis, 9(2), pp.137-163.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization, CoRR, arXiv preprint arXiv:1412.6980.

Lado, R. (1961). Language Testing: The construction and use of foreign language tests, London: Longman.

Landauer, T. K. (2003). Automatic essay assessment, Assessment in education: Principles, policy and practice, vol. 10, n. 3, pp. 295-308.

McCarthy, P. M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment, Behavior research methods, 42(2); pp. 381-392.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55-60.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space, ICLR, https://arxiv.org/abs/1301.3781.

Miłkowski, M. (2010). Developing an open-source, rule-based proofreading tool. Software: Practice and Experience, 40(7), pp. 543-566.

Morris, A. C., Maier, V. & Green, P. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In Eighth International Conference on Spoken Language Processing.

Ong, T. (2017). Facebook's translations are now powered completely by AI.

Page, E. B. (1968). The use of the computer in analyzing student essays», International Review of Education, vol. 14, n. 2, pp. 210-225.

Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) , pp. 5206-5210.

Pellegrino, J. W., Chudowsky, N. & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment. Washington DC: National Academies Press.

Pennington, J., Socher, R. & Manning, C. D. (2014). GloVe: Global vectors for word representation. In: Association for Computational Linguistics, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1532-1543.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N. & Vesely, K. (2011). The Kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society.

Proisl, T. & Schöch, C. (2020). Textcomplexity. Linguistic and stylistic complexity. [Accessed on the 26th May 2021]. https://github.com/tsproisl/textcomplexity.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.

Ravanelli, M., Brakel, P., Omologo, M. & Bengio, Y. (2018). Light gated recurrent units for speech recognition. IEEE Transactions on Emerging Topics in Computational Intelligence, 2(2), pp. 92-102.

Richards, J. C. (2015). Error analysis: Perspectives on second language acquisition. Routledge.

Saito, K. (2017). Effects of sound, vocabulary, and grammar learning aptitude on adult second language speech attainment in foreign language classrooms. Language Learning, 67(3), pp. 665-693.

Sak, H., Senior, A., Rao, K., Beaufays, F. & Schalkwyk, J. (2015). Google voice search: faster and more accurate. Google Research blog.

Selinker, L. (1972). Interlanguage. International Review of Applied Linguistics, 10, pp. 209–31.

Sheen, R. (1996). The advantage of exploiting contrastive analysis in teaching and learning a foreign language. IRAL: International Review of Applied Linguistics in Language Teaching, 34(3), 183.

Smith, C. (2017). iOS 10: Siri now works in third-party apps, comes with extra AI features. BGR. [Accessed on the 15th May 2021]. https://bgr.com/tech/ios-10-siri-third-party-apps-4914313/

Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Żyła-Hoppe, M., Banaszczak, J., Augustyniak, L., Mizgajski, J. & Carmiel, Y. (2020). WER we are and WER we think we are. arXiv preprint arXiv:2010.03432.

Taguchi, N. (2017). Interlanguage pragmatics: A historical sketch and future directions. In: Barron, A., Gu, Y. & Steen, G. (eds.), The Routledge handbook of pragmatics, London: Routledge, pp. 153-167.

Taylor, A., Marcus, M. & Santorini, B. (2003). The Penn treebank: an overview. In: Abeillè, A. (eds.). Treebanks. Text, Speech and Language Technology, vol 20, Dordrecht: Springer, pp. 5-22.

Townshend, B., Bernstein, J., Todic, O. & Warren, E. (1998). Estimation of Spoken Language Proficiency. In: Proceedings of the ESCA Workshop STiLL: 'Speech Technology in Language Learning', Sweden: Marholmen, pp. 179-182.

Ullman, M. T. (2005). A cognitive neuroscience perspective on second language acquisition: The declarative/procedural model. In: Sanz, C. (ed.), Mind and context in adult second language acquisition, Washington: Georgetown University Press, pp. 141-78.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., N Gomez, A., Kaiser, U. & Polosukhin, I. (2017). Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, R. and Garnett, R. (eds.), Advances in Neural Information Processing Systems 30, pp. 5998–6008.

Wang, Y., Luan, H., Yuan, J., Wang, B. & Lin, H. (2020). LAIX Corpus of Chinese Learner English: Towards a Benchmark for L2 English ASR. In: Proceedings Interspeech 2020, pp. 414-418.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S. , von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38-45.

Xi, X., Higgins, D., Zechner, K. & Williamson, D. M. (2008). Automated Scoring of Spontaneous Speech Using SpeechRaterSM v1.0 - Educational Testing Service Research Report No. RR-08-62, Princeton: ETS,

Yannakoudakis, H., Briscoe, T. & Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, June 19-24, 2011. Association for Computational Linguistics, pp. 180-189.

Appendix A

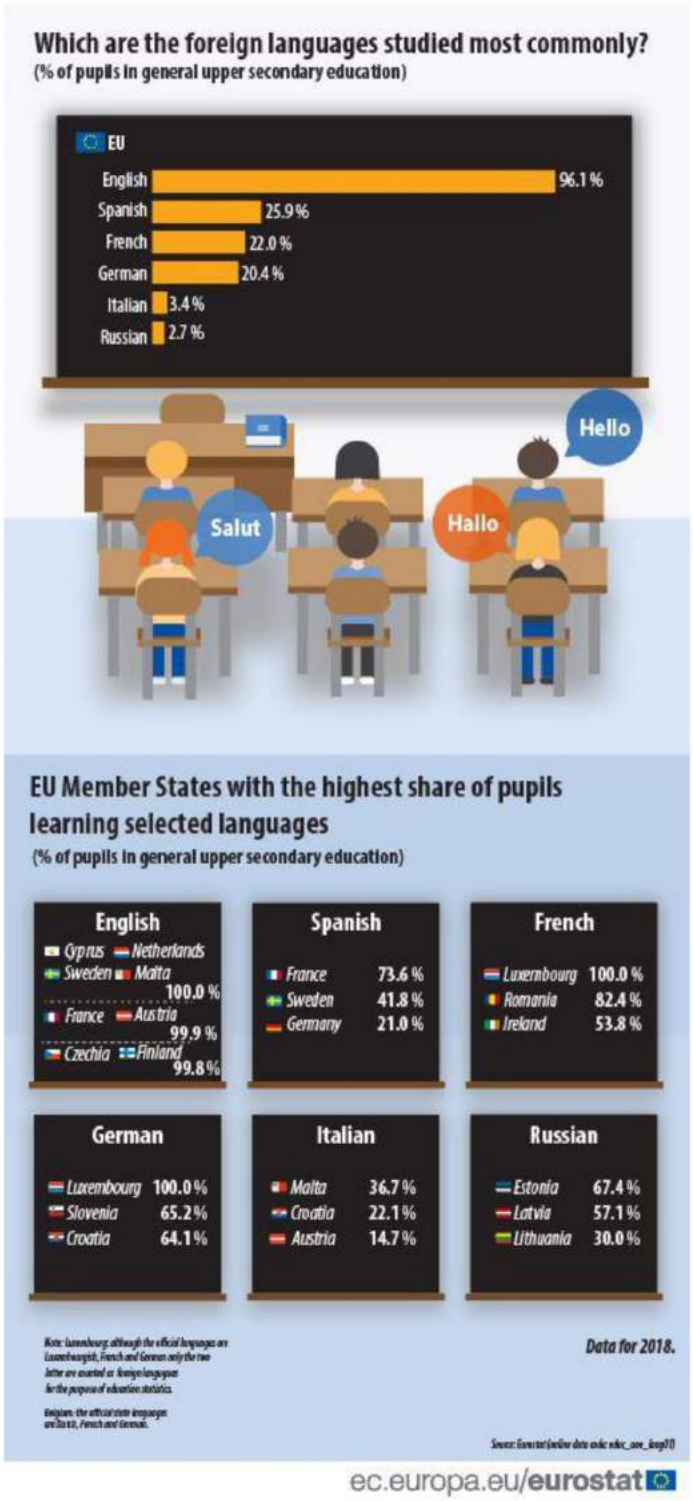


Fig. 1 Secondary education pupils learning English, Spanish, French, German, Italian and Russian in EU member states in 2018.

From EUROSTAT (2020). "Foreign language learning statistics". (https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Foreign_language_learning_statistics).

Proficient User	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
Independent User	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Fig. 2 Original global scale released by the Council of Europe in 2001.

From Council of Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment, Cambridge, Cambridge University Press.

Englishtown	1-3	4-6	7-9	10-12	13-15	16
Cambridge Esol	-	KET	PET	FCE	CAE	-
IELTS	-	<3	4-5	5-6	6-7	>7
TOEFL iBT	-	-	57-86	87-109	110-120	-
TOEIC Listening & Reading	120-220	225-545	550-780	785-940	945	-
TOEIC Speaking & Writing	40-70	80-110	120-140	150-190	200	-
CEFR	A1	A2	B1	B2	C1	C2

Fig. 3 EFCAMDAT 16 levels of competence mapped to common proficiency standards.

From Huan,Y., Geertzen, J., Baker, R., Korhonen, A. & Alexopoulou, T. (2017). The EF Cambridge Open Language Database (EFCAMDAT), information for users, pp. 3.

Appendix B

	Ampiezza del lessico	Padronanza del lessico	Grammatica	Kohärenz	Appropriatezza sociolinguistica	Ortografia
C2	SPETTRO repertorio vastissimo PECULIARITÀ Padronanza di espressioni colloquiali/idiomatiche, da cui consapevolezza dei livelli di connotazione semantica	IN GENERALE Costantemente corretto & adeguato	REPERTORIO E CORRETTEZZA anche in forme linguistiche complesse → costante controllo grammaticale	TESTO • coeso • coerente CONNETTIVI uso appropriato di una (ampia) gamma di connettivi per strutturare e rendere coeso il testo	IN GENERALE Coglie pienamente le implicazioni sociolinguistiche/socioculturali del linguaggio di un parlante nativo e reagisce in modo adeguato ESPRESSIONI IDIOMATICHE Buona padronanza di espressioni idiomatiche/colloquiali e consapevolezza dei livelli di connotazione semantica	Nessun errore ortografico
C1	SPETTRO vasto repertorio lessicale PECULIARITÀ • Buona padronanza di espressioni colloquiali/idiomatiche; • Lacune prontamente superate tramite circonlocuzioni; • Strategie di evitamento poco evidenti	ERRORI Occasionali sbagli di minore entità, ma nessun errore lessicale significativo	IN GENERALE Costantemente elevato livello di correttezza CORRETTEZZA Gli errori sono… • rari • poco evidenti	TESTO • ben strutturato, chiaro CONNETTIVI Padroneggia • schemi organizzativi • connettivi • espressioni coesive	ESPRESSIONI IDIOMATICHE Riconosce un'ampia gamma di espressioni idiomatiche/colloquiali DIMENSIONE DIAFASICA (situativa) Coglie i cambiamenti di registro ATTI LINGUISTICI • è in grado di usare la lingua per scopi sociali in modo flessibile ed efficace • è in grado di esprimere emozioni, fare allusioni e battute umoristiche	• Impaginazione, strutturazione in paragrafi e punteggiatura coerenti e funzionali • Ortografia corretta, a parte qualche sbaglio occasionale
B2	SPETTRO buon repertorio lessicale relativo al proprio settore & ad argomenti generali PECULIARITÀ • Le formulazioni vengono variate per evitare ripetizioni • Cionostante, lacune lessicali possono richiedere circonlocuzioni	IN GENERALE Correttezza lessicale generalmente elevata ERRORI • Qualche confusione/scelta lessicale scorretta • Nessun ostacolo alla comunicazione	IN GENERALE buona padronanza grammaticale CORRETTEZZA Gli errori non provocano fraintendimenti	TESTO • coerente, chiaro • in un intervento lungo possono presentarsi „salti" logici CONNETTIVI numero limitato	DIMENSIONE SITUATIVA Si esprime in modo adeguato alla situazione IN GENERALE • nessun errore di formulazione grossolano. • Non si rende ridicolo	• È in grado di stendere un testo coerente e chiaro • Rispetta gli standard convenzionali di impaginazione e strutturazione in paragrafi • Ortografia e punteggiatura sono ragionevolmente corrette, ma possono presentare tracce dell'influenza della L1
B1	SPETTRO lessico sufficiente per esprimersi su quasi tutti gli argomenti della vita di tutti i giorni (ad es. famiglia, hobby, interessi, lavoro, viaggi, attualità) PECULIARITÀ Circonlocuzioni	IN GENERALE/SPETTRO Buona padronanza del lessico elementare ERRORI Errori gravi in caso di pensieri più complessi/argomenti & situazioni non familiari	REPERTORIO & CORRETTEZZA formule di routine e strutture di uso frequente relative alle situazioni più prevedibili → ragionevolmente corretto	TESTO • sequenza lineare per punti • serie di elementi relativamente brevi e semplici	ATTI LINGUISTICI Ampia gamma di atti linguistici utilizzando le espressioni più comuni DIMENSIONE SITUATIVA Registro „neutro" REGOLE DI CORTESIA È consapevole delle più importanti regole di cortesia e delle differenze fra usi e costumi più significative e le rispetta	• è in grado di stendere un testo nel complesso comprensibile • Ortografia, punteggiatura e impaginazione sono corrette quanto basta per essere quasi sempre comprensibili.
A2	SPETTRO/FUNZIONI Lessico sufficiente a... • esprimere bisogni comunicativi di base • far fronte a bisogni semplici di „sopravvivenza"	IN GENERALE/SPETTRO Repertorio ristretto funzionale a esprimere bisogni concreti della vita quotidiana.	REPERTORIO & CORRETTEZZA alcune strutture semplici → errori gravi/sistematici (ad es. confonde i tempi verbali, dimentica di segnalare gli accordi) CHIAREZZA/COMPRENSIBILITÀ Solitamente chiaro ciò che cerca di dire	TESTO frasi semplici/gruppi di parole CONNETTIVI semplici quali "e", "ma" e "perchè"	ATTI LINGUISTICI Atti linguistici di base, ad es. salutare, rivolgere la parola a qualcuno, invitare, chiedere scusa in scambi comunicativi molto brevi REGOLE DI CORTESIA Formule convenzionali correnti per salutare e rivolgere la parola a qualcuno	COPIARE Brevi frasi su argomenti correnti (ad es. indicazioni per arrivare in un posto) SCRIVERE • riproduzione della fonetica di parole brevi che fanno parte del suo vocabolario orale, ma con ortografia non necessariamente del tutto corretta
A1	SPETTRO • Repertorio lessicale di base fatto di singole parole ed espressioni • Riferimento a situazioni concrete	- al di sotto del livello A2 -	REPERTORIO & CORRETTEZZA qualche semplice struttura grammaticale e semplice modello sintattico imparati a memoria → padronanza limitata	TESTO parole/gruppi di parole CONNETTIVI lineari, molto elementari quali "e" o "allora" per collegare parole o gruppi di parole	IN GENERALE Contatti sociali di base (ad es. dire „per favore/grazie/scusi", per presentarsi, ecc.) REGOLE DI CORTESIA Più semplici formule convenzionali correnti per salutare e congedarsi	COPIARE parole e brevi espressioni conosciute (ad es. avvisi, istruzioni…) SCHREIBEN Sillabare il proprio indirizzo, nazionalità e altri dati personali

Fig. 1 Original MERLIN evaluation grid for Italian written exams.

From MERLIN project, Griglia di valutazione (2014), http://MERLIN-platform.eu.

	Wortschatz: Spektrum	Wortschatz: Beherrschung	Grammatik	Kohärenz	Soziolinguistische Angemessenheit	Orthographie
C2	SPEKTRUM Sehr reicher Wortschatz BESONDERHEITEN Beherrschung umgsprachl./idiomatischer Wendungen, dabei Bewusstsein über Konnotationen	ALLGEMEIN Durchgängig korrekt & angemessen	REPERTOIRE & KORREKTHEIT auch bei komplexen Sprachmitteln → durchgehende Beherrschung der Grammatik	TEXT • Gut gegliedert • Zusammenhängend KONNEKTOREN setzt Vielfalt an K. zur Gliederung & Verknüpfung angemessen ein	ALLGEMEIN soziolinguistische/soziokulturelle Implikationen von L1-Sprechern richtig eingeschätzt & entsprechend reagiert IDIOMATIK Gute Kenntnisse idiom./alltagssprachl. Wendungen mit Konnotationen	keine orthographischen Fehler
C1	SPEKTRUM Großer Wortschatz BESONDERHEITEN • Gute Beherrschung umgsprachl./ idiomatischer. Wendungen; • Bei Lücken problemlos Umschreibungen; • Vermeidungsstrategien selten	FEHLER Gelegentliche kleinere Schnitzer, aber keine größeren Fehler	ALLGEMEIN Beständig hohes Maß an Korrektheit KORREKTHEIT Fehler sind… • Selten • Kaum auffallend/störend	TEXT • gut strukturiert, klar KONNEKTOREN beherrscht K. • zur Gliederung • zur inhaltlichen und • zur sprachlichen Verknüpfung	IDIOMATIK Großes Spektrum idiom./alltagssprachl. Redewendungen erkannt SITUATIVE DIMENSION Wechsel im Register richtig eingeschätzt SPRACHFUNKTIONEN • Kann Sprache zu geselligen Zwecken flexibel/effektiv einsetzen • Kann Emotionen ausdrücken, Anspielungen und Scherze machen	• Gestaltung, Gliederung, Zeichensetzung konsistent & hilfreich • Rechtschreibung richtig, gelegentliches Verschreiben
B2	SPEKTRUM Großer Wortschatz im eigenen Sachgebiet & bei allgemeinen Themen BESONDERHEITEN • Formulierungen werden variiert, um Wiederholungen zu vermeiden. • Trotzdem können Lücken zu Umschreibungen führen	ALLGEMEIN Im Allgemeinen große Korrektheit FEHLER • Einige Verwechslungen /falsche Wortwahl • Durch Fehler keine Behinderung der Kommunikation	ALLGEMEIN Gute Beherrschung der Grammatik KORREKTHEIT Durch Fehler entstehen keine Missverständnisse	TEXT • zusammenhängend, klar • evtl. sprunghaft bei längeren Beiträgen KONNEKTOREN Begrenzte Anzahl	SITUATIVE DIMENSION Situationsangemessener Ausdruck ALLGEMEIN • Keine krassen Formulierungsfehler. • Belustigt nicht	• Schreibt zusammenhängend, klar verständlich. • Übliche Konventionen der Gestaltung & Gliederung (Absätze) eingehalten • Rechtschreibung/Zeichensetzung hinreichend korrekt, dabei evtl. L1-Einflüsse
B1	SPEKTRUM Ausreichend großer Wortschatz bei den meisten Themen des Alltagslebens (bspw. Familie, Hobbys, Interessen, Arbeit, Reisen, aktuelle Ereignisse) BESONDERHEITEN Dabei treten Umschreibungen auf	ALLGEMEIN/SPEKTRUM Gute Beherrschung des Grundwortschatzes FEHLER Elementare Fehler bei komplexeren Sachverhalten/wenig vertrauten Themen & Situationen	REPERTOIRE & KORREKTHEIT häufig verwendete Redefloskeln & Wendungen, die an eher vorhersehbare Situationen gebunden sind → ausreichend korrekt	TEXT • zusammenhängend, linear • kurze, einfache Einzelelemente	SPRACHFUNKTIONEN Breites Spektrum von Sprachfunktionen mit gebräuchlichste Redemitteln SITUATIVE DIMENSION Neutrales Register HÖFLICHKEITSKONVENTIONEN kennt wichtigste Höflichkeitskonventionen & Unterschiede zw. Sitten und Gebräuchen und handelt entsprechend	• Schreibt zusammenhängend. • Rechtschreibung, Zeichensetzung und Gestaltung exakt genug, dass man sie meistens verstehen kann.
A2	SPEKTRUM/FUNKTIONEN Wortschatz ausreichend für … • elementare Kommunikationsbedürfnisse • einfache Grundbedürfnisse	ALLGEMEIN/SPEKTRUM Beherrscht einen begrenzten Wortschatz in Zusammenhang mit konkreten Alltagsbedürfnissen.	REPERTOIRE & KORREKTHEIT Einige einfache Strukturen → elementare/ systematische Fehler (z.B. Zeitformen vermischt, keine Subjekt-Verb-Kongruenz markiert) VERSTÄNDLICHKEIT In der Regel klar, was ausgedrückt werden soll	TEXT Einfache Sätze/Wortgruppen KONNEKTOREN Einfache K. wie 'und', 'aber' und 'weil'	SPRACHFUNKTIONEN elementare Sprachfunktionen, z.B. Begrüßung, Anrede, Einladung, Entschuldigung, in sehr kurzen Kontaktgesprächen HÖFLICHKEITSKONVENTIONEN gebräuchliche Höflichkeitsformeln der Begrüßung und Anrede	ABSCHREIBEN kurze Sätze über alltägliche Themen (z.B. Wegbeschreibungen) SCHREIBEN • kurze Wörter aus mündlichem Wortschatz, dabei ("phonetische") Wiedergabe nicht unbedingt orthographisch korrekt
A1	SPEKTRUM • Elementarer Vorrat an einzelnen Wörtern und Wendungen • Dabei Bezug auf bestimmte konkrete Situationen	- unterhalb von A2 -	REPERTOIRE & KORREKTHEIT einige wenige einfache grammatischer Strukturen und Satzmuster, auswendig gelernt → begrenzt beherrscht	TEXT Wörter/Wortgruppen KONNEKTOREN Sehr einfache lineare K. wie 'und' oder 'dann' bei Wörtern/ Wortgruppen	ALLGEMEIN Elementarer sozialer Kontakt (z.B. „bitte"/"danke" sagen, sich vorstellen entschuldigen) wird hergestellt HÖFLICHKEITSKONVENTIONEN einfachste alltägliche Höflichkeitsformeln zur Begrüßung / Verabschiedung	ABSCHREIBEN vertraute Wörter, kurze Redewendungen (z.B. Schilder, Anweisungen…) SCHREIBEN Buchstabieren: Adresse, Nationalität u.a. Angaben zur Person

Fig. 2 Original MERLIN evaluation grid for German written exams.

From MERLIN project, Griglia di valutazione tedesco (2014), http://MERLIN-platform.eu.

GRID FOR THE EVALUATION OF ORAL PRODUCTION B2

unibz Freie Universität Bozen · Libera Università di Bolzano · Università Liedia de Bulsan

POINTS	COMMUNICATIVE EFFECTIVENESS [CE]	POINTS	LEXIS [L]	POINTS	GRAMMAR[GR]	POINTS	FLUENCY, PRONUNCIATION [FPI]
5	• Completes all aspects of the task satisfactorily. • Expresses himself/herself clearly and in a structured way; uses cohesive devices effectively. • Arguments are made precisely and are supported with details and examples.	5	• Has a broad lexical range and uses a considerable amount of subject-specific vocabulary. • Uses fixed phrases and collocations correctly. • Compensates for any gaps by paraphrasing appropriately. • Occasional interference from other languages.	5	• Uses a wide range of structures at this level. • Good control of grammar: occasional slips are self-corrected.	5	• Clear pronunciation. • Occasional phonological slips that are self-corrected. • Uses intonation to emphasize important points. • Expresses himself/herself effortlessly and spontaneously, so that the speech can be followed without effort.
4	• Completes all aspects of the tasks adequately, even if some points are more developed than others. • Mostly expresses himself/herself clearly and in a structured way, mostly uses cohesive devices effectively. • Arguments are made quite precisely and are supported with details.	4	• Has an adequate range of vocabulary; occasionally uses subject-specific vocabulary. • Almost always uses fixed phrases and collocations correctly. • Tries to compensate for lexical gaps but not always successfully. • Occasional interference from other languages.	4	• Generally uses frequent structures correctly. • Uses grammar satisfactorily; errors mostly occur in complex utterances.	4	• Clear pronunciation. • Occasional phonological errors, which are not always self-corrected. • Uses intonation quite effectively to emphasize important points. • Generally expresses himself/herself easily and spontaneously; occasionally hesitates in more complex utterances.
3	• Partially or approximatively completes the task required. • Satisfactorily expresses himself/herself clearly and in a structured way, albeit with some minor uncertainties. • Arguments are not made completely clearly; there may be some unimportant details.	3	• Has an adequate range of vocabulary. • Uses paraphrasing, can sometimes be vague. • Occasional interference from other languages.	3	• Has a sufficient range of grammar and tends to use simple forms. • Adequate use of grammar. Errors occur frequently, but do not lead to misunderstanding.	3	• Clear pronunciation which is sometimes unnatural • Obvious phonological errors which do not effect understanding • Uses intonation effectively to emphasize important points, but not always successfully. • Generally expresses himself/herself effortlessly; an effort needs to be made to understand the speech at times.
2	• Completes the tasks only to a limited extent. • The structure of the speech is not clearly apparent and confusing in parts. • Reasoning is poor.	2	• Interference from other languages is frequent. • Excessively uses generic and vague terminology.	2	• Only uses lower level structures. • Control of grammar is inadequate; errors occasionally prevent understanding.	2	• Phonological errors and difficulties in articulation can impair understanding. • Intonation is not used to emphasize important points. • Gaps and pauses impede fluency of speech.
1	• The assigned tasks are not completed. • Expresses himself/herself too simplistically and mostly cannot be understood • Ideas are not supported with arguments.	1	• Almost always uses only basic vocabulary. • Has great difficulty finding adequate terms to fulfil the task. • Frequent lexical errors and interference from other languages.	1	• Also has difficulties with lower level structures. • The amount and type of errors prevent understanding of the speech.	1	• Phonological errors and difficulties in articulation are very frequent. • It is necessary to interpret what is being said. • Hesitations are continuous throughoutand the discourse is fragmented.
0	• The speech is incomprehensible or cannot be evaluated.	0	• The speech is incomprehensible or cannot be evaluated.	0	• The speech is incomprehensible or cannot be evaluated.	0	• The speech is incomprehensible or cannot be evaluated.

Fig. 3 Grid for the evaluation of oral productions B2 level – Free University of Bozen-Bolzano