University of Trento

Department of Psychology and Cognitive Science

Doctoral school in Cognitive Science
XXXV Cycle

~ · ~

Academic Year 2022–2023

Automatic Assessment of L2 Spoken English

Supervisors
Massimo Zancanaro
Marco Matassoni

PhD Candidate
Stefano Bannò

A mamma, papà e Simone

Acknowledgments

First, I would like to thank my FBK supervisor, Marco Matassoni. I thank him for sharpening my way of thinking and shaping me into a better scientist and person. This work would not have been possible without his guidance, vision and mentorship. Meeting two other amazing scientists gave a turning point to my research project. Mark Gales and Kate Knill taught me that you can be extremely creative even when crunching numbers. I thank all three of them for their wisdom, understanding and infinite patience. I would also like to thank my supervisor from the University of Trento, Massimo Zancanaro, for his constant support over the past three and a half years. I am also very grateful to Catia Cucchiarini and Rutuja Ubale for reviewing my thesis and giving me their precious and insightful advice to improve it.

Most of this work was carried out at the SpeechTeK Lab of FBK, where interacting with Daniele Falavigna, Roberto Gretter, Alessio Brutti, Maurizio Omologo, Mohamed Nabih, Umberto Cappellazzo, Michela Rais and Seraphina Fong has been fundamental for my research. I want to express my deepest gratitude to them. I also must acknowledge IPRASE and ISIT for their work in the Trentino Language Testing campaigns. Special thanks go to the Department of Psychology and Cognitive Science of the University of Trento and the Digital Society research centre of FBK.

The central part — in many ways — of my PhD took place at the ALTA Institute of the University of Cambridge. In addition to Mark Gales and Kate Knill, I would like to thank Yiting Lu, Vyas and Vatsal Raina, Mengjie Qian, Adian Liusie, Yassir Fathullah, Linlin Wang and Kostas Kyriakopoulos for their invaluable help. I also thank Cambridge Assessment, who provided the funding and data for the experiments conducted while I was working at the ALTA Institute.

Finally, I would like to express my love and gratitude to mamma Rosa, papà Paolo and Simone for believing in me even when I didn't.

Abstract

In an increasingly interconnected world where English has become the lingua franca of business, culture, entertainment, and academia, learners of English as a second language (L2) have been steadily growing. This has contributed to an increasing demand for automatic spoken language assessment systems for formal settings and practice situations in Computer-Assisted Language Learning. One common misunderstanding about automated assessment is the assumption that machines should replicate the human process of assessment. Instead, computers are programmed to identify, extract, and quantify features in learners' productions, which are subsequently combined and weighted in a multidimensional space to predict a proficiency level or grade. In this regard, transferring human assessment knowledge and skills into an automatic system is a challenging task since this operation should take into account the complexity and the specificities of the proficiency construct.

This PhD thesis presents research conducted on methods and techniques for the automatic assessment and feedback of L2 spoken English, mainly focusing on the application of deep learning approaches. In addition to overall proficiency grades, the main forms of feedback explored in this thesis are feedback on grammatical accuracy and assessment related to particular aspects of proficiency (e.g., grammar, pronunciation, rhythm, fluency, etc.).

The first study explores the use of written data and the impact of features extracted through grammatical error detection on proficiency assessment, while the second illustrates a pipeline which starts from disfluency detection and removal, passes through grammatical error correction, and ends with proficiency assessment. Grammar, as well as rhythm, pronunciation, and lexical and semantic aspects, is also considered in the third study, which investigates whether it is possible to use systems targeting specific facets of proficiency analytically when only holistic scores are available. Finally, in the last two studies, we investigate the use of self-supervised learning speech representations for both holistic and analytic proficiency assessment.

While aiming at enhancing the performance of state-of-the-art automatic systems, the present work pays particular attention to the validity and interpretability of assessment both holistically and analytically and intends to pave the way to a more profound and insightful knowledge and understanding of automatic systems for speaking assessment and feedback.

Contents

List of Figures **xii**

List of Tables **xv**

Introduction **1**

1 Theoretical framework **7**

1.1 From structuralism to the CEFR 7

1.2 Assessment of speaking 10

1.2.1 Linguistic competence 14

1.2.2 Sociolinguistic competence 24

1.2.3 Pragmatic competence 27

2 Automatic speaking assessment **33**

2.1 Introduction 33

2.1.1 Advantages, limitations, and challenges of automatic assessment 34

2.1.2 Historical background 37

2.2 Linguistic competence 39

2.2.1 Grammar 39

2.2.2 Pronunciation 42

2.2.3 Vocabulary 46

2.3 Sociolinguistic competence 49

2.4 Pragmatic competence 50

2.4.1 Coherence and cohesion 50

2.4.2 Fluency 52

3 Data **55**

3.1 Publicly available data 55

3.1.1 Written corpora 55

3.1.2 Written and spoken corpora 58

3.1.3 Spoken corpora 59

3.2 Non-publicly available data 59

3.2.1 Written corpora 59

3.2.2 Written and spoken corpora 60

3.2.3 Spoken corpora 61

3.3 Other spoken corpora 62

4 GED, GEC, and assessment **65**

4.1 Study 1: Cross-corpora experiments of speaking assessment and grammatical error detection 66

4.1.1 Introduction 66

4.1.2 Data 67

4.1.3 Model architectures 71

4.1.4 Experiments and results 73

4.1.5 Conclusions 76

4.2 Study 2: Using grammatical error correction for speaking assessment 78

4.2.1 Introduction 78

4.2.2 Data 79

4.2.3 Disfluency detection 82

4.2.4 GEC 83

4.2.5 Proficiency assessment 85

4.2.6 Conclusions 92

5 View-specific assessment **95**

5.1 Study 3: View-specific assessment 96

5.1.1 Introduction 96

5.1.2 View-specific training 98

5.1.3 Single-view graders 99

5.1.4 Data and experimental setup 102

5.1.5 Experimental results and analysis 105

5.1.6 Conclusions 108

6 SSL-based assessment **111**

6.1 Study 4: Speaking assessment using wav2vec 2.0 (Part 1) 111

6.1.1 Introduction 111

6.1.2 Data 113

6.1.3 Model architectures 115

6.1.4 Experiments and results 118

6.1.5 Conclusions 122

6.2 Study 5: Speaking assessment using wav2vec 2.0 (Part 2) 124

6.2.1 Introduction 124

6.2.2 Data 124

6.2.3 Model architectures 125

6.2.4 Experimental results and analysis 127

6.2.5 Conclusions 131

Discussion and conclusions **133**

References **141**

Appendix A

TLT-school and TLT-GEC question prompts **183**

Appendix B

Linguaskill question prompts **185**

List of Figures

1.1 Diagram of communicative competence as outlined in this thesis. 14

2.1 Pipeline of a typical automatic system for speaking assessment. 34
2.2 Delivery and scoring possibilities in L2 speaking assessment. 35

4.1 Diagram of the proposed training pipeline based on textual input (i.e, the written train set). The grader is then used to predict proficiency scores on manual and ASR transcriptions (i.e., the spoken test set). 67
4.2 EFEX model architecture. 72
4.3 Grader architecture. 73
4.4 MSE variation across scores on manual transcriptions and ASR output text. . . . 75
4.5 The pipeline proposed in this study. 78
4.6 Precision and Recall curves: the model used in this study versus the model used in Lu et al. (2022). 83
4.7 Detailed diagram of the proposed pipeline. 87
4.8 MSE variation of the two graders across formal correctness scores (manual transcriptions). 89
4.9 Bar charts showing the 10 most common ERRANT edit labels on manual transcriptions (above) and ASR transcriptions (below). 91

5.1 View-specific training. 98
5.2 Example of text, GEC edit, and POS sequence. 101
5.3 5 most common ERRANT edit labels across three systems: manual transcriptions manually corrected, manual transcriptions automatically corrected, and ASR transcriptions automatically corrected. 104

5.4 RMSE variation across proficiency levels. 107
5.5 Heatmap of the results of the post-hoc Nemenyi test. 108

6.1 Heatmap of the results of the post-hoc Nemenyi test on the analytic subscores of the TLT-school dataset. 116
6.2 Confusion matrices of CEFR proficiency levels for the two grading systems (predicted labels on X-axis, true labels on Y-axis) on the ICNALE test set. 119
6.4 MSE variation of the wav2vec2-based and BERT-based (manual transcription) graders across scores for lexical richness and complexity. 122
6.5 The three systems considered in this study: a) standard grader, b) BERT-based grader, and c) wav2vec2-based grader. 125

List of Tables

1.1 CEFR levels description. 9

2.1 Example of written GEC. 41
2.2 Example of spoken GEC. 41

3.1 EFCAMDAT error tagset. 56
3.2 TLT-school proficiency indicators for speaking and writing. 61

4.1 EFCAMDAT error tagset without codes related to spelling, punctuation and orthographic errors. 68
4.2 Symmetric KL Divergence between distributions of counts from all 17 error labels in EFCAMDAT. 68
4.3 The 5 error classes used in the study. 69
4.4 Mean of the ratio (number of errors divided by number of words) of each error class and their sum for each proficiency level. 69
4.5 Statistics (number of answers and word counts) for the three test sets: ICNALE (Written and Spoken), CLC-FCE, TLT-school (Written and Spoken). 71
4.6 Statistics (number of answers and word counts) for the TLT-school spoken test set across test scores. 71
4.7 Model architectures and hyperparameters. The number of epochs in brackets refer to the EFEX-enriched model. 73
4.8 EFEX performance in terms of PCC on EFCAMDAT. 74
4.9 Results on the ICNALE test dataset (MSE and PCC). 74
4.10 Results on the TLT test dataset (MSE and PCC): baseline; baseline + fine-tuning; baseline + EFEX labels; baseline + EFEX labels + fine-tuning. 75

4.11 Frobenius norm values of EFEX vectors across score ranges. 76
4.12 Corpora statistics. Note that the table reporting the statistics on the datasets used for proficiency assessment shows the number of responses (which may consist of more than one sentences) as opposed to the number of sentences reported in the tables above. 80
4.13 DD+spoken GEC. The disfluencies are indicated in bold. 82
4.14 Results of DD on the TLT-GEC test set and LIN-MAN in terms of Precision, Recall, and F_1 Score. 83
4.15 Results of GEC on CLC-FCE test set and TLT-GEC test set in terms of M^2 and GLEU (**dsf** = transcriptions with disfluencies; **flt** = transcriptions with disfluencies manually removed; **autoflt** = transcriptions with disfluencies automatically removed). 84
4.16 Symmetric KL Divergence between distributions of counts from all 38 ERRANT edit labels in the TLT-school data across formal correctness scores. 85
4.17 Results on the TLT-school test set of the GEC-based (**GEC**), the BERT-based (**BERT**), and their combinations on the task of predicting the holistic score and the score related to formal correctness using manual transcriptions. 88
4.18 β coefficients of the GEC-based grader ($\beta_{\mathtt{g_c}}$) and the BERT-based grader ($\beta_{\mathtt{b_t}}$), and intercept (β_0) in the *shallow* combination for the holistic and formal correctness scores (manual transcriptions). 89
4.19 Results on the TLT-school test set of the GEC-based (**GEC**), the BERT-based (**BERT**), and their combinations on the task of predicting the holistic score and the score related to formal correctness using ASR transcriptions. 90
4.20 β coefficients of the GEC-based grader ($\beta_{\mathtt{g_c}}$) and the BERT-based grader ($\beta_{\mathtt{b_t}}$), and intercept (β_0) in the *shallow* combination for the holistic and formal correctness scores (ASR transcriptions). 92

5.1 Comparison of the performance of the GEC model on manual and ASR transcriptions in terms of M^2. 103
5.2 Symmetric KL Divergence between distributions of counts from all ERRANT edit labels in the manually annotated subset across proficiency levels. 104
5.3 Symmetric KL Divergence between distributions of counts from all ERRANT edit labels in the full training set (ASR transcriptions) across proficiency levels. . . . 105

5.4 Performance of the single-view graders and baseline in terms of RMSE. Individual models VS ensembles. 105
5.5 RMSE and β coefficients of linear regression model with different combinations. 106
5.6 Comparison of the performance of the baseline, text grader, and linear regression model. 108

6.1 Number of answers for each CEFR proficiency level in ICNALE. 114
6.2 Number of answers for each score range in TLT-school. 115
6.3 Hyperparameters of the individual wav2vec2-based graders. 117
6.4 Results on the ICNALE test set of the BERT-based and wav2vec2-based graders in terms of accuracy and weighted F_1 score. 119
6.5 Results on TLT-school test set (holistic score) of the BERT-based grader (manual and ASR transcriptions) and the wav2vec2-based grader in terms of PCC, SRC and MSE. 120
6.6 Results on TLT-school test set (lexical richness and complexity) of the BERT-based (manual transcriptions), the wav2vec2-based graders, and their combinations in terms of PCC, SRC and MSE. 122
6.7 RMSE results on the five parts of the LinGen exam. 127
6.8 Submission-level performance on LinGen and LinBus. 128
6.9 β coefficients of per-part linear regression model for the standard ($\mathtt{s}_\mathtt{d}^{\otimes}$), BERT ($\mathtt{b}_\mathtt{t}^{\otimes}$), wav2vec2 ($\mathtt{w}_\mathtt{v}^{\otimes}$), and combination ($\mathtt{s}_\mathtt{d}{\otimes}\mathtt{b}_\mathtt{t}{\otimes}\mathtt{w}_\mathtt{v}$) estimated on the calibration data. 129
6.10 Results on overall grades on LinGen and LinBus using per-part linear regression estimated on the calibration data. 131

Introduction

In an interconnected world where English has become the lingua franca of culture, entertainment, business, and academia, there has been a growing demand for learning English as a second language (L2) over the last few decades. It has been estimated that it is used by approximately two billion people daily at various proficiency levels (Howson, 2013). Given these premises, speaking ability has become a crucial language skill often defined as essential for social inclusion and integration at all levels (Derwing & Munro, 2009) and can be considered at the core of the four-skills model of writing, speaking, listening, and reading in language education curricula. Its importance is to be attributed, at least in part, to the growing influence of the communicative model in language teaching and assessment (Fulcher, 2000) (see Chapter 1). Therefore, it comes as no surprise that there has been an increasing interest in methods and techniques for automating the otherwise cumbersome, time-consuming, and expensive process of spoken language proficiency assessment both for formal settings and for practice situations in Computer-Assisted Language Learning (CALL).

The discipline of language testing and assessment has its roots in the 1960s (Lado, 1961), but automatic speaking assessment is an even younger field of research (Bernstein et al., 1990) and still has many open questions. This thesis aims to explore novel automatic approaches to automatic speaking proficiency assessment and feedback of L2 learners of English and answer some of these questions.

In their early days, automatic systems for speaking assessment only targeted read-aloud speech (see Chapter 2), and this might have contributed to fueling the common misconception that automatic proficiency assessment would be roughly equivalent to automatic pronunciation assessment or, at least, to devoting a great deal of attention to the acoustic aspects of language at the expense of other equally important features, especially between the 1990s and the early 2000s. On the contrary, proficiency is a multifaceted construct composed of formal and content-

related aspects that have been variously defined and weighted in the succession of language assessment models from the beginning of the scientific era of language testing and assessment to the consolidation of the communicative approach (Canale & Swain, 1980), which has informed language teachers, testers, and researchers for the last 40 years. In light of the complex and multidimensional nature of proficiency, not only is it paramount to provide reliable, valid, and accurate assessment of L2 learners, but also to give them feedback on specific aspects of proficiency, highlighting their strengths and weaknesses. In addition to holistic proficiency scores, the main forms of feedback explored in this thesis are grades related to particular facets of proficiency (i.e., grammar, pronunciation, rhythm, fluency, etc.) and feedback on grammatical errors. We conduct our investigation by mainly focusing on the application of deep learning approaches, which have been shown to bring considerable improvements in this field (see Chapter 2). In particular, self-supervised learning (SSL) representations of speech and text have been proven to be exceptionally powerful and yield remarkable results, but they come at the cost of low explainability, thus potentially compromising the validity of results. In this thesis, we also attempt to address this issue.

Another common problem in the field of automatic speaking assessment is the lack of publicly available data specifically designed and annotated for this purpose (see Chapter 3) since the transcription, annotation, and scoring process of speech recordings could be costly and time-consuming. The transcription operation could be performed automatically, thus alleviating this effort, but this may result in inaccurate representations, as it is well known that automatic speech recognition (ASR) of L2 learner speech still constitutes a difficult task. A further issue often occurring in the scoring phase is that learners' proficiency is typically assessed holistically as opposed to analytically (i.e., by focusing on individual facets of proficiency). Therefore, learners might receive indications on how to improve their language skills from a global perspective but not with particular attention to specific aspects of proficiency. These are other problematic aspects that we address in this work.

In light of the theoretical aspects, state-of-the-art approaches, their advantages, issues, and limitations mentioned above and illustrated in detail in the following chapters, we pose the following research questions that we address in this thesis:

1. How can we increase the performance of automatic speaking assessment systems based on objective elements present in the data?

2. How can we increase the validity and interpretability of results and provide informative

feedback to learners, teachers, and testers?

3. How can we assess communicative competence in speaking automatically?

And considering implementation aspects specifically:

4. How can we use written data in order to assess speaking proficiency?

5. How can we assess speaking proficiency automatically, avoiding transcriptions?

In Chapter 1, we provide a review of the history of L2 assessment from its origins to the present, describing some of the most influential language assessment models with particular attention to the assessment of speaking considering the underlying aspects of communicative competence. The componential aspects of the communicative model also constitute the frame of the core of Chapter 2, which starts with an introduction about the general aspects of automatic speaking assessment, discusses advantages, challenges, and limitations, and includes a brief historical background. In Chapter 3, we describe the data used in our studies, considering both spoken and written corpora. For completeness, we also include an outline of other corpora which have not been considered in our experiments. Chapter 4 contains two studies which investigate the relationships and interconnections between mastery of grammar and proficiency assessment. While Study 1 explores the use of written data and the impact of features extracted through grammatical error detection (GED) on speaking proficiency assessment, Study 2 illustrates a pipeline which starts from disfluency detection (DD), passes through grammatical error correction (GEC), and ends with speaking proficiency assessment. In Chapter 5, Study 3 investigates whether it is possible to use systems targeting specific facets of proficiency when only holistic scores are available. In Chapter 6, we explore approaches based on SSL speech representations in two studies. In Study 5, we investigate the tasks of predicting holistic and analytic proficiency using a relatively small amount of data and investigating possible combinations of different models. In Study 6, we extend the work illustrated in Study 5, using a larger amount of data derived from a multi-part language examination and providing further comparisons and combinations. In the final chapter, we discuss the findings, implications, and limitations of the experimental results and future perspectives and summarise the conclusions of this thesis.

List of publications

Published

- Gretter, R., Matassoni, M., **Bannò, S.**, & Falavigna, D. (2020). TLT-school: a Corpus of Non Native Children Speech. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 378-385). URL: aclanthology.org/2020.lrec-1.47

- **Bannò, S.**, Matassoni, M., & Simakova S. (2021). Towards error-based strategies for automatically assessing ESL learners' proficiency. In *Collated Papers for the ALTE 7th International Conference, Madrid* (pp. 148-153). URL: alte.org/resources/Documents/ALTE%207th%20International%20Conference%20Madrid%20June%202021.pdf

- **Bannò, S.**, & Matassoni, M. (2022). Cross-corpora experiments of automatic proficiency assessment and error detection for spoken English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (pp. 82-91). doi: 10.18653/v1/2022.bea-1.12

- Lu, Y., **Bannò, S.**, & Gales, M. J. F. (2022). On assessing and developing spoken 'grammatical error correction' systems. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (pp. 51-60). doi: 10.18653/v1/2022.bea-1.9

- **Bannò, S.**, Balusu, B., Gales, M. J. F., Knill, K. M., & Kyriakopoulos, K. (2022). View-specific assessment of L2 spoken English. In *Proceedings of Interspeech 2022* (pp. 4471-4475). doi: 10.21437/Interspeech.2022-10691[1]

- **Bannò, S.**, & Matassoni, M. (2023). Proficiency assessment of L2 spoken English using wav2vec 2.0. In *2022 IEEE Spoken Language Technology Workshop (SLT)* (pp. 1088-1095). doi: 10.1109/SLT54892.2023.10023019

Accepted

- **Bannò, S.**, Rais, M., Matassoni, M. (2023). Towards automatic spoken grammatical error correction of L2 learners of English. Submitted and accepted to *Workshop "AI ed*

[1]This paper was also presented at the poster session of UK Speech 2022 (conferences.inf.ed.ac.uk/ukspeech2022/tech-prog.html) and some of its parts, as well as parts of Lu et al. (2022), were presented by Dr. Kate Knill at the International Symposium on Chinese Spoken Language Processing (ISCSLP 2022) in her keynote speech (iscslp2022.org/).

Educazione" - Ital-IA 2023.

Submitted

- **Bannò, S.**, Knill, K. M., Matassoni, M., Raina, V., & Gales, M. J. F. (2023). Assessment of L2 Oral Proficiency Using Self-Supervised Speech Representation Learning. Submitted to *Interspeech 2023*.
- **Bannò, S.**, & Matassoni, M. (2022). Automating the assessment of communicative competence in speaking: current trends, challenges, and future perspectives. Submitted to *Language Assessment Quarterly*.
- **Bannò, S.**, & Matassoni, M. (2023). Back to grammar: using grammatical error correction to automatically assess L2 speaking proficiency. Submitted to *Speech Communication*.

Chapter 1

Theoretical framework

In this chapter, we briefly review the history of second language assessment from its origins to the present, focusing on some of the most influential language assessment models. The second part of the chapter focuses on the assessment of speaking and reviews various studies conducted in this area and is divided into sections covering the individual aspects which compose the communicative competence model.

1.1 From structuralism to the CEFR

The origins of the field of L2 assessment date back to the influential work of Lado (1961), who believed that the problems of learning a new language could be described and explained by comparing the learners' first language (L1) language and their target L2 language, consistently with his structuralist perspective of language and contrastive linguistics. Language was taught — and therefore assessed — as a set of isolated and discrete elements, starting from a contrastive analysis of decontextualised phonemes, lexicon, and grammar. In an attempt to measure language objectively, the principles of structural linguistics were integrated with psychometrically based testing. Therefore, tests typically included multiple-choice, true-false, and other types of objective items. This model of language testing, also referred to as the skills-and-elements approach and articulated mainly by Lado (1961) and Carroll (1961, 1968), made a clear distinction between skills and elements of proficiency. In this approach, the aspect related to the "skills" included listening comprehension, spoken production, reading, and writing, whereas the set of "elements" included vocabulary, pronunciation, grammatical structure, and cultural meanings.

This approach was extremely influential and informed a generation of large-scale L2 assessments in the United States. Although such a perspective is no longer as fashionable as it was in the 1960s and 1970s, several elements — not strictly related to language testing and assessment — originated from structuralism and contrastive linguistics are still valid, such as language transfer and interference, which play a major role in second language acquisition and assessment.[1] Furthermore, specifically with regard to L2 language testing and assessment, the four-skills model (i.e., listening, speaking, reading, and writing) conceived in the 1960s is still at the core of most language tests and curricula.

Starting from the late 1970s, following the changes in language teaching and language use, the approach to language assessment had also changed and, in contrast to the structuralist view, a psycholinguistic-sociolinguistic trend emerged and shifted focus from an assessment based on individual language elements to a holistic view of language proficiency. Inspired by research into the nature of intelligence, specifically Spearman's theory of general intelligence (Spearman, 1904), Oller (1979) introduced the so-called "unitary trait hypothesis", according to which language proficiency is fundamentally an individual holistic ability. As a result, such an ability could be best assessed by means of global and integrative measures, such as cloze and dictation (Oller, 1979) and context-based and specific purpose tests (Morrow, 1977, 1979; Carroll, 1978). Learners were required to prove their ability to use the various aspects of linguistic knowledge (grammar, vocabulary, spelling, etc.) in combination. Although Oller (1983) himself eventually recognised that the unitary competence hypothesis was unfounded, his work had a lasting influence in the field of second language testing and assessment.

The subsequent paradigm shift in language testing and assessment was inspired by the forward-looking work on communicative competence by Hymes (1972), later refined and framed in the so-called communicative approach by Canale & Swain (1980), Canale (1983) and further by Bachman (1990) and Bachman & Palmer (1996). According to Canale's model (1983), language is used to communicate meaning, which encompasses:

- grammatical competence: the ability to use language accurately and correctly, including accurate construction of words and sentences and correct lexis, spelling, and pronunciation;
- sociolinguistic competence: the ability to use and understand language based on different contexts, including elements such as register choice;

[1] The construct of interference is to be attributed to Weinreich (1953), who first formulated a theorisation of the processes of contact and interference, whereby the first indicates the encounter of two or more language varieties in the competence of a speaker in reference to the potential exposure to interlinguistic influence, while the latter refers to the actual realisation of such encounter in a speaker's utterance.

- strategic competence: the ability to use verbal and non-verbal communication strategies in order to "compensate for breakdowns in communication due to performance variables or insufficient competence" and to "enhance the rhetorical effect of utterance" (Canale, 1983, p. 339);
- discourse competence: the ability to combine and interpret appropriate forms and meanings, applying coherence and cohesion rules appropriately, in order to produce unified texts in different modes (spontaneous conversation, argumentative essay, narrative essay, etc.).

CEFR level	**Level description**
A1	Beginner
A2	Elementary
B1	Intermediate
B2	Upper intermediate
C1	Advanced
C2	Upper advanced

Table 1.1: CEFR levels description.

In the late 1990s, the communicative approach was fixed in the Common European Framework of Reference (CEFR) (Council of Europe, 2001), which was meant to provide "a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc., across Europe" and to "overcome the barriers to communication among professionals working in the field of modern languages arising from different educational systems in Europe" (p. 1). The approach outlined in the CEFR is often referred to as "action-oriented", implying that L2 learners are primarily seen as "social agents", i.e., members of society who act and have tasks to achieve in a specific environment within a specific field of action and under specific circumstances. The six CEFR levels are structured according to 'can-do' descriptors of language proficiency outcomes, especially in relation to communicative competence, and range from A1 to C2 (see Table 1.1). We report an example of a proficiency descriptor drawn from Overall Oral Production at B2 level:

- "B2. Can give clear, detailed descriptions on a wide range of subjects, related to his/her field of interest, presenting it as a linear sequence of points" (Council of Europe, 2001, p. 27).

Despite some limitations (Weir, 2005; Alderson, 2007; Hulstijn, 2007), the CEFR has become an established standard and has informed language teaching, testing, and assessment for the

last 20 years, gradually expanding from Europe to the rest of the world. Drawing on different competence models developed in the 1980s, the CEFR presents its scales for various aspects of communicative language competence under three headings:[2]

- linguistic competence, which includes general linguistic range, grammatical accuracy, orthographic control, phonological control, vocabulary control, and vocabulary range;
- sociolinguistic competence, which involves only one indicator, i.e., sociolinguistic appropriateness;
- pragmatic competence, which encompasses fluency, coherence and cohesion, propositional precision, thematic development, turn-taking, and flexibility.

1.2 Assessment of speaking

Globalisation, technological advances, and recent changes in educational and working habits and lifestyles due to the COVID-19 pandemic have brought together — both physically and virtually — more and more people from various cultural and linguistic backgrounds. In such an interconnected world, L2 learners of English are steadily increasing not only in English-speaking countries but also in other areas of the world where English is the lingua franca of business, culture, entertainment, and academia, and speaking ability has become a crucial language skill often highlighted as indispensable for social inclusion and integration at all levels (Derwing & Munro, 2009).

Therefore, speaking might arguably be considered at the core of the four-skills model of writing, speaking, listening, and reading, and has the goal of "developing learners' fluency and accuracy, as well as their sociocultural communicative competence requiring adapting the language from context to context and from genre to genre" (Hinkel, 2010). The importance of speaking is also to be attributed, at least in part, to the growing influence of the communicative model in language teaching and assessment (Fulcher, 2000). It is worth noting that in the field of general linguistics, several scholars stated the primacy of spoken language compared to written language. De Saussure wrote that "[l]anguage and writing are two distinct systems of signs; the second exists for the sole purpose of representing the first" (de Saussure, 2011, p. 23); Bloomfield went as far as to declare that "writing is not language, but merely a way of recording

[2] We refer to the updated version of the CEFR, whose conceptual framework remains valid (Council of Europe, 2020).

language by means of visible marks" (Bloomfield, 1933, p. 21); similarly, but less radically, Lyons stated that "the spoken language is primary and [...] writing is essentially a means of representing speech in another medium" (Lyons, 1968, p. 38). Jespersen considered speaking a primary function but associated it with listening as opposed to the skills of writing and reading: "the spoken and heard word is the primary form for language, and of far greater importance than the secondary form used in writing (printing) and reading" (Jespersen, 1924, p. 2). Specifically in the field of language learning, some years later, Lundeberg seemed to echo Jespersen's position and asserted that "oral-aural skills are today recognized as desirable and attainable objectives in the instructional program. A great many teachers and administrators [...] rank the attainment of ear and tongue skills very high among their objectives. The layman, especially the parent, would often have this practical phase of language study placed first in the list" (Lundeberg, 1929, p. 193). More than three decades later, Lado also defined speaking as "the most highly prized language skill" (Lado, 1961, p. 239). In addition to pre-structuralist and structuralist viewpoints, the centrality of speaking ability has been recently reiterated, as it has been defined as the primary mean for a learner to acquire a language (Lazaraton, 2014). However, despite its importance, its inclusion in large-scale assessment tests has often been unclear and ambiguous. While standardised tests, such as the Certificate of Proficiency in English (CPE) created in 1913 in Cambridge, contain a compulsory speaking part,[3] other high-stakes language exams have included mandatory speaking tasks only recently. For example, the Test of English as a Foreign Language (TOEFL) incorporated a speaking part only in 2005 (Weir et al., 2013).[4]

Despite (and because of) its supposed primacy, speaking assessment comes with some intrinsic challenges. Given the transient and ethereal nature of speech, the speaking ability "is possibly the most difficult skill to teach, the most difficult skill to assess and the most difficult skill to investigate" (Lowie et al., 2018, p. 105), especially when it comes to assessing a learner's performance by means of a direct speaking test, i.e., a face-to-face oral interview between the learner and one or multiple interlocutors. The alternative, enabled by technological advances, is the so-called semi-direct speaking test, which involves the use of a recording device that captures learners' answers without the presence of a human interviewer, and might often result

[3]The CPE even contained a conversation task other than more conventional read-aloud and oral dictation tasks.

[4]This is quite surprising, considering that already in 1944 there was a certain insistence on prioritising spoken language essentially for practical reasons such as military training in the United States: "The primary concern is ability to understand the spoken language and to speak the foreign tongue. In elementary and intermediate courses, reading and writing are introduced, if at all, only in the service of these paramount abilities. The urgency of the world situation does not permit of erudite theorizing in English about the grammatical structure of the language for two years before attempting to converse or to understand telephone conversations" (Kaulfers, 1944, p. 137).

in a more accurate and efficient assessment, which can be performed at a later time (hence the term "semi-direct") by a human evaluator or an automatic system. Historically, the American assessment tradition has been theoretically driven by psychometric criteria more than the British tradition, with the first preferring semi-direct testing, e.g., in TOEFL, thus ensuring assessment reliability, validity, and fairness by replicating the same testing conditions across learners. On the other hand, the latter has typically preferred to choose direct testing as the ideal assessment mode, e.g., in the International English Language Testing System (IELTS), thus highlighting authenticity and interaction. Despite evident differences between the two assessment traditions, it is difficult to decide which one is better at a global level, as both show advantages and disadvantages (Isaacs, 2017a). In fact, although test takers seemingly tend to favour direct testing because of the possibility for the interviewer and interviewee to interact during the exam, it should be noted that semi-direct testing is often more efficient and generally cheaper (Qian, 2009). There would be a third mode of speaking assessment, which is referred to as indirect and consists of a paper-and-pencil exam that aims at assessing abilities underlying the speaking skills the examiner intends to target. According to Lado (1961), there would be a strong correlation between the written and oral productions of the tested word or sentence, but such a claim has been proven to be unfounded and indirect tests are now considered inaccurate and unreliable for assessing spoken proficiency (O'Loughlin, 2001).

If we consider the CEFR model outlined in the previous section, when specifically assessing spoken proficiency, the three competences mentioned above should be adapted to this specific skill. Therefore,

- linguistic competence would result in the ability to make accurate use of the grammatical rules that underpin spoken language, to pronounce words correctly, and to use a variety of expressions appropriately;
- through sociolinguistic competence, learners should prove their ability to use the appropriate register based on the context showing awareness of politeness conventions and dialect and accent differences;
- pragmatic competence consists of the speaker proving how to interact in a real-world communicative situation and handle the flow of conversation by using compensatory strategies both verbally (e.g., hesitation, questioning, etc.) and non-verbally (e.g., gestures) and showing the ability to fluently formulate coherent and cohesive ideas.

It should be noted that between these three competences there are evident overlaps and grey

areas since, in some cases, it is objectively difficult to draw clear lines of demarcation between one skill and another. For example, learners' use of vocabulary, which falls under the umbrella of linguistic competence, may affect the appropriateness of idiomatic expressions, which, as a matter of fact, should be part of sociolinguistic competence. Similarly, sociopragmatics may be considered part of sociolinguistic competence as well as pragmatic competence. In light of this, especially sociolinguistic and pragmatic competences and their indicators have been arranged differently in the succession of models of communicative competence. While both Canale & Swain (1980) and Canale (1983) cover pragmatics under sociolinguistic competence and discourse competence, Bachman & Palmer's model (1996) incorporates what the authors call "sociolinguistic knowledge" and "functional knowledge" into a wider area labelled "pragmatic knowledge". Instead, it seems that the CEFR model follows the distinction made by Leech (1983) between pragmalinguistics and sociopragmatics, whereby pragmalinguistics pertains to the ability to use relevant linguistic devices in order to perform a specific speech act, whilst sociopragmatics can be described as the ability to perform a speech act which is appropriate to a specific situation or context (Grabowski, 2016). As a result, "pragmalinguistic failure is basically a *linguistic* problem, caused by differences in the linguistic encoding of pragmatic source", while "sociopragmatic failure stems from cross-culturally different perceptions of what constitutes appropriate linguistic behavior" (Thomas, 1983, p. 99). In this thesis, this type of distinction is adopted, notwithstanding that some intersections between competences, skills and indicators are intrinsically inevitable.

In the following paragraphs, linguistic competence is treated by analysing its three main elements, i.e., grammar, pronunciation, and vocabulary. Sociolinguistic competence is articulated by focusing on idiomaticity and sociopragmatics. As regards pragmatic competence, the reader will notice that some of its aspects, as described in the CEFR, are not treated in detail or are incorporated into other elements. In particular, propositional precision and thematic development tend to pertain to coherence and cohesion, whereas flexibility appears to be more related to fluency aspects.[5] Turn-taking is a distinctive element of conversational tasks, which are not investigated in this thesis. Therefore, we compressed the aspects of pragmatic competence into two main sections: coherence and cohesion and fluency (see Fig. 1.1).

[5]The authors of the CEFR (Council of Europe, 2020, p. 142) acknowledge that "[f]luency [...] has a broader, holistic meaning [...] and a narrower, technical and more psycholinguistic meaning [...]. The broader interpretation would include "Propositional precision", "Flexibility", and at least to some extent "Thematic development" and "Coherence/ cohesion"." Other problematic aspects of pragmatic (and, to some extent, sociolinguistic) competence are discussed further in this and the next chapter.

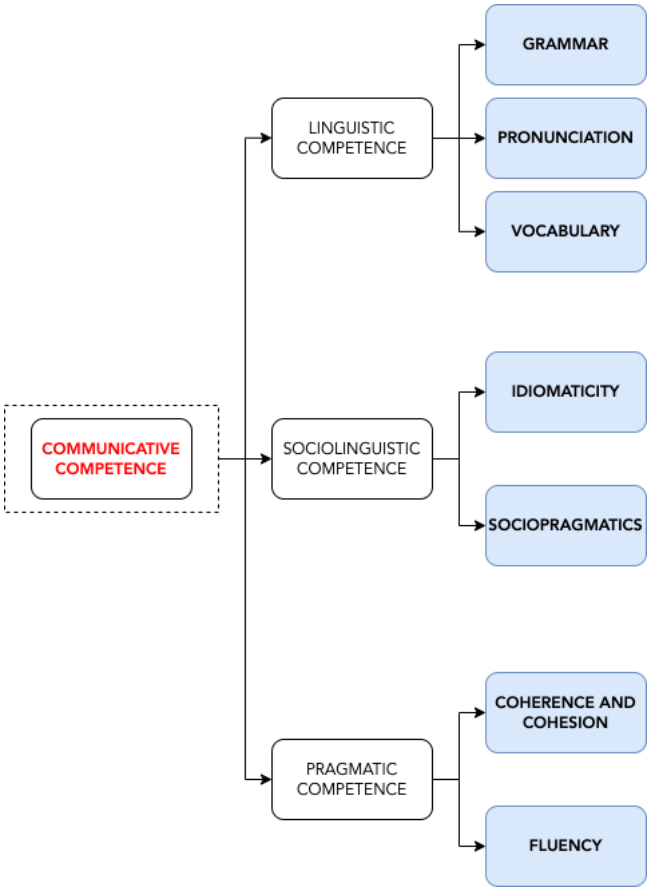


Figure 1.1: Diagram of communicative competence as outlined in this thesis.

1.2.1 Linguistic competence

Grammar

As previously said, according to the structuralist approach, the problems of learning a new language could "be predicted as described in most cases by a systematic linguistic comparison of the two language structures" (Lado, 1961, p. 24), i.e., the learner's L1 and L2. As a result, grammar plays an important role in this construct, especially in relation to those grammatical structures and errors that are traceable to specific contrasts between a learner's L1 language and English. Lado's skills-and-elements construct of L2 proficiency is at the foundation of the Comprehensive English Language Test (CELT) (Harris & Palmer, 1986), in which proficiency is assessed based on discrete grammatical and lexical elements. When the goal is the assessment of individual isolated forms, this approach still remains beneficial, but it excludes the possibility

of assessing grammatical forms in relation to their semantic and pragmatic meanings (Purpura, 2014).

In response to and in continuation of contrastive analysis, at the end of the 1960s, the seminal work of Corder (1967) set the foundation for error analysis and provided an effective distinction between the concepts of error and mistake. According to Corder, an error is a failure in learners' competence, whereas a mistake is a failure in learners' performance. As a corollary of this, second language learners are aware of their own mistakes but not of their own errors. Therefore, errors are systematic and are useful to highlight a learner's knowledge and gaps, whereas mistakes occur randomly and are due to "memory lapses, physical states such as tiredness and psychological conditions such as strong emotion" (Corder, 1981, p. 10). Other features that are helpful to distinguish errors from mistakes are self-correction (or corrigibility) and intentionality, i.e., mistakes are (or can be) self-corrected and are committed deliberately, while errors are not (James, 1998). Corder (1981) also considered the concept of error from a developmental perspective, i.e., errors typical of any learner, independently of their L1, at a particular stage in learning English. As a result, errors would be highly useful indicators of a learner's proficiency level.[6] However, error analysis rapidly began losing momentum, as its critics highlighted several methodological and functional shortcomings, such as the employment of heterogeneous and unstructured learner data, the fuzziness of error categories, the difficulty with handling the so-called strategy of avoidance (i.e., learners can substitute certain language elements that they are not sure how to use with others that they feel more confident with), the sometimes too restricted focus merely on what learners cannot do, and the fact that the L2 representation provided by error analysis would often result in a static picture (Ellis, 1994; Scholfield, 1995; Harley, 1980; van Els et al., 1984).

In the late 1990s, since computers started becoming more powerful and more accessible and large-scale corpora started to be commonly used for studies on second language acquisition (Granger et al., 2002), testing and assessment (Barker, 2006), error analysis experienced a new dawn, introduced with the term "Computer-aided error analysis" by Dagneaux et al. (1998). The use of technological instruments made error analysis more accurate and systematic and made it possible to address critical issues mentioned above.

In the same years, the communicative approach was increasingly consolidating. Although it

[6] From this perspective, errors are highly significant for three reasons: "First, to tell the teacher how far towards the goal the learner has progressed and, consequently, what remains for him to learn. Second, they provide to the researcher evidence of how language is learned or acquired, what strategies or procedures the learner is employing in his discovery of the language. Thirdly, to the learner because we can regard the making of errors as a device the learner uses in order to learn" (Corder, 1981, p. 12).

might seem that this approach privileges communication at the expense of formal correctness, errors still play a major role in language teaching and assessment (Pfingsthorn, 2013). In a study focused on grammatical errors across proficiency levels in written and spoken productions of Japanese learners of English, Abe (2007) found that certain errors systematically decrease as proficiency levels increase. Specifically, in the spoken data, the accuracy rates for the use of articles and for some types of prepositions (subordinating prepositions and prepositions in phrasal verbs) dramatically increased throughout proficiency levels. Moreover, the study reported that verbal errors were strongly connected to lower-level learners, whereas nominal errors characterised advanced-level learners.

As regards written production, Hawkins & Buttery (2010) found that certain errors tend to follow progressive learning patterns across proficiency levels, i.e., they gradually tend to decrease as proficiency levels increase. For example, errors related to derivation of determiners (e.g., *Shes name is Esther*) or to form of determiners (e.g., *I have an computer*). On the other hand, other types of errors have "inverted U-patterns", i.e., errors increase towards the middle proficiency levels and decline again by C2 level. This category includes noun agreement errors (e.g., *One of my mate*) and missing noun errors (e.g., *It is an interesting*). The "inverted U" is a common pattern since, in lower levels of proficiency, the distribution of a given item which is unknown or new to L2 learners tends to be sparse; therefore, the error rate of that item is generally low. As proficiency levels increase, the item tends to be used more widely and frequently, and, as a result, it is often used inappropriately or incorrectly; therefore, the error rate starts rising. Typically, due to error correction and increasing practice with and exposure to the item, it eventually stabilises and becomes a marker of competence for a specific CEFR level. Similarly, in an article addressing the issue of L2 accuracy developmental trajectories, Thewissen (2013) investigated 45 types of errors and associated them with proficiency levels. However, unlike Hawkins & Buttery (2010), no U-shaped developmental pattern was found, but this absence could be also due to the relatively small amount of data considered in the study and the exclusion of A1 and A2 learners' texts.

It is also important to mention the English Grammar Project, which is a database of more than 1,200 statements derived from a large written corpus and includes various grammatical structures linked to CEFR levels. Interestingly, the researchers involved in the project are also planning to replicate this type of study on spoken data (O'Keeffe & Mark, 2017).

Specifically for spoken production, an interesting correlation between learners' proficiency levels and grammatical accuracy was found in Iwashita et al. (2008) and De Jong et al. (2012).

In the first study, the measure related to global grammatical accuracy is one of the four variables (the other three being speech rate, a measure related to vocabulary, and a global pronunciation measure) that influence proficiency scores the most. In the second study, an indicator related to knowledge of grammar strongly correlated with proficiency levels.

Grammatical proficiency is not only determined by grammatical accuracy but also by syntactic complexity, which comprises the range of complexity of the syntactic structures produced by a learner (Ortega, 2003; Lu, 2011). Iwashita (2006) reported that the length of T-units[7] and the number of clauses per T-unit are efficient indicators to predict learner proficiency. In addition to investigating global grammatical accuracy, the previously cited study by Iwashita et al. (2008) also found that proficiency levels of the candidates taking part in the speaking section of the TOEFL iBT were correlated with the mean length of their utterances and, more interestingly, with the number of verb phrases per T-unit. The study by Lambert & Nakamura (2019) also found that four clause combination strategies (coordination, nominal subordination, adverbial subordination, and relative subordination) varied systematically across proficiency levels in six communication tasks performed by Japanese learners of English.

Pronunciation

Pronunciation also falls under the umbrella of linguistic competence in the CEFR, but it has not been given much attention in well-known communicative models of language assessment (Canale & Swain, 1980; Bachman, 1990; Bachman & Palmer, 1996), despite their primary focus on getting the message across. This lack is particularly unexpected, considering that pronunciation is one the most perceptually important elements of spoken language: it has been demonstrated that listeners with no prior linguistic training are able to distinguish between L1 and L2 speech under non-optimal conditions, i.e., only by listening to a 30-millisecond recording (Flege, 1984), to a recording containing speech in a foreign language (Wester & Mayo, 2014) even when this language is unfamiliar (Major, 2007), and even to a recording played backwards (Munro et al., 2010). Furthermore, even religious and folk literature is full of examples that testify to the — literally — vital importance of pronunciation: Spolsky (1995) mentions the Shibboleth test narrated in the Book of Judges (12:5-6) as one of the earliest documented pronunciation assessment test, which, however, had no educational purpose and had instead lethal consequences for those who failed it.[8] Similar linguistic anecdotes are widespread in other cultures, ages, and places, e.g.,

[7] A T-unit is a dominant clause with one or more subordinate clauses attached to it (Hunt, 1965).

[8] The term Shibboleth comes from the Hebrew word *shibbóleth*, which indicates either the part of a plant containing grain or a torrent. In the biblical story, the inhabitants of Gilead defeated the tribe of Ephraim

the Sicilian Vespers and the Matins of Bruges, and contribute to emphasising the importance of pronunciation for speakers and listeners.[9] Despite all this, pronunciation has often been neglected in the field of second language testing and assessment, and, therefore, Lado's (1961) contribution, albeit mostly outdated, has remained the only extensive work on the subject until the recent publication of the works by Isaacs & Trofimovich (2016), Kang et al. (2017), and Levis et al. (2022). One of the reasons for the exclusion, or at least disregard, of "the Cinderella of language teaching" (Kelly, 1969, p. 87) in several assessment frameworks is attributed to the belief that excessive attention to pronunciation may constitute an obstacle to achieving objectives such as communicative effectiveness (Celce-Murcia et al., 2010). Another important issue lies in the fact that the construct of pronunciation has often been ill-defined and has suffered from a dichotomic approach, caught between the two fires of the "nativeness principle" and the "intelligibility principle" (Levis, 2005): the first maintaining that the goal of language teaching with respect to pronunciation should be to eradicate L1 influence from speech and attain L1-like pronunciation; the latter holding that its aim should be intelligibility. Intelligibility has been variously defined, but most researchers refer to it as "the extent to which a speaker's message is actually understood by a listener" (Munro & Derwing, 1995, p. 76), and it is generally measured by the ratio of an L2 speaker's utterance that a listener can correctly transcribe.[10] On the other hand, comprehensibility "is defined as listeners' *perceptions* of how easily they understand L2 speech" (Isaacs, 2014, p. 5). Although in the pre-scientific era of language assessment, the trend was in favour of the achievement of a L1-like accent (Kaulfers, 1944), nowadays, most L2 researchers endorse the "intelligibility principle" (Isaacs, 2014) because of several reasons: first, there is no necessity for L2 learners to sound like L1 speakers in order to pursue integration and educational or professional success (Derwing & Munro, 2009); secondly, since many L1 English speakers themselves do not speak standard varieties, such as General American English, Received Pronunciation, or General Australian English, it is hard (and perhaps futile) to keep defining the meaning of 'native-like' in its traditional acceptation (Seidlhofer, 2018); finally,

and invaded its territories. The survivors among the Ephraimites tried to cross the River Jordan back to their settlements, but the inhabitants of Gilead occupied the fords of the river. In order to identify and eliminate these Ephraimite fugitives, the Gileadites required them to utter the word shibboleth (/ʃiˈbɔlɛt/). In the dialect of the Ephraimites, however, the initial consonant sound was pronounced as /s/, making the word sound like /siˈbɔlɛt/.

[9] Similarly to the story narrated in the Book of Judges, during the uprising of the Sicilian Vespers in 1282, the inhabitants of Sicily killed the French invaders who could not pronounce the word *ciciri* ("chickpeas") appropriately. Likewise, according to an anecdote related to the Matins of Bruges, before the Battle of the Golden Spurs in 1302, the Flemish identified and killed the French based on their pronunciation of the syntagm *schild en vriend* ("shield and friend").

[10] With respect to intelligibility, Lado's (1961) question is still valid and has not been properly addressed: "This standard, however, is hard to define. Intelligible to native speakers, but what native speakers? A native speaker that has been in contact with foreign speakers will understand that sound entirely foreign to another native speaker."

accent and identity are interconnected, therefore removing any traces of L1 accent may not be desirable for learners (Gatbonton & Trofimovich, 2008).[11] The difficulty in building a construct of pronunciation is also an issue for internationally recognised language tests, such as the TOEFL iBT and the IELTS, since their descriptors related to this specific skill are generic and unclear. For instance, the level 3 descriptor of the "Integrated Speaking Rubrics" of the first reads: "Speech is generally clear, with some fluidity of expression, but it exhibits minor difficulties with pronunciation, intonation, or pacing and may require some listener effort at times" (Educational Testing Service, 2009, p. 187), moreover making an ambiguous distinction between pronunciation and intonation. Similarly, the band 6 pronunciation descriptor of the public version of the IELTS speaking scale reads: "uses a range of pronunciation features with mixed control; shows some effective use of features but this is not sustained; can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times".[12]

Only in recent years, studies on pronunciation have regained momentum, and there has been a growing interest in investigating pronunciation both holistically and analytically considering its two componential phenomena, i.e., segmental (individual phonemes) and suprasegmental (rhythm, intonation, prosody, word stress) features, mainly due to the impact of the findings and applications in the field of automatic assessment and in the literature related to international teaching assistants (Isaacs, 2014).

In L2 speech, segmentals are often altered due to the influence of L1 transfer on both vowel and consonant sounds. Such deviations occur as replacements or alternations of a sound or even as additions, deletions, and reordering of segments or syllables (Anderson-Hsieh et al., 1992). For instance, it is well-known that Italian L2 English learners' speech is often characterised by epenthetic schwas, i.e., the addition of /ə/ at the end of words ending in a consonant or a consonant cluster (Broniś, 2016; Grice et al., 2018). Another example is described in Kartushina et al. (2015), in which Italian speakers have difficulties producing the vowel contrast /ʌ/ and /ɑ/ (e.g., *hut* and *hot*, respectively), while Spanish and Korean speakers find it hard to realise the vowel contrast /ɪ/ and /i/ (e.g., *fit* and *feet*, respectively). Several studies have investigated the impact of segmental deviations on intelligibility and comprehensibility, coming to different conclusions: non-standard segments, as well as non-standard syllable stress patterns, significantly affect intelligibility (Zielinski, 2008); certain consonant contrasts (e.g., /ʃ/ versus /s/) are more disturbing to comprehensibility and intelligibility than others (e.g., /f/ versus /θ/) (Isaacs, 2014); Derwing

[11] On the other hand, some learners might want to achieve L2 accent-free speech due to stigmatisation of accents and regional varieties (Moyer, 2013).

[12] `ielts.org/-/media/pdfs/speaking-band-descriptors.ashx`

& Munro (1997), instead, show that, although L1 listeners identify segmental deviations as the primary source of a foreign accent, they find that errors have a limited effect on comprehensibility. Another interesting contribution stems from the application of the functional load principle to language teaching and assessment. Functional load (or phonemic load) refers to the relevance of certain features in making distinctions in a given language. Therefore, features characterised by a high functional load differentiate more words in a given language. The misuse or mishearing of such features may be problematic for intelligibility and comprehensibility. In English, an example can be drawn from contrasts that have a high functional load, such as /d/ and /t/, since they affect the distinction between many different lexical items; and contrasts that have a low functional load, such as /ð/ and /θ/, since they are subject to regional variation and listeners are more likely to 'adjust' their perception (Brown, 1988). This criterion was used by Kang & Moran (2014) to label learners' errors ranging from A2 to C2 on some monologic speaking tasks taken from four Cambridge English exams. As proficiency levels increased, a significant decrease in high functional load errors was found.

Studies comparing the impact of segmental and suprasegmental errors also have come to different conclusions: while it appears that suprasegmental errors affect intelligibility more than segmental errors in the study conducted by Anderson-Hsieh et al. (1992), other studies have shown that intelligibility and comprehensibility of L2 speech are compromised by both types of errors (Jenkins, 2009; Hahn, 2004; Saito et al., 2016; Bøhn & Hansen, 2017). The term "suprasegmentals" is often used as a synonym for "prosody", and it refers to pronunciation features such as rhythm, intonation, lexical stress or word stress (i.e., the stress placed on syllables within words), and sentence stress or prosodic stress (i.e., the stress placed on words within sentences). Particularly, prosodic stress can be considered a foundational element of rhythm and marks the words that are crucial for understanding an utterance. Speakers shift stress in some words or clusters of words in various ways to produce rhythm in their speech, as well as they can change their intonation (i.e., variation of vocal pitch) in order to convey grammatical information (e.g., to mark sentences and clauses or to differentiate questions from statements) and to express emotions and personal attitude (Crystal, 2011; Chen et al., 2004). Intonation is also used to communicate meaning and prioritise crucial information (Levis, 1999). Despite a non-unanimous consensus (Arvaniti, 2009), a conventional manner to label languages based on their rhythmic pattern is to distinguish between syllable-timed (e.g., Italian, French, Spanish, Turkish, Mandarin Chinese, Japanese, and Korean) and stress-timed languages (e.g., German, English, Russian and Arabic): while the first have syllables with approximately the same stress and the same duration,

the latter have syllables that may have various duration and are characterised by the recurrence of stressed syllables at constant intervals of time (Taylor, 1981; Ramus et al., 1999). Since there are remarkable differences in how intonation is used in different languages to convey differences in meaning, speakers may encounter severe difficulties handling this feature correctly when facing their target language (Cruz-Ferreira, 1987; Wennerstrom, 1994; Levis, 1999). Similarly, the acquisition of stress-timed English rhythm can be challenging for L2 speakers of both syllable-timed (Anderson-Hsieh & Venkatagiri, 1994) and stress-timed languages (Setter, 2006). In a study examining the impact of segmental and suprasegmental features on the oral proficiency of L2 speakers, Kang (2013) found that pitch variables and stress were the most relevant features, as they accounted for 30.9% of the variance in proficiency scores (fluency took 26.7%, tone choice 4.5% and segmental errors 8%).

Vocabulary

Pronunciation is not the only factor affecting comprehensibility, but also grammar and vocabulary play a major role, especially for L2 learners above a certain proficiency level (Isaacs & Trofimovich, 2012). The psycholinguistic contribution by De Jong et al. (2012) mentioned above also found that intonation and knowledge of vocabulary accounted for 75% of the variance of speaking ability, and this comes as no surprise since "words are the basic building blocks of language, the units of meaning from which larger structures such as sentences, paragraphs, and whole texts are formed"(Read, 2000, p. 1). However, remarkably, the assessment of vocabulary was only selectively investigated at the beginning of the scientific era of language assessment, whereas much more attention was paid to the contrastive analysis of sounds and grammar. This lack was mainly due to the influence of structural linguistics, according to which language is a structure that can be divided into hierarchically organised systems: at the foundation lies phonology, then morphology, and finally syntax. In such a framework, semantics and lexicon received scant attention (Lennon, 2008).[13]

When vocabulary knowledge was assessed, it was mainly tested through the so-called discrete-point approach, i.e., assessing a learner's knowledge of one specific linguistic element — phonology, morphology, syntax, and vocabulary — at a time, generally using multiple-choice questions. Vocabulary tests using this approach were subject to various criticisms. First, it was not possible to deduce an exhaustive judgement on a learner's vocabulary simply based on the score of such

[13]Charles Fries, one of Lado's mentors at the University of Michigan, wrote: "[...] the chief problem is not at first that of learning vocabulary items. It is, first, the mastery of the sound system [...] second, the mastery of the features of arrangements that constitute the structure of the language" (Fries, 1945, p. 3).

a test. Secondly, language proficiency is not only about receptive abilities but also implies that a learner makes effective use of vocabulary for communicative purposes productively in both speaking and writing. Thirdly, in the real world, words do not occur aseptically and randomly in isolated sentences but are integrated into specific contexts. Finally, learners do not have to understand every word of a written or a spoken sentence, and they can use compensatory strategies, e.g., they can guess the meaning of the words they do not understand based on other contextual information or prior knowledge or they can simply ignore such words (Read, 2000).

When the psycholinguistic-sociolinguistic approach gained momentum and started integrating measures such as cloze and dictation, vocabulary started receiving more attention, but, according to the unitary competence hypothesis, it was encapsulated in the holistic view of language proficiency and was tested together with other elements of language by means of so-called tests of integrative skills (Oller, 1973).

However, it was the 1980s that represented the watershed in vocabulary assessment since, in this period, a small group of researchers started to publish studies on defined procedures and measures aiming at assessing specific aspects of vocabulary use and knowledge (Anderson & Freebody, 1981, 1983; Nation, 1983; Meara & Buxton, 1987). These seminal works were something of an exception, given that, on the one hand, the field of second language acquisition was primarily concerned with the investigation of the acquisition by learners of morphological and syntactic features, whereas, on the other hand, the advent of the communicative approach shifted the attention of language assessment researchers from knowledge of grammatical and lexical elements to the performance of real-world-like tasks (Read, 2013). Read (2000) classified vocabulary assessment according to 6 dimensions arranged in antonymic pairs: discrete versus embedded, selective versus comprehensive, and context-independent versus context-dependent. The first binomial refers to the construct underlying a given vocabulary test. By using an embedded measure, vocabulary is only one of many facets that contribute to the assessment of a larger construct of language proficiency, whereas a discrete test considers vocabulary separately from other aspects of language. The second dimension refers to the range of lexical items to be included in the vocabulary test. This is selective if test-takers have to prove their vocabulary knowledge on a set of selected target words, considered individually or integrated into a specific context. On the other hand, the full range of vocabulary is assessed when using a comprehensive measure. The dimension context-independent versus context-dependent speaks for itself, but it is important to stress that "it is necessary to broaden the notion of context to include whole texts and, more generally, discourse" (Read, 2000, p. 11). Due to the widespread acceptance of the

communicative approach, it is straightforward to conclude that current trends in language testing and assessment tend to privilege embedded, comprehensive, and context-dependent measures of vocabulary assessment. In this regard, the CEFR distinguishes between vocabulary range and vocabulary control. The first "concerns the breadth and variety of expressions used" (Council of Europe, 2020, p. 131), and it applies to both reception and production, while the latter refers to "the user/learner's ability to choose an appropriate expression from their repertoire" (Council of Europe, 2020, p. 132). These indicators are generally operationalised along the dimensions of lexical diversity and lexical sophistication, and their development and integration were considerably enhanced by the relatively recent application of corpus linguistics in the field of language testing and assessment (Barker, 2006).

Lexical diversity has as its object the range of vocabulary used by the learner (Yu, 2010; Lu, 2012) and is typically estimated using various features such as the number of different words (NDW), the number of word types uttered or written, the type-token ratio (TTR),[14] or the D measure[15] in a spoken or written text. The connection between measures of lexical diversity and language proficiency, which finds its ideal application in the field of writing assessment (Treffers-Daller et al., 2018), has also been investigated for speech assessment in various ways. In a study mentioned earlier aiming at deconstructing comprehensibility of French learners of English, Isaacs & Trofimovich (2012) found that, out of 19 features targeting fluency, pronunciation, grammar, and vocabulary, the frequency of word types had the strongest correlation with the judgements of comprehensibility of 60 raters, with token frequency being the second most correlated feature. In another study examining the relationship of lexical richness to the quality of oral narratives by L2 learners of English, Lu (2012) found that NDW, TTR, D measure, and Corrected TTR (Carroll, 1964) correlated with human ratings. The word type frequency and the D measure were also investigated and were found to account for a great proportion of variance in human scores of transcriptions of TOEFL iBT Independent speaking sections (Crossley & McNamara, 2013).

The focus of lexical sophistication is the depth and breadth of lexical knowledge and "is often simply described as the number of "unusual" words in a sample" (Baese-Berk et al., 2021, p. 4). It is generally operationalised using features related to word frequency and familiarity, such as the Lexical Frequency Profile, i.e., "the percentage of words a learner uses at different

[14]The TTR is computed by dividing the total number of word types by the total number of word tokens in a text. Each unique word is referred to as a word type, while the individual occurrences of word type are called word tokens. For example, in the sentence "A rose is a rose is a rose", there are three word types and eight word tokens.

[15]Since text length heavily affects the TTR, the D measure (Malvern et al., 2004), a complex mathematical transformation of the TTR, is used to avoid sample size effects.

vocabulary frequency levels" (Laufer & Nation, 1995, p. 311) based on a word frequency list taken from a corpus used as a reference point. For writing assessment, the English Vocabulary Profile (Capel, 2015), a resource that describes words and phrases used by English learners at different CEFR levels, has been employed to assign proficiency bands to 90 essays, finding a strong correlation between the clusters obtained using the vocabulary profile and the human-assigned CEFR levels (Leńko-Szymańska, 2015). In a previously cited work by Iwashita et al. (2008), the authors found that both the proportion of low- and high-frequency word tokens and the proportion of low- and high-frequency word types were associated with different proficiency levels in the TOEFL iBT speaking sections. Lu (2012) also found a correlation between human scores and sophistication features related to verb usage.

1.2.2 Sociolinguistic competence

Idiomaticity

Vocabulary is intertwined with the sociocultural context in which learners act as "social agents". Therefore, one cannot help but notice that vocabulary can also serve as a valuable indicator for assessing sociolinguistic competence (Read, 2000). For instance, the main feature to characterise register is the use of distinctive words and phrases (McCarthy, 1990).

In particular, the CEFR stresses the importance of "employing idiomatic expressions, allusive usage and humour" and "recognising sociocultural cues, especially those pointing to differences, and acting accordingly" (Council of Europe, 2020, p. 136). Formulaic expressions are a fundamental component of L1 speakers' everyday conversations (Pawley & Syder, 1983) to the extent that Jackendoff (1995) stated that each individual is able to store at least as many fixed expressions as single words in his or her mental lexicon, and they are regarded as a primary feature that a learner should acquire in order to achieve L1-like idiomaticity and fluency (Wray, 2002; Kecskes, 2007), although formulaic language tends to be less common in L2, as L2 learners generally find the acquisition of idiomatic expressions to be difficult (Ellis et al., 2008). Two underlying properties of idioms are considered to be crucial for their acquisition: cross-language overlap and transparency (Cucchiarini et al., 2022). The first refers to the extent to which an L2 idiom has an equivalent in L1 (e.g., the English *don't cry over spilt milk* versus the Italian equivalent *non piangere sul latte versato*), whereas the latter explains the degree of clarity of an idiom on the basis of the meaning of its individual words, e.g., the expression *cold turkey* is characterised by a high degree of opaqueness, since its meaning (i.e., "the period of extreme suffering

that comes immediately after a person has stopped taking a drug on which they depend")[16] cannot be understood or even guessed from its two constituting words, whereas an L2 learner can (more) easily associate the figurative and literal meanings of *the tip of an iceberg*. Indeed, idioms carry pieces of history and culture, and, while some cultures and languages share similar sayings and idiomatic expressions mainly because of historical reasons, this might not be the case for the relationship between English and the L1 of many L2 learners. In a study on the use and comprehension of English L2 idioms by advanced Spanish learners, Irujo (1986) reported that learners had no difficulties using identical idioms, while they found it hard to produce similar or different idioms. On the contrary, at a receptive level, they found identical or similar idioms easier to understand, whereas different idioms were still challenging to comprehension. Idiomatic expressions can constitute such a problem that they might be avoided by learners, as in the study conducted by Laufer (2000), in which Hebrew-speaking learners of English tended to avoid idioms which do not have counterparts in Hebrew and idioms which have partially formal similarity more than idioms which are completely similar and idioms which are formally different. In a comparative study between Malay and English, Charteris-Black (2002) distinguished between conceptual and linguistic similarities and differences and found that formally and linguistically similar figurative expressions were the easiest both for comprehension and production at the expense of idioms which had conceptual differences but equivalent linguistic form and idioms which were both formally and linguistically different. It has been demonstrated that developing a repertoire of formulaic expressions can help learners improve their oral proficiency, as shown in Boers et al. (2006). In this study involving 32 college students taking a 22-hour course, a group was made aware of idiomatic expressions, whereas the others were exposed to the traditional grammar-lexis approach. Two blind evaluators found that the experimental group was more proficient than the control group and that the number of idiomatic expressions correlated well with proficiency scores. In a similar study including Dutch learners of English and Spanish, Stengers et al. (2011) arrived at similar conclusions, however, finding a higher correlation between the count of formulaic expressions and proficiency ratings for English than for Spanish. This was probably due to numerous inflectional errors made by the learners of Spanish, which negatively influenced the evaluators' judgement on the use of formulaic expressions.

[16] `dictionary.cambridge.org/dictionary/english/cold-turkey`

Sociopragmatics

Often misunderstood and assimilated into pragmatic competence,[17] the sociolinguistic competence outlined in the CEFR also seems to have some overlaps with the concept of sociopragmatics (Sickinger & Schneider, 2014), as described by Leech (1983), whose definition was previously reported. The main focus of this construct is on learners' knowledge and ability of social norms, conventions and relationships, politeness and appropriateness, and reciprocal rights and obligations, and it has its roots in Leech's (1983) Politeness Principle, which, in short, relates to a speaker's will to speak appropriately in a given situation while still getting the intended message across. Leech (1983, p. 81) formulates the principle using a negative and a positive form. The first reads: "Minimize (other things being equal) the expression of impolite beliefs", and it concerns the minimisation of impoliteness in impolite illocutions, e.g., ordering. On the other hand, the latter focuses on ways to "maximize (other things being equal) the expression of polite beliefs", and it refers to the maximisation of polite illocutions, e.g., offering and thanking. The concept of linguistic politeness was further expanded by Brown & Levinson (1987), identifying three sociolinguistic aspects (power, distance, and absolute ranking of imposition) as major variables accounting for variation in speakers' linguistic choices. In other words, an interlocutor should evaluate a given context based on such three variables and subsequently select a strategy which will optimise the association between the linguistic features and elements employed to convey the intended meaning and the suitable level of politeness expected in that context. One of the commonly used instruments for assessing politeness is the discourse completion test (DCT), which generally consists of a prompt containing a situation description, a stimulus posed in the form of a question, and a blank space for test-takers to write their answers. On the one hand, DCTs are extremely practical, as the written form allows researchers to systematically change the three sociolinguistic variables mentioned above and administer tests to numerous participants at once. On the other hand, it is evident that DCTs cannot serve as accurate renditions of actual conversations since "there is no discourse-internal context, responses are not constructed under the time pressure of an online communicative situation, and respondents have been shown to write what they actually do say in reality" (Roever, 2014, p. 3).

The first and most influential work on the assessment of L2 pragmatics was Hudson et al.'s (1995) battery of six tests, which included several DCT items proposed in various forms (oral, written, and multiple-choice), a role-play, and two self-assessment sections. The tests included

[17] Already before the creation of CEFR, the sociolinguistic competence theorised in the communicative competence framework was criticised for its inconsistent definition (Zuskin, 1993).

Japanese-speaking learners of English and targeted the three most investigated speech acts, i.e., apology, request, and refusal. Subsequently, they were rated by L1 speaker evaluators on a five-point scale ranging from "very unsatisfactory" to "completely appropriate" on the basis of six criteria: ability to use the correct speech act; typical expressions; amount of speech used and information given; and levels of formality, directness, and politeness. After a try-out on 25 Japanese-speaking learners of English, Hudson (2001) reported a high correlation of the assigned scores between the oral DCT — which was found the most difficult — and the written DCT, whereas a low correlation was obtained between the DCT items and the role-play.

Hudson et al.'s (1995) framework was later adopted in other studies (Brown, 2001; Brown & Ahn, 2011), most of which appeared problematic when dealing with the multiple-choice DCTs, as they tended to have low reliabilities. This was probably due to the difficulty of producing incorrect options that were evidently unacceptable to all raters without being too rude or abnormal, thus biasing the tests.

In order to address this issue, Liu (2006) involved learners directly in the development of the scenarios used in the multiple-choice DCT items that were later administered to L2 learners and L1 speakers. Learners' responses were labelled as incorrect, while L1 speakers' responses were marked as correct. The test was then run on 200 Mandarin-speaking learners of English, and a high inter-rater agreement was obtained (around 0.90).

More recently, Sickinger & Schneider (2014) tried to compensate for the underspecifications and inaccuracies of the sociolinguistic and pragmatic competences outlined in the CEFR by building a profile for pragmatic competence called PRA.PRO, along the lines of the already mentioned profiles, i.e., the English Vocabulary Profile (Capel, 2015) for the domain of vocabulary and the English Grammar Profile (O'Keeffe & Mark, 2017) for grammatical competence.

1.2.3 Pragmatic competence

Coherence and cohesion

As has been said already, it is difficult to make a clear distinction between sociolinguistic and pragmatic competence, and even the developers of the CEFR (Council of Europe, 2020, p. 138) in the paragraph related to pragmatic competence admit that "[k]nowledge of interactional and transactional schemata relates also to sociocultural competence and is to some extent treated under "Sociolinguistic appropriateness" on the one hand and "General linguistic range" and "Vocabulary range" on the other". Similarly, the operationalisation of coherence and cohesion

also relies on measures of syntactic complexity (see Section 1.2.1).

There is no unanimous agreement on the definition of coherence since it has been connected to "continuity of sense" (de Beaugrande & Dressler, 1981, p. 115), referred to as "the relationships that link the ideas in a text to create meaning" (Lee, 2002, p. 135), and "the quality of the mental representation of the text that is created by the reader" (McNamara et al., 2010, p. 60). Especially this last definition gives a rather illustrative idea of the fuzziness of this aspect of proficiency and of its intangible and ephemeral nature. On the other hand, cohesion is a more objective aspect of discourse, and it refers to "the grammatical and/or lexical relationships between the different elements of a text" (Richard et al., 1985, p. 45) or, similarly, to "the mutual connection of components of surface text" (Bell, 1993, p. 165). However, despite this disambiguation, the two terms are often used interchangeably.

For coherence and cohesion, the CEFR also provides quite broad and generic definitions, although there is indeed a clear difference between C2 level ("Can create coherent and cohesive text making full and appropriate use of a variety of organisational patterns and a wide range of cohesive devices") and A1 level ("Can link words/signs or groups of words/signs with very basic linear connectors (e.g. "and" or "then")") (Council of Europe, 2020, p. 165).

Discourse competence of L2 learners has so far received scarce attention (Purpura, 2008; Kormos, 2011), especially with regard to speaking performance. In a foundational work by Brown et al. (2005) aiming at building the TOEFL iBT Speaking scoring rubrics, 20 answers to an independent speaking task and 20 answers to an integrated speaking task from speakers of five proficiency levels were analysed and annotated considering the number of clauses and T-units. They reported that both the average number of T-units and the average number of clauses per ten utterances varied across the five proficiency levels. Another interesting piece of analysis conducted in their study focused on discourse organisation, providing an ideal scheme for responses, including mandatory and optional elements (in square brackets):

- independent speaking task: [Introduction] → Opinion → Reasons for opinion → [Examples] → [Opinion 2] → [Reasons for opinion] → [Examples] → [Conclusion]
- integrated speaking task (Level 1): [Introduction] → Problem → Solution → Complication → Solution → [Conclusion]
- integrated speaking task (Level 2): Process → Outcome → [Evaluation][18]

[18]The steps outlined in Level 2 occur for each mandatory step in Level 1.

The results of the independent speaking task showed that proficient speakers tended to use complex structures and provide explanatory examples, whereas learners of lower proficiency levels often omit reasons for their opinions. Similarly, for the integrative speaking task, high-level test-takers employed elaborate structures with well-illustrated examples, logical connections, and clearly understandable statements, while less proficient speakers used less sophisticated structures.

Finally, the authors conducted a specific analysis on the use of logical connectives, assuming that proficient speakers would use more logical connectives than low-level test-takers, but, unfortunately, the amount of data was not sufficient to perform statistical comparisons between proficiency levels.

In a more recent study, Iwashita & Vasquez (2015) investigated cohesive (use of reference, ellipsis and substitution, lexical cohesion, conjunctions) and coherence (text generic structure and theme-rheme development) devices in 58 speech samples of IELTS Speaking Part 2. The results showed that higher-level test-takers used a wider range of conjunctions and employed referential expressions more accurately than the lower-level test-takers. Instead, other features did not have a dissimilar distribution across proficiency levels.

For other works on coherence and cohesion, the reader may refer to the paragraph on syntactic complexity in Section 1.2.1.

Fluency

As coherence, the concept of fluency has also been defined in various ways. In its broadest sense, fluency is commonly used as a synonym for oral proficiency (Lennon, 1990; Chambers, 1997), but its narrow definition designates a specific paramount aspect of proficiency. Far from being a monolith, fluency is, in fact, a multifaceted skill (Suzuki & Kormos, 2022). Fillmore (1979) provides a fourfold definition of L1 fluency:

- the ability to occupy time with talk;
- the ability to produce coherent, reasoned, and semantically dense speech;
- the ability to be pragmatic in one's speech depending on the context;
- the ability to talk with imagination and creativity, using metaphors, puns and jokes by means of sounds and meanings.

The higher-level features of L1 speech provided in this definition, such as creativity and use of metaphors, are absent from the conceptualisation of L2 fluency provided by Derwing et al. (2009, p. 534), in which fluency is defined as "temporal aspects of oral production that influence the degree of fluidity in speech (e.g., pauses, hesitation phenomena, speech rate)". Lennon (1990, p. 391) provides another definition considered from the standpoint of the interlocutor: "a listener's impression that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently". These definitions are complementary to each other since the study of fluency needs to investigate both the temporal elements of speech production and the listener's perception (Derwing et al., 2009).

In line with this perspective are the three senses of fluency elaborated by Segalowitz (2010, 2016): cognitive fluency, utterance fluency, and perceived fluency. The first refers to the speed and efficiency of the cognitive processes which are in charge of L2 speaking performance. The second seems to overlap with the definition by Derwing et al. (2009), and it designates the fluidity of the recognisable speech as distinguished by measurable temporal elements (e.g., filled and silent pauses, syllable rate, and hesitation rate). Finally, perceived fluency refers to the subjective perception and judgment of L2 speakers' fluency.

Turning specifically to utterance fluency, an essential study by Tavakoli & Skehan (2005) provided another dimensionality of fluency by focusing on three aspects: breakdown fluency, which is related to silences and pauses, e.g., number, duration and location of unfilled pauses, filled pauses, and overall amount of silence; speed fluency, which is connected to measures such as amount of speech, articulation rate, speech rate, time ratio and mean length of run; and repair fluency, which encompasses phenomena such as replacement, reformulation, repetition, and false starts.[19]

As regards breakdown fluency, higher-level learners' speech is generally characterised by a lower relative frequency of unfilled pauses and a lower ratio of pause time to speech time (Derwing et al., 2004; Kormos & Dénes, 2004; Rossiter, 2009; Bosker et al., 2014; Suzuki & Kormos, 2020; Suzuki et al., 2021). In fact, there is a potential issue that needs to be taken into account when linking pausing behaviour to linguistic proficiency, i.e., pauses could also be frequent in L1 speech. In order to distinguish L1 from L2 disfluencies, a key feature is not the total number of disfluencies but their distribution: multiple studies have found that L2 speakers tend to

[19] De Jong (2017) points out that this taxonomy misses an essential aspect of fluency, i.e., turn-taking fluency, which occurs when a speaker needs to interact with one or more interlocutors in dialogues. Turn-taking is also a component of pragmatic competence in the CEFR. However, despite its importance, it has received relatively scarce attention in research.

pause more often within clauses (Davies, 2003; Tavakoli, 2011; Kahng, 2014), Analysis of Speech (AS) units[20] (Skehan & Foster, 2007), or constituents (Riazantseva, 2001) than L1 speakers. Furthermore, a strong negative correlation between fluency ratings and frequency of pauses was found both in Rossiter (2009) and in Bosker et al. (2014).

With respect to speed fluency, speech rate (i.e., the number of syllables articulated per minute) and mean length of run (i.e., the average words or syllables per speech chunk contained within pauses) are the features that have been mostly investigated. In a study that has been cited several times in the previous paragraphs, Iwashita et al. (2008) reported that the best predictor of overall proficiency scores was speech rate. Similarly, Ginther et al. (2010) found strong correlations between holistic proficiency scores and various measures of fluency, such as articulation rate, speech rate, and mean length of run. Other studies (Lennon, 1990; Riggenbach, 1991; Cucchiarini et al., 2002) also found that speech rate is a good predictor of analytic fluency ratings.

The presence of self-corrections, false starts, reformulations, and repetitions may also be a common feature in L1 speech. Compared to the previous two sub-types of utterance fluency, repair fluency has been found to contribute only marginally to fluency ratings or overall proficiency (Kormos & Dénes, 2004; Brown et al., 2005; Suzuki et al., 2021).

It is also important to mention the impact of discourse markers on L2 proficiency. In the work by Huang et al. (2023), the authors investigated the developmental patterns of three discourse markers (i.e., *well*, *you know* and *like*) in the speech of learners ranging between A2 and C1. They found that the frequency of discourse markers (especially *well* and *you know*) increased as the fluency level increased almost reaching L1-like levels.

The multifaceted nature of fluency has led to different types of constructs in the context of language assessment. In the IELTS speaking rubrics, which are composed of four indicators (i.e., "fluency and coherence", "lexical resource", "grammatical range and accuracy", and "pronunciation"), fluency is combined with coherence[21]. The higher-level bands mention the terms "hesitation", "self-correction", and "repetition", whereas the lower-level ones refer to length of pauses.

Instead, the TOEFL has three main categories, i.e. "delivery", "language use", and "topic development" (Educational Testing Service, 2009). Fluency is mentioned both under the broader indicator of "delivery", which encompasses aspects of fluency and pronunciation, and in the de-

[20] "An AS-unit is a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster et al., 2000, p. 365).

[21] `ielts.org/-/media/pdfs/speaking-band-descriptors.ashx`

scriptors of "language use", which includes elements of accuracy, complexity, and fluency. For instance, for Score 4, the descriptor of "delivery" reads: "Generally well-paced flow (fluid expression)", but fluency is also mentioned under the descriptor of "language use" for Score 3, which reads: "Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures used. This may affect overall fluency, but it does not seriously interfere with the communication of the message".

Finally, the Pearson Test of English (PTE) Academic has fluency ratings that are obtained by comparing the test-taker's responses to the responses from an L1 speaker database, and the measures used in this comparison are all related to the duration of speech events, such as pauses between words, segments per articulation time, words per time, response latency, and a combination of these measures (De Jong, 2017). Therefore, unlike the IELTS and the TOEFL scales, the PTE Academic considers fluency as an individual and separate construct, and its descriptors are very specific and detailed. For example, at level 4 (on a scale from 0 to 5), the descriptor for oral fluency reads: "Speech has an acceptable rhythm with appropriate phrasing and word emphasis. There is no more than one hesitation, one repetition or a false start. There are no significant nonnative phonological simplifications" (Tavakoli et al., 2017, p. 8).

Chapter 2

Automatic speaking assessment

The first part of this chapter introduces the general characteristics of automatic speaking assessment, illustrating its typical pipeline and describing its advantages and limitations. A brief timeline traces the history of automatic assessment and also includes some fundamental studies on writing assessment. The same structure proposed in the previous chapter and based on the aspects which constitute the communicative competence model is employed again in this chapter and encompasses various studies on proficiency assessment conducted through automatic and semi-automatic approaches.

2.1 Introduction

In recent years, the growing number of L2 learners of English on a global scale (Howson, 2013) has led to an increasing demand for automated spoken language assessment systems for applications in the context of CALL.

One common misunderstanding about automated assessment is the assumption that machines should replicate the human process of assessment. Instead, computers are programmed to identify, extract, and quantify features in spoken and written productions. Such features are subsequently combined and weighted in a multidimensional space in order to predict a proficiency level or grade. The pipeline of a typical automatic system for speaking assessment is shown in Figure 2.1 and consists of three main components: an ASR, a feature extractor, and a grader.

The ASR module converts the audio signal of human speech into written format. Although the use of end-to-end systems is more and more common and has brought significant improvements,

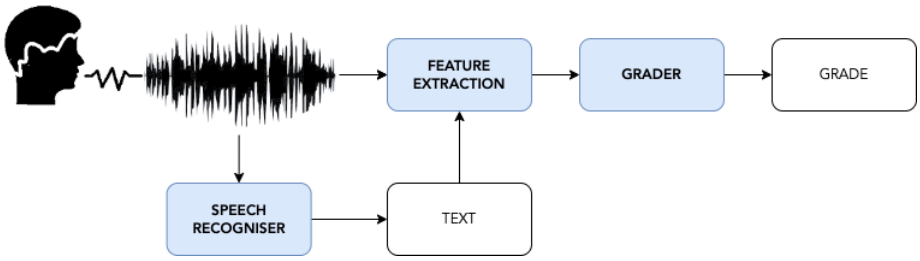


Figure 2.1: Pipeline of a typical automatic system for speaking assessment.

this module has typically consisted of an acoustic model and a language model. The first models the relationship between the audio signal and the phonemes or words, while the latter infers the word sequences that are likely to be uttered. A representative example of how a classic ASR system works is described in Lieberman et al. (2005, p. 1), in which two transcriptions are considered as possible outputs from an acoustic model: "wreck a nice beach you sing calm incense" and "recognise speech using common sense". The language model will then select the second option as more plausible based on its probability estimates.

In the feature extraction module, features that are relevant for the construct to be assessed are automatically extracted from both the speech signal and the transcriptions obtained through the ASR system. In this way, such features serve as proxies for human assessment criteria.

At the end of the pipeline, these features are used by the grader to make predictions of proficiency levels or grades.

2.1.1 Advantages, limitations, and challenges of automatic assessment

Figure 2.2 represents the various possibilities of delivery and scoring in L2 speaking assessment, with the models of quadrant 2 (examiner-delivered and examiner-scored tests) and quadrant 3 (computer-delivered and computer-scored) being the most commonly used.

Among the advantages of the first approach, first of all, there is the possibility of using a broad test construct, as a wide range of skills can be assessed by means of multiple elicitation systems, i.e., direct questions, role-play activities, collaborating with a partner to solve a problem, etc. Secondly, face-to-face tests are expected to mirror real-world interactions. Thirdly, they generally have very positive washback[1] since when the test-takers are preparing for the test, they improve their speaking skills which are crucial for everyday communication (Galaczi, 2010).

[1]The term "washback" refers to the influence of tests on curriculum design and teaching and learning practices (Alderson & Wall, 1993).

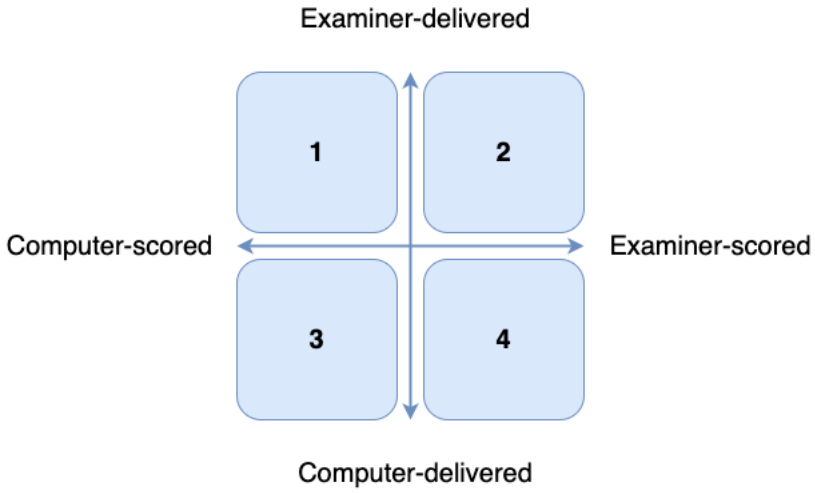


Figure 2.2: Delivery and scoring possibilities in L2 speaking assessment.

On the other hand, one of the compelling reasons for automatic assessment is the need to evaluate and provide feedback to increasing numbers of L2 learners and return results in a timely manner. Almost a century ago, Lundeberg (1929, p. 195) had already realised that individual oral assessment was "cumbersome and time-consuming". Returning to today, to provide a comparison, for the speaking parts of the IELTS and TOEFL iBT, test-takers receive score reports within 13 days, whereas the PTE Academic, which is fully automated, returns results within five business days (Isaacs, 2017b).

Secondly, compared to human graders, not only can automatic systems ensure greater speed, but they can do it at a lower cost since the recruitment and training of new human experts are expensive and can provide only a small increase in performance (Wang et al., 2018).

Finally, the use of automatic assessment methods can improve reliability, consistency, and objectivity of scoring and feedback since machines are not susceptible to rater effects and — more simply — to tiredness (Engelhard, 2002; Zhang, 2013; Van Moere & Downey, 2017). Moreover, machines have been found to be generally better at evaluating specific linguistic phenomena, whilst humans tend to focus on more global aspects of proficiency. For example, Enright & Quinlan (2010) suggested that, for writing, human raters might achieve higher results when assessing ideas, content, and organisation, whereas automatic systems might have better performances when evaluating microfeatures at the grammatical, syntactic, lexical, and discourse levels. Similarly, also for speaking, Loukina et al. (2015) found that human evaluators can have

difficulty in efficiently distinguishing particular phenomena, such as word-level pronunciation accuracy. In a study on the potential complementarity of human and automatic scoring, Davis & Papageorgiou (2021) found that composite grades calculated from various combinations of human and automatic analytic grades were equally or more reliable than human holistic grades.

However, beyond these crucial advantages, automatic assessment systems also have several issues and limitations. First, specifically automatic assessment of speaking is much more challenging than automatic assessment of writing since the first part of the pipeline of a typical automatic system for speaking assessment, i.e., the ASR module (see Figure 2.1), might have a certain word error rate (WER), i.e., a common metric employed to measure the performance of ASR systems and calculated as:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, and N is the number of words in the reference. While minimising the WER might be challenging when working with L1 speech, it is even harder with L2 learner speech since it might contain pronunciation errors, grammatical errors, code-switched words, and other typical fluency- and pronunciation-related features of L2 speech.

Another issue related to this first limitation is that ASR systems usually achieve higher performances on controlled tasks, such as reading aloud or shadowing, i.e., a repetition task in which candidates only hear the utterances they need to reproduce (Hamada, 2019), whereas they generally obtain lower results on spontaneous speech tasks. This is due to the possibility of using pattern matching or force-alignment approaches since the test-takers' utterances are known or, at least, highly predictable.

In a review article on fully automated systems for speaking assessment, Isaacs (2017b) mentions another limitation of such systems, i.e., their heavy dependence on time-based and frequency-based features, mainly connected to fluency and pronunciation, to the detriment of other formal and content-related elements of proficiency, such as grammatical accuracy and complexity, lexical richness and complexity, discourse organisation, and content development. However, it should be noted that this is not necessarily an intrinsic issue since, especially in recent years, automated systems can be used quite effectively also to assess higher-level aspects of proficiency. Instead, the narrow focus on fluency- and pronunciation-related features is often due to extrinsic reasons, such as the scarce availability of annotated data, design choices, or the

use of constructs that prioritise these aspects of proficiency. This extremely specific attention to such features is evident, for example, in the number of papers on pronunciation presented in the two sessions "Applications in Transcription, Education and Learning" at Interspeech 2022: 9 out of 15 papers focused on mispronunciation detection or issues related to pronunciation. While research on automatic scoring played a major role in revitalising studies on L2 pronunciation assessment (Isaacs, 2014, 2018), the scientific community working on automatic approaches should be careful not to fossilise and reduce proficiency to the speaker's capacity to pronounce words correctly.

Finally, there is an issue that concerns users' reception and understanding of automated assessment systems: with the gradually increasing application of methods and techniques of automatic assessment both for high-stakes language exams and for private practice, we have witnessed a certain scepticism and aversion to such systems in speaking (Neri et al., 2003) and writing (Attali, 2007), especially by educators and researchers, to the extent that in 2013 the site HumanReaders.Org[2] launched an online petition called "Professionals Against Machine Scoring of Student Essays in High-Stakes Assessment", which found the support of many distinguished scholars, including Noam Chomsky, and was mentioned in *The New York Times* and in many other newspapers (Stevenson, 2016). Although the campaign specifically attacked automatic essay scoring, it cannot be excluded that similar reactions will target automatic speaking assessment in the future. In particular, the criticism mostly targets technical and logistic problems, such as the financial and timely cost of automatic scoring systems and issues regarding their validity (Yang et al., 2002). However, while some of these issues do still pose significant challenges, some others have been efficiently dealt with, as we will explain in the next paragraphs.

2.1.2 Historical background

As should be clear at this point, an excursus into the history of automatic speaking assessment cannot avoid mentioning the evolution of its counterpart, i.e., automatic writing assessment, since the two fields have multiple aspects in common, especially in relation to the adaptation of techniques of textual analysis and information extraction, which are borrowed from automatic writing assessment and employed on ASR transcriptions.

The roots of the field of automated scoring of language proficiency can be traced back to the work of Page (1966, 1968) on automatic essay scoring. His Project Essay Grade was a system that evaluated writing skills based only on proxy traits: hand-written texts had to be manually entered

[2]The site is no longer active.

into a computer, and a scoring algorithm then quantified superficial linguistic features, such as essay length, average word length, count of punctuation, count of pronouns and prepositions, etc. Across the following decades, the field of automated scoring of writing has expanded and improved, and more significant studies have been conducted from the 1990s and early 2000s as computational techniques and software technology have increased their power (Landauer, 2003). The most widely known automated scoring systems for essays include the e-rater®, developed by Educational Testing Service (ETS) (Burstein, 2002; Attali & Burstein, 2006), IntelliMetric™by Vantage Learning (Rudner et al., 2006), and the Intelligent Essay Assessor™, built at Pearson Knowledge Technologies (Landauer et al., 2002).[3]

The 1990s also represented a turning point in the field of automated scoring of spoken proficiency, which initially focused on automatic evaluation of segmental pronunciation quality (Bernstein et al., 1990), basically through a comparison of the segments of the learner's speech signal and the segments developed from a database of L1 speech. Early approaches to automatic assessment of speech consisted of simple speaking tasks (e.g., reading a word or a sentence out loud), mainly due to limitations of the automatic speech recognition component of the scoring system. Pronunciation assessment is also the focus of Cucchiarini et al. (1997), in which various features such as acoustic scores from a Hidden Markov Model (HMM), the total duration of speech with and without pauses, mean segment duration, and speech rate were employed to score Dutch pronunciation skills. Franco et al. (2000) used similar features for assessing L2 English pronunciation and included an ASR system specifically adapted to L2 speech in order to reduce the WER. All these three studies evaluated their assessment systems on read-aloud speech.

Conversely, in the 2000s, ETS introduced SpeechRater, which could score spontaneous speech in addition to read speech (Xi et al., 2008; Zechner et al., 2009; Higgins et al., 2011) based on features related to pronunciation, fluency, vocabulary, and grammar. SpeechRater is one of the best-known oral proficiency commercial test engines and is still used in the TOEFL speaking test. Another well-known commercial test engine is Versant, originally called SET-10 or PhonePass (Townshend et al., 1998; Bernstein & Cheng, 2007), built by Ordinate Corporation and now employed in the Pearson PTE Academic. It is interesting to note that the two test engines are based on different constructs. For Pearson, speaking ability is considered as a "real-time activity that requires planning, formulating, articulating, and monitoring" and, consequently, test scores should represent test-takers' ability to use "core language component process

[3]A detailed account of these systems goes beyond the scope of this thesis but can be found in Warschauer & Ware (2006). A more recent survey on the state of the art of automatic essay grading can be found in Ke & Ng (2019).

in real time by quantifying the ease with which the speaker can access and retrieve lexical items, build phrases and clause structures, and articulate responses, without conscious attention to the linguistic code" (Downey et al., 2008, p. 161-162). Instead, ETS defines speaking ability as "the use of oral language to interact directly and immediately with others" (Butler et al., 2000, p. 2). Drawing on the contents of Section 1.1, one can observe that the first opted for a psycholinguistic construct, while the latter chose to adopt a communicative construct.[4]

In recent years, deep neural network (DNN) approaches have brought significant improvements in the field of automatic assessment for both writing (Alikaniotis et al., 2016) and speaking (Qian et al., 2012; Evanini et al., 2018), such that end-to-end neural-based techniques outperformed SpeechRater (Chen et al., 2018). In particular, the application of DNNs on ASR systems has been shown to be highly effective, to the extent that some systems have obtained results on L1 transcriptions which are comparable or equal to those achieved by human transcribers (Saon et al., 2017; Xiong et al., 2017), although transcribing L2 speech is still problematic, in some cases also due to the low level of human-to-human agreement (Qian, Lange, & Evanini, 2019). Another crucial advancement was brought by the application of word embedding techniques, such as word2vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019), on automatic assessment tasks (Qian et al., 2019; Raina et al., 2020; Wang et al., 2021). More recently, the use of speech embeddings, such as wav2vec 2.0 (Baevski et al., 2020; Hsu, Sriram, et al., 2021) and HuBERT (Hsu, Bolte, et al., 2021), has been investigated for mispronunciation detection and diagnosis (Peng et al., 2021; Wu et al., 2021; Xu et al., 2021) and automatic pronunciation assessment (Kim et al., 2022) and is explored for the tasks of analytic and holistic proficiency assessment in this thesis.

In the next paragraphs, we will illustrate the state of the art of automatic speaking assessment following the competence model outlined in Chapter 1.

2.2 Linguistic competence

2.2.1 Grammar

The exploration of grammatical accuracy is a rather new area of study since, in its early days, automatic speaking assessment focused on restricted speech and did not require to assess grammatical proficiency. Despite the inclusion of spontaneous speech tasks, the investigation of grammar in the field of automatic speaking assessment has been the focus of relatively few studies,

[4]See Litman et al. (2018) for a detailed analysis and comparison of the two systems.

mainly for grammatical error detection (GED), starting from the isolated and pioneering work by Izumi et al. (2003) on manual transcriptions of Japanese learners of English to several recent fully automated approaches (Knill et al., 2019; Lu, Gales, Knill, Manakul, Wang, & Wang, 2019; Caines et al., 2020), and grammatical error correction (GEC) (Lu et al., 2020), but only to a lesser extent in relation to 'pure' scoring and assessment.

Instead, it has received more attention in the field of automatic essay scoring. Grammatical errors are one of the features employed in Yannakoudakis et al. (2011) along with lexical, part-of-speech (POS) and syntactic features for automatically assessing L2 English exam scripts, and they were found to significantly improve the correlation between true scores and predicted ones. Gamon et al. (2013) used Leacock & Chodorow's (2002) findings on the influence of grammatical errors on TOEFL scores for automatic essay scoring and feedback. In particular, Leacock & Chodorow (2002) reported that the variety of errors, rather than the overall error count, influences the score. In addition to this, they found that subject-verb agreement errors, wrong formations of modal verbs, and determiner-noun agreement errors predicted lower scores best. Similarly, grammatical errors are one of the features explored in the work of Vajjala (2018), in which spelling and grammar errors are automatically extracted by LanguageTool.[5] In this case, the feature related to grammatical error rate was found to have little impact on the classification performance. Similar experiments were conducted again by Vajjala & Rama (2018) with German, Czech, and Italian, including errors as a feature. This work was reproduced by Caines & Buttery (2020), who applied such experiments also to English and Spanish written corpora. Another research conducted on L2 English written examinations found that grammatical error detection highly influences automatic essay scoring (Cummins & Rei, 2018). Recently, the work described by Ballier et al. (2019) has investigated the possibility of predicting CEFR proficiency levels based on manually annotated errors of essays by L1 French and Spanish learners of English, but their study did not employ deep learning techniques. However, they identified that certain types of errors, such as punctuation, spelling, and verb tense errors, are characteristic of specific CEFR proficiency levels.

Written GEC has become an established area of study: four shared tasks have already been organised in the last 15 years, i.e., the HOO 2011 Pilot Shared Task (Dale & Kilgarriff, 2011), the CoNLL-2013 Shared Task (Ng et al., 2013), the CoNLL-2014 Shared Task (Ng et al., 2014), and the BEA-2019 Shared Task (Bryant et al., 2019); the ERRor ANnotation Toolkit (ERRANT) (Bryant et al., 2017), a tool for automatically extracting grammatical error edits from

[5] `languagetool.org`

parallel original and corrected sentences, has become quite commonly used in this field; and at least two metrics have been investigated and employed for evaluating the performance of GEC systems, namely MaxMatch (M^2) (Dahlmeier & Ng, 2012) and General Language Evaluation Understanding (GLEU) (Napoles et al., 2015). Table 2.1 shows an example of written GEC. Most current researchers have investigated models based on neural machine translation (NMT). In particular, Yuan & Briscoe (2016) were the first to apply this approach to GEC using a recurrent neural network (RNN). For two recent surveys on GEC, the reader can refer to Wang et al. (2021) and Bryant et al. (2022).

In (Original)	He see the thief is catched by policeman the last night.
Out (Corrected)	He saw the thief caught by a policeman last night.

Table 2.1: Example of written GEC.

Instead, spoken GEC is a relatively new area of research, mainly due to the scarce availability of specifically designed and annotated data. As can be seen in the example reported in Table 2.2, it is also more challenging than written GEC in that L2 spoken grammar contains disfluencies and errors, which in some cases might differ from the ones made by L2 learners in their written productions. In several studies proposed by the researchers of the Automated Language Teaching and Assessment (ALTA) Institute at the Department of Engineering of the University of Cambridge, Long Short Term Memory (LSTM)-based models trained on a large amount of written L2 learner data were investigated for GEC, GED, and disfluency detection (DD) (Knill et al., 2019; Lu, Gales, Knill, Manakul, & Wang, 2019; Lu et al., 2020). The results indicated that the use of DD improved the performance of both GEC and GED.

In (Original)	uhm he see the the thief is catched by policeman the la- last night
Out (Corrected)	he saw the thief caught by a policeman last night

Table 2.2: Example of spoken GEC.

In the context of automatic speaking assessment, the investigation of features related to grammatical accuracy has received scant attention. In the work by Supnithi et al. (2003), a classification model is used to predict 9 proficiency levels of manual transcriptions of Japanese learners of English using a feature set that modelled fluency, vocabulary, sociolinguistic appropriateness, and grammatical accuracy. Hasan & Khaing (2008) also conducted experiments of proficiency classification on the same corpus by only focusing on features related to grammatical accuracy and fluency.

On the other hand, grammatical complexity has been subjected to a more thorough investigation. Biber et al. (2016) and Lu (2017) employed some automatic techniques based on natural language processing (NLP), e.g., POS taggers and syntactic parsers, to extract features related to grammatical complexity, but neither of these studies used a fully automated process for feature extraction. In Chen & Zechner (2011) and Chen & Yoon (2012), features related to syntactic complexity were extracted by means of deep syntactic analysis and investigated for an automatic speaking assessment system. Some examples of the syntactic features employed in Chen & Zechner (2011) are mean length of sentences, mean length of T-units, mean number of noun phrases per sentence, mean number of passives per sentence, and mean number of dependent infinitives per T-unit. Both studies found that errors at the ASR module stage negatively influenced the performance of the scorer, as the correlation between syntactic complexity features and oral proficiency scores was lower when the ASR output text was used as opposed to manual transcriptions. Instead of using deep syntactic analysis, Yoon & Bhat (2012) and Bhat et al. (2014) proposed a method which aimed at capturing variations in the distribution of morpho-syntactic features based on POS tags across proficiency levels. They calculated the vector similarity between the POS sequences of the transcription of a given spoken response and the responses of a learners' corpus annotated with proficiency levels. They assumed that lower-level test-takers tend to use simple grammatical forms, whilst proficient speakers have a sophisticated repertoire of grammatical expressions and that such differences can be reflected in the distribution of POS tags. Both these studies demonstrated the greater robustness of features based on POS tags against ASR errors compared to features based on deep syntactic analysis. However, both types of features are included in ETS's SpeechRater (Yoon et al., 2019).

2.2.2 Pronunciation

The main focus of automatic speaking assessment was pronunciation in its early days "with a flurry of activities in late 90's to early 2000" (Witt, 2012, p. 1), as has been mentioned in Section 2.1.2. The spoken responses of L2 learners were typically compared to responses from a corpus of L1 speakers, and they only included read-aloud speech (Bernstein et al., 1990; Neumeyer et al., 1996; Franco et al., 1997). In these studies, various measures have been employed for automatic scoring: HMM log-likelihood scores, timing scores, segment duration scores, phone classification error scores, and phone log-posterior probability scores. In the same period, the Goodness of Pronunciation (GOP) algorithm was introduced and became one of the most popular for pronunciation assessment (Witt, 1999; Witt & Young, 2000). GOP is based on the posterior probability

that a given phone corresponds to the phoneme that should have been uttered according to its canonical pronunciation, thus:

$$GOP = \frac{1}{M} \sum_{n}^{M} \ln p(q_n|o_n)$$

where q_n is the n-th phone in a given speech segment, o_n is the corresponding speech segment, and M is the overall number of phones contained in the speech segment. If we consider again the two principles outlined in Section 1.2.1, i.e., the "nativeness principle" and the "intelligibility principle", we cannot help but notice that these initial studies related to a construct of pronunciation proficiency conceptualised as similarity to L1 speech, which has been gradually falling out of favour in the language teaching and assessment community (Isaacs, 2014; Levis, 2020), as mentioned in Section 1.2.1. However, despite the increasingly broad orientation towards intelligibility, approaches involving alignments and comparisons between learners' responses and L1 realisations have also been investigated in recent times in the field of automatic assessment (Kamimura & Takano, 2019; Karhila et al., 2019), and several resources and applications in Computer-Assisted Pronunciation Training (CAPT) aim to help learners reach an L1-like level of pronunciation rather than to enhance intelligibility and comprehensibility (Pennington & Rogerson-Revell, 2019b). For example, ELSA Speak,[6] a well-known language learning application, only focuses on segmental aspects of pronunciation and is explicitly designed with the final goal of reducing accented speech (Becker & Edalatishams, 2019). On the other hand, Duolingo,[7] arguably the most famous language learning application at the moment, does not seem to penalise accented speech, although it does not explicitly set intelligibility and comprehensibility as primary targets over accent reduction (Hirschi, 2020).

Apart from a construct-related issue, some other limitations of the operationalisation of the "nativeness principle" are the need for a large amount of both L1 and L2 data, the tendency to depend on text, and sensitivity to the vocal characteristics of the L1 speakers, which may penalise fluent and intelligible speakers with accents that are not featured in the L1 speaker training data.

While the early findings in automatic pronunciation assessment contributed to boosting research interest also in human pronunciation assessment (Isaacs, 2014), in the early 2000s, research activities on automatic assessment started slowing down to regain momentum at the end of the decade. In particular, in 2007, a special interest group called Speech and Language Technology for Education (SLaTE) was founded within the International Speech Communication Associa-

[6] elsaspeak.com/
[7] duolingo.com/

tion (ISCA) (Witt, 2012). Around the same years, a number of researchers proposed various approaches to pronunciation assessment without comparisons to models trained on L1 speech for segmental (Minematsu et al., 2006; Van Doremalen et al., 2009) and suprasegmental pronunciation (Hönig et al., 2010), although their studies still used L1 speech as a reference. A somewhat ambiguous position fluctuating between 'nativeness' and 'intelligibility' has also characterised commercial test engines. In this regard, Isaacs (2017b) noticed that, despite their claim to assess intelligibility, in fact, many automated speaking tests pay considerable attention to conformity to L1 speech norms and pronunciation accuracy.

An interesting preliminary approach to intelligibility assessment is based on the identification of the most critical pronunciation errors of L2 learners. In this regard, in their study on Japanese learners of English, Raux & Kawahara (2002) used a probabilistic algorithm to connect intelligibility to error rates with particular attention to errors that affect intelligibility most (ten types of insertion, deletion, and substitution errors). Along similar lines, in a study on automatic pronunciation assessment of L2 learners of Dutch, Cucchiarini et al. (2007) investigated a technique which could prevent the ASR system from indiscriminately targeting all types of learner errors. They only selected relevant types of errors according to five criteria, i.e., errors should be:

- common across speakers from various L1 backgrounds;
- perceptually salient;
- potentially representing an obstacle to communication;
- frequent;
- persistent over time.

Another approach to pronunciation assessment which tried to overcome the dichotomies 'correct' versus 'mispronounced' and 'native' versus 'non-native' is the one introduced by Wei et al. (2009). In their study, they proposed to build several parallel acoustic models (called Pronunciation Space Models) covering the entire pronunciation space of a phone in order to represent pronunciation variations across different proficiency levels.

The operationalisation and automatisation of the "intelligibility principle" certainly avoids the problems related to L1 speech data but raises questions about how intelligibility should be measured and who should measure it.[8] This is a clear example of how some flaws and open

[8] See note 10 in Chapter 1.

questions, as they remain unresolved in the theoretical construct of intelligibility, inevitably emerge also in the context of automatic assessment.

If we resume the history of automatic pronunciation assessment, we can see that the 2010s represented a crucial step forward, especially due to the advent of DNNs. The first study to use DNN acoustic models for improving mispronunciation detection was conducted by Qian et al. (2012), showing a significant improvement on their Gaussian mixture model (GMM)-HMM baseline. Recently, the use of DNNs for automatic pronunciation assessment has been investigated in several papers (Yu et al., 2015; Chen et al., 2018; Lin & Wang, 2021). In the study by Yu et al. (2015), high-level abstractions were learned from time-sequence features using a bidirectional LSTM model and were combined with time-aggregated features extracted with SpeechRater in a Multilayer Perceptron (MLP). Interestingly, they used a combination of time-aggregated features which include different aspects of proficiency (i.e., pronunciation, fluency, intonation, rhythm, vocabulary use, and grammar) to predict intelligibility scores annotated by experts ranging from 1 (largely unintelligible) to 4 (highly intelligible). However, the features related to pronunciation were calculated using a corpus of L1 speech. Therefore, there seems to be an issue related to construct definition, as it is not clear whether the authors meant to target "nativeness" or "intelligibility". In Chen et al. (2018), a baseline fed with the same set of SpeechRater features is compared to a bidirectional LSTM model with an attention mechanism to predict overall proficiency with both lexical and acoustic cues. Their proposed approach outperformed the hand-crafted features baseline, but the authors acknowledged the need to ensure the explainability of DNN-based approaches, which is often problematic but constitutes a major aspect in a field such as L2 assessment, especially for high-stakes language exams. Similarly, in an experiment on pronunciation assessment of Chinese learners of English on a scale from 1 (hardly understandable) to 5 (L1-like), Lin & Wang (2021) used a model consisting of two encoders, one for text and the other for audio, combined based on an attention mechanism.

The application of DNNs to automatic pronunciation assessment was also investigated for potentially suitable approaches for assessing intelligibility. Drawing on the seminal works by Asakawa et al. (2005) and Minematsu et al. (2006), Kyriakopoulos et al. (2018) proposed an attention-based method using phone distance features, whereby each phone is defined in relation to the realisation of each of the others, thus representing pronunciation in a manner that should be compact and independent of speaker attributes and comparisons with L1 speech. The same authors proposed a similar approach to assess rhythm proficiency based on durations of phones and silences, grouped into consonant and inter-consonant intervals (Kyriakopoulos et al., 2019).

Both these assessment systems are used in Study 3 (see Section 5.1).

Recent studies have demonstrated that self-supervised learning (SSL) effectively works in different downstream tasks of speech processing applications, such as ASR, emotion recognition, keyword spotting, intent classification, speaker identification, and speaker diarisation (Baevski et al., 2020; S.-W Yang et al., 2021). In these works, contextual representations were applied by means of pre-trained models. In particular, they demonstrated that these models are able to capture a wide range of speech-related features and linguistic information, such as audio, fluency, suprasegmental pronunciation, and semantic and syntactic text-based features for L1, L2, read and spontaneous speech (Singla et al., 2022). In CALL, SSL has been used for mispronunciation detection and diagnosis (Wu et al., 2021; Xu et al., 2021; Peng et al., 2021) and automatic pronunciation assessment (Kim et al., 2022). Specifically, in this last study, various pre-trained and fine-tuned versions of wav2vec 2.0 and HuBERT, as well as baselines based on hand-crafted features, were used for the task of predicting human scores of L2 read speech from two learner corpora. One dataset contained proficiency scores about prosody and fluency, whereas the other was annotated with five pronunciation measures concerning pronunciation, segmental accuracy, pauses, stress, and intonation. Their best-performing model on both datasets was a grader based on HuBERT Large fine-tuned on L2 speech. It appeared that fine-tuning brought interesting improvements to the HuBERT-based models, whereas the ones based on wav2vec 2.0 showed only moderate performance gains when fine-tuned.

2.2.3 Vocabulary

Like grammar, vocabulary was also not a subject of research in the early stages of automatic speaking assessment since the first systems only considered read-aloud speech. In the study by Supnithi et al. (2003) mentioned earlier, a classifier was fed with features related to grammatical accuracy, fluency, sociolinguistic appropriateness, and vocabulary to predict 9 proficiency levels of a corpus consisting of manual transcriptions of Japanese learners of English. They employed 8 types of frequency-based vocabulary features, of which two were most effective, namely the frequency of all words produced by the test-taker and the frequency of a list of words annotated with a level based on a vocabulary profile. Conversely, in Zechner et al. (2009), measures such as TTR and the number of word types were employed for SpeechRater, and a low correlation was found with oral proficiency scores. Similarly to the first study cited in this section, Crossley et al. (2011) also conducted a study focused on lexical proficiency using a dataset consisting of manual transcriptions of 29 learners of English from various L1 backgrounds and a section

of transcriptions extracted from a corpus of L1 speakers. They investigated the correlations of various lexical indices extracted with Coh-Metrix (Graesser et al., 2004), a tool that we will briefly describe in Section 2.4.1, with lexical proficiency ratings ranging from 1 (low) to 5 (high) annotated by human experts, and found that lexical diversity explained over 45% variance of the human scores, followed by word imagability, word familiarity, and hypernymy.[9]

In addition to measures of lexical diversity, lexical sophistication features have also been explored in the field of automatic speaking assessment. Yoon et al. (2012) investigated the use of vocabulary profile to extract features of lexical sophistication for proficiency assessment of spontaneous speech and found interesting correlations with oral proficiency scores, although they reported that both response length and task type strongly affected the correlations. Kyle & Crossley (2015) introduced the Tool for Automatic Analysis of Lexical Sophistication (TAALES), which computes 135 lexical indices. In their study, they found that 5 measures of lexical complexity accounted for more than 50% of the variance in the human ratings of the spoken dataset (manual transcriptions) considered in their study. Their tool has been recently updated (Kyle et al., 2018) and has been used in several papers investigating the connection between vocabulary and spoken proficiency (Uchihara & Clenton, 2020; Tavakoli & Uchihara, 2020; Saito, 2020). Although not strictly related to vocabulary, it is important to mention the work by Qian, Lange, Evanini, Pugh, et al. (2019), in which a monologic task and a simulated dialogic task are assessed employing neural network approaches. Specifically, three dimensions of proficiency, i.e., delivery, language use, and content, are scored using three attention-based bidirectional LSTM RNN systems. These aspects are investigated both individually and holistically, i.e., after fusing their respective subscores into an overall score. The authors show that their approaches outperform the conventional approaches to spoken language assessment. Furthermore, they demonstrate that the correlations of the automatically predicted scores with the scores assigned by human experts are higher than human-to-human correlations for both the monologic task and the dialogic task.

As has been mentioned in Section 2.1.2, a major paradigm shift occurred in the world of NLP — and consequently in the field of automatic speaking assessment — when word embedding techniques such as word2vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019) were introduced. As these models are pre-trained on a large quantity of textual data, they provide

[9]These measures are based on human annotations derived from other databases. In particular, word imagability and word familiarity are based on human word judgements extracted from the MRC Psycholinguistic Database (Wilson, 1988). The first refers to the degree of evocativeness of a word: the word *dog* is highly imagable because it evokes images easily, whilst the word *nonetheless* can hardly produce a mental image. The second refers to the degree of familiarity, e.g., the word *while* has a mean familiarity score of 5.43, whereas the word *eat* has a score of 6.71. The hypernymy index is a measure of word specificity and is based on the WordNet hypernymy values (Fellbaum, 1998).

extremely powerful representations of syntactic and semantic aspects of words and sentences. In 2019, the Spoken CALL Shared Task (Baur et al., 2019) addressed the automatic assessment of sentences produced by young Swiss German learners of English using spoken or written (i.e., transcriptions) input modalities. In the latter case, some approaches based on word embedding techniques were also investigated. For their neural-based scoring system, Qian et al. (2019) used sentence similarities between ASR transcriptions and the corresponding answers included in a reference grammar, i.e., a list of correct responses provided by the challenge organisers. Their work also reported the performances of different types of word embeddings, i.e., word2vec and doc2vec (Le & Mikolov, 2014). The approach proposed by Sokhatskyi et al. (2019) used a scoring system that is based on a neural network fed with word vectors obtained through the concatenation of BERT and a neural-based language model trained on the datasets provided for the shared task. The authors also investigated other types of word embeddings, namely word2vec, doc2vec, ELMO (Peters et al., 2018), and Universal Sentence Encoder (Cer et al., 2018). The work by Raina et al. (2020) investigated the sensitivity of spoken language assessment systems to a universal black-box attack on the ASR transcriptions. In the first part of their experiments, they reported the results of four DNN-based graders for the task of predicting the proficiency levels of a spontaneous speaking section of a Business English test. The graders were the two systems described in Kyriakopoulos et al. (2018) and Kyriakopoulos et al. (2019), which we briefly described in Section 2.2.2, a feature-based Gaussian Process-based grader (Wang et al., 2018), and a BERT-based grader. The BERT-based grader showed the best results across almost all metrics. This grader is also used in Study 3 (see Section 5.1) and Study 5 (see Section 6.2). In a paper by Craighead et al. (2020), an LSTM grader and a BERT-based grader were compared for the task of predicting the proficiency scores of spoken responses collected from candidates taking Cambridge Assessment's BULATS exams.[10] Both systems were investigated for multi-task learning with language modelling, L1 identification, part-of-speech tagging, and universal dependency tagging as auxiliary objectives. Their best-performing system was the BERT-based grader with L1 prediction as an auxiliary task. The study by Wang et al. (2021) also compared two transformer-based models, i.e., BERT and XLNet (Yang et al., 2019), to an attention-based LSTM-RNN as well as a Support Vector Regressor (SVR) fed with hand-crafted content-related features for the assessment of spoken proficiency with a particular focus on content relevance using both manual and ASR transcriptions. For this reason, they investigated prompt-aware (i.e., incorporating prompt text) and prompt-unaware models (i.e., consisting of the response only),

[10] `https://www.cambridgeenglish.org/exams-and-tests/bulats/`

and they found that their best-performing model was the prompt-aware BERT-based grader.

2.3 Sociolinguistic competence

Sociolinguistic competence is arguably one of the least investigated facets of L2 proficiency in the field of automatic assessment, mainly due to longstanding issues in the theoretical construct (Zuskin, 1993). The fact "that the role of pragmatic competence (including the sociolinguistic component) in the CEFR is still critically underspecified" (Sickinger & Schneider, 2014, p. 118) has obvious repercussions on the operationalisation and automatisation of this competence. Secondly, due to its specific nature, it lends itself more to qualitative than quantitative analysis. In a forward-looking paper we have already mentioned, Supnithi et al. (2003) used various features across fluency, vocabulary, grammatical accuracy, and sociolinguistic appropriateness to predict the proficiency levels on a corpus of manual transcriptions of Japanese learners of English. They essentially defined sociolinguistic appropriateness in relation to 'success of communication', for which the reciprocal interaction between interviewer and interviewee is crucial, identifying two features: the frequency of words uttered by the interviewer and the turn count between interviewer and interviewee. Although the construct of sociolinguistic appropriateness was reduced to only one of multiple aspects, it is remarkable that this competence was considered in such an early study.

Other studies explored sociolinguistic competence, specifically focusing on formulaic and idiomatic expressions. Despite its focus on L2 writing proficiency, we find the study proposed by Bestgen (2017) noteworthy. The author automatically extracted multiple measures related to formulaicity as well as lexical diversity and sophistication and reported that one of the three formulaic measures was the most correlated with the text quality scores of two L2 corpora. Other studies we have mentioned in Section 2.2.3 have employed semi-automatic techniques (e.g., TAALES) to extract lexical and idiomatic measures from manual transcriptions of L2 learners. In particular, Saito (2020) analysed the manual transcriptions of 85 Japanese learners of English assessed for global comprehensibility and lexical appropriateness by targeting the use of collocations. The study reported strong correlations between both scores and low-frequency combinations, including abstract, infrequent, and complex words. The study by Tavakoli & Uchihara (2020) explored the connection between fluency and the use of multi-word sequences in a dataset of manual transcriptions of 56 learners of English across four proficiency levels ranging from low B1 to C1. They found a) a positive correlation between high-frequency n-grams

and articulation rate, b) a negative correlation between n-gram proportion and frequency of mid-clause pauses, and c) that n-gram associational strength had a positive correlation with the frequency of pauses located at the end of clauses and a negative correlation with repair frequency. Interestingly, they also found that lower-level learners used n-grams verbatim, whereas proficient speakers used them appropriately in a variety of forms. Similarly, Uchihara et al. (2022) investigated the relationship between collocation knowledge and oral proficiency, which was measured by means of human ratings and objective measures of fluency and lexical richness. The study found that the use of low-frequency collocations was correlated positively with speech rate and negatively with the number of silent pauses, and, in general, speakers who used them were perceived as more fluent. Instead, the use of strongly associated collocations was positively correlated with the number of sophisticated lexical items and the perception of higher lexical proficiency. This study was also conducted on manual transcriptions using automatic tools to extract fluency-related and lexical measures.

Finally, although related to L2 written productions, it is important to mention the VUA and TOEFL Metaphor Detection Shared Task (Leong et al., 2020). A correlation between proficiency and the number of metaphors was found in the L2 corpus considered in the challenge. Interestingly, more than half of the participating systems leveraged BERT for metaphor identification. For this task, the usefulness and effectiveness of a BERT-based approach had already been shown by Mao et al. (2019).

2.4 Pragmatic competence

2.4.1 Coherence and cohesion

Approaches to automatically assess the discourse coherence of textual data have been extensively investigated in the context of applications such as document paraphrasing and summarisation, text readability assessment, and natural language generation. Latent Semantic Analysis (LSA) (Landauer et al., 1998) was the technique proposed by Foltz et al. (1998) to measure textual coherence by computing the semantic relatedness between adjacent segments of text. LSA, as well as other components (e.g., POS taggers, syntactic parsers, lexicons, etc.), is also at the core of Coh-Metrix (Graesser et al., 2004), a tool which analyses textual data across over 200 measures of cohesion and other aspects of language. As mentioned in Section 2.2.3, Coh-Metrix has been employed in the context of L2 speaking proficiency assessment (Crossley et al., 2011).

However, not many works have investigated discourse-level features in this specific context.

On the other hand, coherence and cohesion have been investigated in several studies on automatic essay grading, which we find essential to mention in our review, as we have already done for some other aspects of proficiency in the previous sections. In the study by Higgins et al. (2004), the coherence of L1 student essays is evaluated by computing the semantic relatedness between essay questions and discourse elements of the essays using LSA and another vector-based approach for semantic representation called Random Indexing (Sahlgren, 2005). The coherence scores of essays were also the target of an automatic essay grader presented in Burstein et al. (2010), which combined features related to grammatical errors and word usage with the features extracted with the entity-based coherence algorithm introduced in Barzilay & Lapata (2008). Specifically, in the context of automatic assessment of L2 writing proficiency, Yannakoudakis & Briscoe (2012) presented a systematic analysis of various approaches for assessing coherence and reported that the most predictive features were word length, Incremental Semantic Analysis (Baroni et al., 2007) (i.e. a fully-incremental variation of Random Indexing), local histograms of words obtained from the locally-weighted bag-of-words framework (Lebanon et al., 2007), and a POS-based adaptation of the IBM model 1 (Soricut & Marcu, 2006). The work by Somasundaran et al. (2014) drew on the concept of lexical chains (Morris & Hirst, 1991), i.e., sequences of semantic-related ordered words characterised by synonymy, similarity, and repetition, for evaluating the quality of discourse coherence in essays written by L1 and L2 speakers of English. They indicated that the features based on lexical chaining outperformed previous approaches to the same task and that the best-performing approach was obtained through the combination of these features with other discourse features, such as features representing errors in mechanics, grammar, and word usage, and features obtained from a discourse parser based on Rhetorical Structure Theory (RST) (Mann & Thompson, 1988).

Going back to the field of automatic speaking assessment, in the work by Hassanali et al. (2012), a corpus of child language samples of story retells was annotated for coherence, narrative structure, and narrative quality features by human experts for the task of automatic diagnosis of children with language impairment. They used the manually annotated narrative features and coherence-related features extracted with Coh-Metrix to predict the human-annotated scores. Wang et al. (2013) investigated the use of coherence features for automatically assessing L2 spontaneous speech and found that the addition of such features improved the performance of their automatic assessment system by 10% for the prediction of holistic scores. The prediction of the analytic coherence scores also achieved significant results. RST also inspired the authors

of Wang et al. (2017) for the annotation of a corpus of 600 spoken responses drawn from the TOEFL iBT. Several features were extracted from the RST annotations and were found to have interesting correlations with both the holistic scores and the analytic scores related to discourse coherence. This work was expanded in Wang et al. (2019). The authors increased the number of annotated responses from 600 to 1440. All these manually annotated data were used to train RST parsers, which were subsequently used for inference on ASR transcriptions in order to generate features related to discourse coherence. These features could predict holistic scores with low accuracy (55.9%). Even though their impact on performance was found to be limited, they were finally combined with other types of features in order to enhance the validity of SpeechRater. Finally, it is worth mentioning the Tool for the Automatic Analysis of Cohesion (TAACO) 2.0 (Crossley et al., 2019), a text analysis tool which provides hundreds of indices related to cohesion. In particular, it incorporates semantic similarity features based on LSA, Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and word2vec. In addition to investigating the impact of this tool on writing proficiency, the authors also explored the performance of its features on manual transcriptions of responses obtained from the TOEFL iBT. They reported that the percentage of keywords present in both the prompt and response and the word2vec-based similarity between the prompt and the response were good predictors of speaking proficiency.

2.4.2 Fluency

Unlike coherence and cohesion, fluency has been widely and variously explored in the context of automatic speaking assessment since the end of the 1990s. The studies proposed by the researchers of the Centre for Language and Speech Technology (CLST) at Radboud University (Strik & Cucchiarini, 1999; Cucchiarini et al., 2000) investigated the fluency features of spontaneous speech that can be automatically obtained from the output of an ASR system (e.g., speech rate, articulation rate, number of disfluencies, length and number of pauses, mean length of run, and phonation/time ratio) and their correlation with human-assigned fluency scores. Speech rate appeared to be the best predictor of fluency, although correlations were lower for spontaneous speech than read speech. In the already mentioned paper by Franco et al. (2000), speech rate was also one of the features employed in their automatic scoring system for pronunciation, and its presence in the feature set moderately increased the correlation between the predicted scores and the scores assigned by human experts. In a paper we have mentioned several times, Supnithi et al. (2003) used different features related to vocabulary, grammatical accuracy, sociolinguistic appropriateness, and fluency for the task of predicting the proficiency levels of

Japanese learners of English based on manual transcriptions. The fluency-related features included the frequency of disfluencies (i.e., fillers, repetitions, and self-corrections), the number of sentences included in the response, the duration of the interview, the duration of words, the average sentence length, and the total frequency of all words. As it appears from their ablation study, the fluency-related features seem to have a positive impact on the performance of their proposed models. In its early stages, the feature computation module of ETS's SpeechRater was mostly targeting fluency-related features (Zechner, Higgins, & Xi, 2007) such as the number of disfluencies, the ratio between the number of silences and the number of words, various measures related to silences, the number of words per second, the number of types per second, and other features (Zechner & Bejar, 2006; Zechner, Bejar, & Hemat, 2007). The paper by Chen, Tetreault, & Xi (2010) proposed to move beyond the word-level cues used in Zechner & Bejar (2006) and Zechner, Bejar, & Hemat (2007) and focus on structural events such as structures of clauses and disfluencies. They reported interesting correlations between the features derived from the annotations related to structural events and the holistic proficiency scores. In a subsequent study (Chen & Yoon, 2012) which we mentioned in Section 2.2.1 in relation to syntactic complexity measures, the same group of researchers investigated the application of automatically detected (Chen & Yoon, 2011) structural events on ASR transcriptions for automatic speaking assessment. They found that one of the four features derived from structural events (i.e., a pause-related feature) held an interesting correlation with human holistic scores even when used on ASR transcriptions. Fluency-related features (i.e., features related to silences, disfluencies, and various word-level and phone-level aspects) are also a substantial part of the features employed in the Gaussian Process-based grader for spontaneous speech developed by the ALTA Institute (van Dalen et al., 2015). In addition to these features, their proposed grader also leveraged other audio-related features but no features related to content. Although it obtained remarkable results on the task of predicting human scores, the lack of these features would allow test-takers to deceive the system, as acknowledged by the authors, who eventually integrated them into their grading systems, e.g., in the work by Raina et al. (2020) mentioned in Section 2.2.3. The work by Fontan et al. (2018) on automatic assessment of read-aloud speech by Japanese learners of French proposed a system which did not use an ASR module. Instead, they extracted features by means of a Forward-Backward Divergence Segmentation algorithm (Andre-Obrecht, 1988) to segment speech recordings into units at a subphonemic scale. The combination of these features with other more typical measures achieved interesting performances. The interesting aspect of this approach is the portability to other languages, as it only uses low-level features derived from the

audio signal. However, a downside could be the lack of content-related features if this approach were applied to spontaneous speech, as in the previously mentioned study. Other more recent semi-automatic studies on oral fluency have been mentioned in Section 2.3, whereas in Section 2.2.2, we described the experiments on pronunciation assessment using wav2vec 2.0 proposed by Kim et al. (2022), in which the prediction of fluency scores was also investigated.

Furthermore, the application of DD (see Section 2.2.1) has also been investigated for other aspects of CALL. Lu, Gales, Knill, Manakul, & Wang (2019) and Lu et al. (2020) showed that the use of an LSTM DD model applied to ASR transcriptions could improve the performance of spoken GEC, as it would make transcriptions more similar to written language, for which a much more significant amount of labelled data is typically available. Both DD and spoken GEC will be investigated in Study 2 (see Section 4.2).

Finally, although the analysis of conversational agents for language learning goes beyond the scope of this contribution, we cite the promising work by Ramanarayanan (2020), which presented a human-machine dialogue corpus for the assessment of conversational proficiency. The corpus was manually scored along various dimensions of proficiency, including aspects related to interaction, i.e., engagement, turn-taking, repair, and appropriateness.

Chapter 3

Data

In this chapter, we illustrate the data used in our studies. We divide the chapter into two main sections: the first is devoted to the publicly available datasets, whilst the second describes the non-public datasets. As should be clear at this point, automatic speaking assessment often leverages information extracted from written productions, such as letters, articles, reports, and short stories or essays, in addition to spoken data. Therefore, we also include the description of several written corpora in our account. For completeness, we also devote a short paragraph to the outline of publicly available corpora which have not been considered in our experiments.

3.1 Publicly available data

3.1.1 Written corpora

EFCAMDAT

Arguably the largest publicly available[1] L2 learner corpus, the second release of EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2013; Huang et al., 2017, 2018) comprises 1,180,310 scripts written by 174,743 L2 learners as assignments to Englishtown, an online English language school. There are 128 different writing tasks related to several topics, e.g., describing the rules of a game, reporting a news story, illustrating a homemade remedy for fever, writing to a pen pal, etc. The compositions are annotated with a score on a scale from 0 to 100 and a proficiency level from 1 to 16 (mapped to CEFR levels from A1 to C2).[2] The L1s of the

[1] philarion.mml.cam.ac.uk/

[2] englishlive.ef.com/en/how-it-works/levels-and-certificates/

Code	Meaning	Code	Meaning
XC	change from x to y	NSW	no such word
AG	agreement	PH	phraseology
AR	article	PL	plural
D	delete	PO	possessive
PS	part of speech	PR	prepositions
EX	expression of idiom	SI	singular
IS	insert	VT	verb tense
MW	missing word	WC	word choice
WO	word order	AS	add space
CO	combine sentences	C	capitalisation
HL	highlight	NS	new sentence
PU	punctuation	RS	remove space
SP	spelling		

Table 3.1: EFCAMDAT error tagset.

learners are not available but can be inferred from their nationalities (about 200), among which the best-represented are Brazilian, Chinese, Russian, Mexican, German, French, Italian, Saudi Arabian, Taiwanese, and Japanese. Furthermore, the learner scripts are annotated with POS tags and information on grammatical dependencies using the Penn Treebank Tagset (Marcus et al., 1993) and the SyntaxNet parser (Andor et al., 2016) are partially error-tagged by human experts. The error tagset of the corpus consists of 25 types of errors, which are reported in Table 3.1.

CLC-FCE

The Cambridge Learner Corpus - First Certificate English (CLC-FCE) (Yannakoudakis et al., 2011) is a publicly available section[3] of a large proprietary L2 learner corpus, the Cambridge Learner Corpus (CLC) (see below), developed in collaboration between Cambridge University Press and Cambridge Assessment. It includes the scripts of an English language exam aimed at around B2 level of the CEFR. Its 1244 exam scripts contain responses to two different prompts requiring test-takers to write a short answer (e.g. a letter, an article, a report, a short story) and range from 200 to 400 words on average. Each answer has been annotated by human experts with a mark between 0 and 40. Moreover, the dataset contains an overall score assigned to

[3] `ilexir.co.uk/datasets/index.html`

both prompts. Similarly to EFCAMDAT, the CLC-FCE also features manual annotations with information about errors according to a taxonomy of about 80 error types described in Nicholls (2003). An example drawn from the data in XML format is the following:

I am looking forward to <NS type="FV"> *hear | hearing* </NS> *from you.*

in which FV indicates a verb form error and *hear* is corrected to *hearing*.

The corpus consists of a training set of 1141 scripts and a test set of 97 scripts.

BEA-2019 Shared Task data

As already mentioned in the previous chapter, in 2019, a shared task on GEC was organised within the Workshop on Innovative Use of NLP for Building Educational Applications (Bryant et al., 2019). The organisers released a collection of text-based corpora tagged with GEC annotations, which includes the CLC-FCE, described above, a dataset derived from Cambridge English Write & Improve, the Louvain Corpus of Native English Essays (LOCNESS), the Lang-8 Corpus of Learner English, and the National University of Singapore Corpus of Learner English (NUCLE).[4]

Write & Improve is an online platform where L2 learners of English can practise their writing skills (Yannakoudakis et al., 2018).[5] Users can submit their compositions in response to different prompts, and the Write & Improve automatic system provides assessment and feedback. Some of these compositions have been manually annotated with CEFR levels and grammatical error corrections since 2014, resulting in a corpus of 3,600 texts.

The LOCNESS corpus would originally consist of approximately 400 essays written by L1 English undergraduates from the United Kingdom and the United States (Granger, 1998), but the organisers of the shared task excluded compositions longer than 550 words and containing transcription issues. The final selection amounts to 100 essays.

The Lang-8 Corpus of Learner English is extracted from the data collected on the Lang-8 website,[6] on which users are encouraged to correct each other's grammar (Mizumoto et al., 2012; Tajiri et al., 2012).

The NUCLE is a corpus of 1,400 essays composed of Asian undergraduate students enrolled at the National University of Singapore (Dahlmeier et al., 2013). This corpus has already been employed as the training set for the CoNLL-2013 and CoNLL-2014 shared tasks on GEC, mentioned

[4] `cl.cam.ac.uk/research/nl/bea2019st/`
[5] `writeandimprove.com/`
[6] `lang-8.com/`

in Section 2.2.1.

3.1.2 Written and spoken corpora

ICNALE

The International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2011) is a publicly available dataset[7] comprising written and spoken responses of L2 English learners ranging from A2 to B2 and partially of L1 speakers. The L1s of the L2 speakers are not reported, but they can be inferred from their countries of origin: China, Hong Kong, Indonesia, Japan, South Korea, Pakistan, Philippines, Singapore, Thailand, and Taiwan. The CEFR levels were assigned prior to collecting the data, as the ICNALE team required all the learners to take an L2 vocabulary size test and to present their scores previously obtained in English proficiency tests such as TOEFL, TOEIC (Test of English for International Communication), IELTS, etc. On the basis of these two scores, the learners were classified into proficiency levels. The written section consists of 5,600 essays, whereas the spoken section is composed of 4,400 monologues and 4,250 dialogues. In both the written and the spoken parts, learners are required to express their opinion on the following two statements:

- *It is important for college students to have a part-time job.*

- *Smoking should be completely banned at all the restaurants in the country.*

Only a small section of dialogues and essays has been scored by human experts so far and has been included in the ICNALE Global Rating Archives (Ishikawa, 2020), which includes the assessments and scores (on a scale from 0 to 100) of 140 dialogues and 140 essays assigned by 40 human raters.

To the best of our knowledge, ICNALE is the only publicly available L2 learner corpus to include both written and spoken data specifically designed and annotated for L2 research. Furthermore, the spoken section is provided with audio (monologues) and video (dialogues) data, as well as manual transcriptions.

[7] `language.sakura.ne.jp/icnale/download.html`

3.1.3 Spoken corpora

NICT-JLE

The National Institute of Information and Communications Technology - Japanese Learner English (NICT-JLE) corpus was created in 2004 (Izumi et al., 2004), and its latest version was released in 2012.[8] It contains manual transcriptions of approximately 300 hours of oral interviews with Japanese learners of English, but the audio recordings are not available. A subset of the corpus was manually annotated with about 50 types of errors and corrected. Furthermore, this subset includes annotations about proficiency scores (ranging from A1 to B2), code-switched words, and disfluencies, which are labelled under three types: filled pauses, repetitions, and self-corrections.

KIT Speaking Test Corpus

In 2022, a similar corpus to NICT-JLE was released for public use, i.e., the Kyoto Institute of Technology (KIT) Speaking Corpus, which consists of manual transcriptions of interviews with 574 Japanese undergraduate students for a total of approximately 4,448 hours.[9] As in the case of NICT-JLE, the audio recordings of the KIT Speaking Test corpus are not publicly available. The manual annotations follow the tagging methods of NICT-JLE but only include disfluencies, whereas grammatical errors are not annotated. The proficiency level of the test-takers approximately ranges from A1 to B2 (Kanzawa et al., 2022).

3.2 Non-publicly available data

3.2.1 Written corpora

CLC

We have already introduced the CLC in the paragraph on CLC-FCE. The CLC is an ever-growing collection of data obtained from Cambridge English exams. As reported by O'Keeffe & Mark (2017), in 2017, it included 266,600 examination scripts and 143 L1 backgrounds collected in the period between 1993 and 2012. While the CLC-FCE only consists of a subset of exam scripts of B2-level students, the CLC includes data from lower to advanced proficiency levels.

[8] alaginrc.nict.go.jp/nict_jle/index_E.html#license
[9] kitstcorpus.jp/

For information about the error-tagging method employed in this corpus, the reader may refer to the example shown in the paragraph related to CLC-FCE and to Nicholls (2003).

3.2.2 Written and spoken corpora

TLT-school

In Trentino, an autonomous region in the north of Italy, the linguistic proficiency of Italian students has been assessed in recent years through proficiency tests in both English and German (Gretter et al., 2020), involving about 3000 students ranging from 9 to 16 years old, belonging to four different school grades (5^{th}, 8^{th}, 10^{th}, 11^{th}) and three proficiency levels (A1, A2, B1). Since our experiments are conducted only on the B1 section of the written part and on the A2 and B1 sections of the spoken parts of the English portion of the corpus, we do not describe the texts and utterances of the German section, as their analysis goes beyond the scope of this thesis.

The written section consists of 895 answers to 2 question prompts. Test-takers are asked two questions: the first one requires them to write a blog entry in which they have to describe what happened during the day and talk about their plans for the rest of the week, while the second one asks them to write an email to a friend who broke an object borrowed from them.

Only a subset of the spoken section was manually transcribed and annotated with hesitations, truncated words, and code-switched words from L1 (Italian) and L3 (German) and consists of 1022 responses to 13 small talk questions about everyday life situations, 7 for B1 and 6 for A2 (see Appendix A). We provide detailed information about the manual transcriptions and other aspects of the corpus in Gretter et al. (2020). However, it is worth mentioning that some answers are characterised by a number of issues (e.g., presence of words belonging to multiple languages or presence of off-topic answers).

In addition to CEFR levels, the corpus was annotated with proficiency scores. The total score ranges from 0 to 8 in the written section and from 0 to 12 in the spoken section and consists of the sum of the subscores assigned by human experts for each specific proficiency indicator assigned by the human raters (see Table 3.2). For each indicator, human raters could choose 0, 1 or 2 points. Since every utterance was scored by only one expert, it was not possible to evaluate any kind of agreement among experts. Note that the CEFR levels were assigned before the tests and should be considered as expected proficiency levels, whereas the test scores are effectively representing each learner's performance in the exam.

The type and amount of data, as well as the training/test partition, vary depending on each

study. Therefore they will be illustrated in detail in the next chapters.

Speaking	**Writing**
Relevance	Task fulfillment
Formal correctness	Formal correctness and lexical complexity
Lexical richness and complexity	Cohesion
Pronunciation	Narrative and descriptive skills
Fluency	
Communicative effectiveness	

Table 3.2: TLT-school proficiency indicators for speaking and writing.

3.2.3 Spoken corpora

Linguaskill

In our studies, we also used the candidate responses to the 5 spoken parts of the Linguaskill[10] examinations for L2 learners of English. The data were provided by Cambridge English Language Assessment (Ludlow, 2020). Part 1 consists of answers to eight personal questions, of which the first two are not graded. They last about 10 or 20 seconds. Part 2 features a reading-aloud activity, which includes eight sentences of 10 seconds each. Part 3 and Part 4 test the candidates' ability to deliver a long turn, and they are required to speak for up to one minute. While in the former, the candidates should talk about a given topic, in the latter, they are required to describe one or more graphics, such as diagrams, charts, or information sheets. Finally, in Part 5, test-takers should provide their opinions in the form of responses of about 20 seconds to five questions related to a given topic. Appendix B contains examples of question prompts for each part of the exam. Each part contributes 20% to the speaking exam. Therefore, the overall grade is computed as the average of the grades assigned to the five parts, which are on a scale from 1 to 6 based on CEFR proficiency levels.

Datasets of 31475 and 1033 non-overlapping speakers are employed as the training and development/calibration set, respectively.

For evaluation, we consider two test sets, LinGen, of 1049 speakers, and LinBus, of 712 speakers. LinGen contains learners' responses to questions on General English, whereas LinBus includes answers to questions on Business English. Both test sets feature around 30 L1s and are balanced for gender and proficiency level.

[10] cambridgeenglish.org/exams-and-tests/linguaskill/information-about-the-test/

Switchboard

The Switchboard corpus consists of 260 hours of telephone conversations by L1 American English speakers (Godfrey et al., 1992; Meteer et al., 1995). The Penn Treebank 3 (Taylor et al., 2003) tagset provides manual transcriptions and disfluency annotations on the Switchboard corpus, including filled pauses, repetitions, false starts, and discourse markers (e.g., "so", "right", "okay", "well", etc.). Although it is a public dataset, it is only available for a fee.

3.3 Other spoken corpora

In this section, we provide a brief outline of other spoken corpora that have been used in the field of L2 proficiency assessment and were not considered in our study. We excluded these corpora because they only consist of read-aloud speech or they lack specific annotations on proficiency.

ISLE: the Interactive Spoken Language Education (ISLE) corpus (Menzel et al., 2000) consists of 7,714 read utterances collected from 23 German and 23 Italian intermediate-level speakers of English for a total of 9.5 hours. The corpus is available for a fee.

L2-ARCTIC: the L2-ARCTIC corpus[11] is a publicly available collection of recordings of read speech from 24 L2 speakers of English with 6 different L1 backgrounds, i.e., Arabic, Hindi, Korean, Mandarin, Spanish, and Vietnamese, with their respective orthographic and phonetic transcriptions (Zhao et al., 2018). Furthermore, the corpus is manually annotated with three types of mispronunciation errors (substitutions, deletions, and additions).

speechocean762: speechocean762 is a public dataset[12] consisting of 5,000 read-aloud sentences collected from 250 Mandarin Chinese learners of English (Zhang et al., 2021). The corpus is annotated with multidimensional scores on pronunciation accuracy, stress, fluency, and prosody at different levels (i.e., phoneme-level, word-level, and sentence-level).

Spoken CALL Shared Task data: three shared tasks for spoken CALL took place in 2017, 2018, and 2019, one per year.[13] The first (Baur et al., 2017) was organised by the University of Geneva, the University of Birmingham, and the CLST of Radboud University, while the second and third also involved the ALTA Institute of the University of Cambridge (Baur et al., 2018, 2019). The task consists in a binary classification problem whereby text-based or audio-based automatic systems should predict whether a sentence is linguistically and grammatically acceptable or unacceptable. The data were collected from young Swiss German learners of

[11] psi.engr.tamu.edu/l2-arctic-corpus/
[12] openslr.org/101
[13] regulus.unige.ch/spokencallsharedtask_3rdedition/

English by means of a dialogue interface showing them a prompt in German, e.g., "Frag: Zimmer für 3 Nächte" ("Request: room for 3 nights"), thus allowing them to respond to it with a certain degree of freedom. The third edition of the challenge also employed the data released in the previous two, with a training set of 11,919 utterances, a development set of 995 utterances, and a test set of 1,000 utterances.

CrowdED: the CrowdED corpus (Caines et al., 2016) is a publicly available[14] crowdsourced speech corpus of English collected from L1 and L2 speakers of German and English answering questions on business. In its second release, grammatical error annotations were added to a part of the English section of the corpus for a total of 1108 transcriptions and corrections for 383 unique recordings (Caines et al., 2020). A major issue with the annotations is that they were crowdsourced. This aspect may be particularly problematic, especially for grammatical error corrections.

[14] `ortolang.fr/market/corpora/ortolang-000913`

Chapter 4

GED, GEC, and assessment

In this chapter, we describe two studies which explore the interconnections between grammar and proficiency assessment.

In Study 1, we train a feature extractor on L2 learner written data to obtain information related to grammatical accuracy. Subsequently, we use it for inference on spoken data. The work investigates the impact of the feature extractor on speaking proficiency assessment as well as the written-to-spoken approach.

Study 2 explores DD, GEC, and their potential use for proficiency assessment in a cascaded fashion.

A preliminary analysis of the interconnections between grammar and proficiency assessment can be found in Bannò et al. (2021). Study 1 and part of Study 2 were presented at the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022) and can be found in Bannò & Matassoni (2022) and Lu et al. (2022), respectively. In particular, Study 2 builds on the findings of the paper presented at BEA 2022 and reports additional experiments and results, which have been recently gathered together in an article submitted to *Speech Communication*. Part of this work has been recently resumed and illustrated in a contribution which will be presented at the workshop "AI and Education" of Ital-IA 2023.[1]

[1] `ital-ia2023.it/workshop/ai-ed-educazione`

4.1 Study 1: Cross-corpora experiments of speaking assessment and grammatical error detection

4.1.1 Introduction

As mentioned in Chapter 3, a common issue in the field of automatic speaking assessment is the lack of publicly available data specifically designed and annotated for this purpose. Another typical problem is the lack of consistency and coherence in human assessment, as it frequently relies on proficiency indicators that often have biases and are not clearly generalisable, therefore not easily transferable into automatic scoring systems (Zhang, 2013; Engelhard, 2002). Although L2 proficiency cannot be assessed on the mere basis of the presence of grammatical errors in learners' productions, this aspect is highly consistent and plays a major role in language assessment by human experts (see Section 1.2.1). Nonetheless, the impact of errors on automatic speaking assessment has been sporadically investigated, whereas other types of feature-based assessment have been more widely studied and explored (see Section 2.2.1).

In this study, we address the task of automatically predicting the scores of spoken responses of L2 learners using written data and leveraging the presence of grammatical errors, thus addressing both the problems mentioned above: the issue related to the scarce availability of spoken data and the problem of inconsistency in human assessment.

In order to do so, we design a ranking of grammatical error gravity based on the frequency of each human-annotated error in the EFCAMDAT, modelling it across 15 proficiency levels aligned with CEFR levels ranging from A1 to C1; as our purpose is scoring spoken language proficiency, we discard spelling, punctuation, and orthographic errors, and we group errors into 5 categories.

Subsequently, we train a feature extraction model feeding the learners' texts of the EFCAMDAT as inputs and setting the 5 classes of errors as targets for our predictions, and we use this model as an error-feature extractor (EFEX) for inference on the CLC-FCE and ICNALE, thus generating 5 labels corresponding to the 5 classes of errors mentioned above; then, we train a grader on the CLC-FCE injecting the 5 error labels generated by EFEX, and we test it on the spoken annotated section of ICNALE.

Likewise, we use EFEX for inference on the TLT-school corpus. Subsequently, we train a grading system on the written section of the corpus injecting the 5 error labels generated by EFEX, and we test it on the spoken section. Figure 4.1 shows the proposed pipeline.

Finally, we fine-tune our model on a small spoken subset.

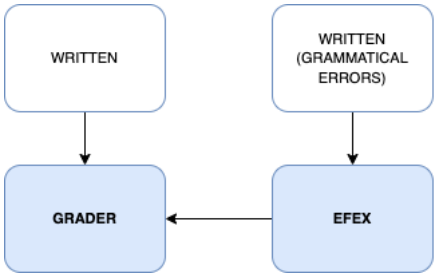


Figure 4.1: Diagram of the proposed training pipeline based on textual input (i.e, the written train set). The grader is then used to predict proficiency scores on manual and ASR transcriptions (i.e., the spoken test set).

4.1.2 Data

EFCAMDAT

As our work investigates the efficacy of errors as features, we only use the error-tagged section of the EFCAMDAT Cleaned Subcorpus (Shatz, 2020), consisting of 498,208 scripts ranging from proficiency level 1 to 15 (mapped to CEFR levels A1 to C1), which we split into training and test set. The error tagset of the corpus consists of 24 types of errors (see Table 3.1), of which we discarded 7 related to punctuation and spelling, as they would be of no use for assessing speech (see Table 4.1). We preliminarily computed the KL-Divergence between the distribution of the 17 error label counts across CEFR proficiency levels in the EFCAMDAT Cleaned Subcorpus. The labels were converted into a smoothed distribution by applying add-one smoothing. The symmetric KL-Divergence was then calculated. Therefore, for error type t_i for proficiency level L_k:

$$\mathrm{P}(t_\mathrm{i}|L_\mathrm{k}) = \frac{\mathrm{cnt}(t_\mathrm{i}, L_\mathrm{k}) + 1}{\sum_{j=1}^{N}(\mathrm{cnt}(t_\mathrm{i}, L_\mathrm{k}) + 1)}$$

where $\mathrm{cnt}(t_\mathrm{i}, L_\mathrm{k})$ is the number of occurrences for a given label at a given grade.

The symmetric KL Divergence was subsequently calculated across proficiency levels:

$$\mathrm{KL}(L_\mathrm{k}|L_\mathrm{l}) = \left(\sum_{i=1}^{N}\mathrm{P}(t_\mathrm{i}|L_\mathrm{k})\log\left(\frac{\mathrm{P}(t_\mathrm{i}|L_\mathrm{k})}{\mathrm{P}(t_\mathrm{i}|L_\mathrm{l})}\right)\right) + \left(\sum_{i=1}^{N}\mathrm{P}(t_\mathrm{i}|L_\mathrm{l})\log\left(\frac{\mathrm{P}(t_\mathrm{i}|L_\mathrm{l})}{\mathrm{P}(t_\mathrm{i}|L_\mathrm{k})}\right)\right)$$

Table 4.2 reports the symmetric KL-Divergence between distributions of counts from all 17 error labels across CEFR proficiency levels. It appears that we can consider errors as criterial features of linguistic proficiency, as there are differences in the distributions of grammatical errors

Code	Meaning	Code	Meaning
XC	change from x to y	NSW	no such word
AG	agreement	PH	phraseology
AR	article	PL	plural
D	delete	PO	possessive
PS	part of speech	PR	prepositions
EX	expression of idiom	SI	singular
IS	insert	VT	verb tense
MW	missing word	WC	word choice
WO	word order		

Table 4.1: EFCAMDAT error tagset without codes related to spelling, punctuation and orthographic errors.

across proficiency levels, to which we can correlate differences in their frequency.

	A1	A2	B1	B2	C1
A1	0.0	0.055	0.065	0.085	0.066
A2	0.055	0.0	0.013	0.029	0.028
B1	0.065	0.013	0.0	0.005	0.009
B2	0.085	0.029	0.005	0.0	0.010
C1	0.066	0.028	0.009	0.010	0.0

Table 4.2: Symmetric KL Divergence between distributions of counts from all 17 error labels in EFCAMDAT.

Ranking of error gravity

In light of this, we analysed the frequency of each type of error across the 15 proficiency levels of the corpus. We calculated it by dividing the sum of all the occurrences of a given type of error in a given proficiency level by the number of texts assigned to a given proficiency level. We then decided to design a ranking of error gravity for each type of error in relation to each proficiency level by introducing a negative bias in the error count when this amounts to 0:

$$b_t = \begin{cases} -1 & 0.1 \leq F_{t,L} < 0.2 \\ -2 & 0.2 \leq F_{t,L} < 0.3 \\ \dots & \\ -9 & 0.9 \leq F_{t,L} < 1.0 \end{cases}$$

where $F_{t,L}$ is the normalised frequency of error type t at proficiency level L; e.g., if $F_{AR,1}$ is 0.2, all the occurrences of error AR at level 1 reporting 0 errors are replaced by -2. The rationale

Errors	**Class**
VT	VT
NSW + PH + EX + MW + WC + WO	LUW
AR + PO + PR + PS	PAP
AG + PL + SI	AG
D + IS + XC	GE

Table 4.3: The 5 error classes used in the study.

behind this idea is to 'award' learners who have not made errors which are frequent in their proficiency level. Subsequently, in order to avoid having a too sparse representation, we grouped the 17 types of errors into 5 classes of errors: verb tense (VT), lexis and use of words (LUW), prepositions, articles, possessives and part of speech (PAP), agreement (AG) and generic errors (GE), as shown in Table 4.3. We divided each of the 5 error counts by the word count in order to also weigh the text length. Finally, the error count in each level is normalised on a scale from 0 to 1.

Before applying our ranking of error gravity and introducing the negative bias, we also computed the mean of the error ratio (i.e., the number of errors divided by the number of words) of each of the 5 classes and of their sum for each proficiency level (see Table 4.4). Furthermore, we performed ANOVA on each of the 5 classes, and we always obtained significant p-values (<0.05), finding that there are significant differences between proficiency levels in terms of errors.

	mean (%)				
	A1	**A2**	**B1**	**B2**	**C1**
LUW	3.67	3.10	2.69	1.96	1.58
PAP	1.63	1.42	1.20	0.99	0.70
AG	0.99	0.49	0.47	0.36	0.31
GE	2.00	1.67	1.29	0.95	0.80
VT	0.31	0.43	0.41	0.36	0.19
total	8.62	7.13	6.08	4.63	3.59

Table 4.4: Mean of the ratio (number of errors divided by number of words) of each error class and their sum for each proficiency level.

ICNALE

In order to test our approach, we used ICNALE. As mentioned already in Section 3.1.2, only a small section of essays and dialogues has been scored by human experts so far and has been included in the ICNALE Global Rating Archives (Ishikawa, 2020), which currently include as-

sessments and scores (on a scale from 0 to 100) of 140 dialogues and 140 essays by 40 human raters. Since not all the dialogues and essays were previously assigned a proficiency level, for our experiments, we selected only the ones classified into CEFR levels and scored by human experts, and we also considered the scored texts and utterances produced by L1 speakers, therefore reducing the written section to 121 essays and the spoken section to 116 dialogues, of which we considered only the learners' utterances (i.e., we removed the parts containing speech uttered by the interviewers). Out of the 40 raters involved in the project, we only selected the L1 speakers with more than 5 years of experience in L2 English teaching and assessment, i.e., 4 raters for the written section and 3 raters for the spoken section. We set the average of these scores as targets. Details about the average and standard deviation of the raters' scores can be found in Ishikawa (2020).

CLC-FCE

Due to the limited amount of annotated data in the ICNALE corpus, we train our models on the CLC-FCE corpus (see Section 3.1.1). The corpus contains the scripts of an English language exam aimed at around B2 level of the CEFR, which is also the highest level of the ICNALE corpus. Note that we eliminated the answers that did not report a score.

TLT-school

Our experiments are conducted on the B1 section of the English written and spoken parts of the corpus. The written section consists of 895 answers to 2 question prompts. Test-takers are asked two questions: the first one requires them to write a blog entry in which they have to describe what happened during the day and talk about their plans for the rest of the week, whereas the second one asks them to write an email to a friend who broke an object borrowed from them. The spoken section considered in this study consists of 442 responses to 7 small talk questions about everyday life situations. It is worth mentioning that some answers are characterized by a number of issues (e.g., presence of words belonging to multiple languages or presence of off-topic answers).

As regards the speech transcriptions, for this set of experiments, we eliminated the annotations related to spontaneous speech phenomena such as hesitations, fragments of words, etc. As for the ASR output text, its word error rate is 41.13% for the B1 subset we used in our experiments; the acoustic and language models are described in Gretter et al. (2019), in which the reader can also find details about the training data used for ASR development. The total score

ranges from 0 to 8 in the written section and from 0 to 12 in the spoken section and consists of the sum of the subscores assigned by human experts for each specific proficiency indicator (i.e., fulfilment, formal correctness and lexical complexity, cohesion, and narrative and descriptive competences for writing; and relevance, formal correctness, lexical complexity, pronunciation, fluency, and communicative effectiveness for speaking). For each indicator, human raters could choose 0, 1 or 2 points. Note that the CEFR levels were assigned before the tests and should be considered as expected proficiency levels, whereas the test scores are effectively representing each learner's performance in the exam. Table 4.6 shows the number of answers and word counts of the TLT-school spoken test set across test scores.

	ICNALE		**CLC**	**TLT**	
	Wr	**Sp**		**Wr**	**Sp**
Train	-	-	2122	594	345
Dev	-	-	160	-	-
Test	121	116	194	301	97
Avg. len	225	186	192	103	28
Max. len	302	455	462	279	221
Min. len	179	23	72	1	1
Score	0-100	0-100	1-40	0-8	0-12

Table 4.5: Statistics (number of answers and word counts) for the three test sets: ICNALE (Written and Spoken), CLC-FCE, TLT-school (Written and Spoken).

Score	**Samples**	**Min. len**	**Max. len**	**Avg. len**
0-3	27	1	100	11.18
3-6	23	9	85	22.00
6-9	14	11	51	27.07
9-12	33	20	196	55.57

Table 4.6: Statistics (number of answers and word counts) for the TLT-school spoken test set across test scores.

4.1.3 Model architectures

We build our models using a BERT architecture (Devlin et al., 2019) in the version provided by the HuggingFace Transformer Library (Wolf et al., 2020).[2] In both the feature extractor and the graders, the BERT layers are frozen.

[2] `huggingface.com/bert-base-uncased`

Feature extractor

In particular, EFEX takes a sequence of token embeddings, i.e., of the answers provided by the learners $[x_1, ..., x_n]$ as inputs and predicts the 'biased' estimate (see formula in Section 4.1.2) of the error rate of each class of error, i.e., VT, LUW, PAP, AG, and GE. Each rate is calculated by a final dense layer, and the model uses mean squared error (MSE) as the loss function. For the GE and LUW outputs, we add one and two extra dense layers, respectively. We used the Adam optimizer (Kingma & Ba, 2015) with learning rate set to 8e-6, batch size 16, validation split 0.1, and we trained our models for 60 epochs. Figure 4.2 shows the architecture of EFEX.

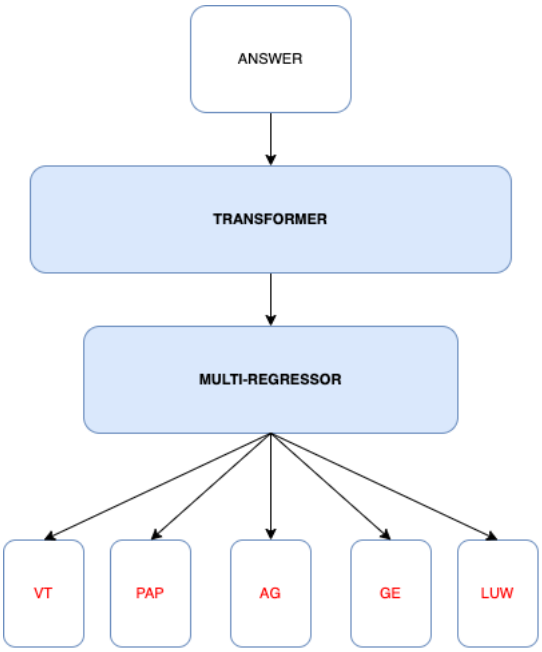


Figure 4.2: EFEX model architecture.

Graders

Before testing the impact of the labels generated by EFEX, we run several experiments on the selected datasets using our simple baseline grading systems, which take only a sequence of token embeddings, i.e., of the answers provided by the test-takers $[x_1, ..., x_n]$, as inputs and predict the total score of each answer normalised on a scale from -1 to 1. Instead, the EFEX-enriched models are fed with the answers combined with a 5-dimensional vector, i.e., the number of classes of errors generated by EFEX, and have the same outputs as the baselines, as shown in Figure 4.3.

In both the baseline and the EFEX-enriched models, the scores are calculated by a final dense

layer, and the model employs MSE as the loss function. The structure and hyper-parameters of the models are shown in Table 4.7. For evaluation, we consider two metrics: MSE and Pearson's correlation coefficient (PCC) between the true scores and the predicted ones.

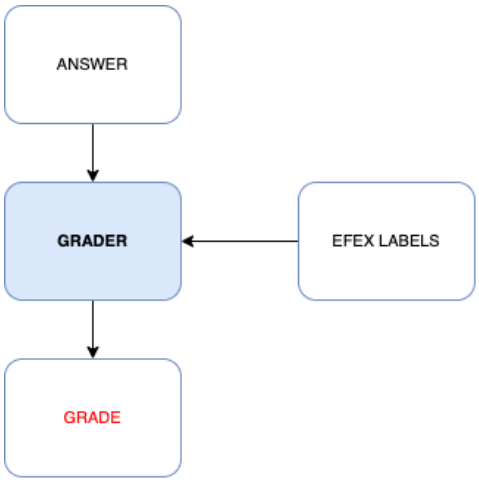


Figure 4.3: Grader architecture.

	TLT	**CLC/ICNALE**
Max. seq. len.	256	512
Learning rate	9e-6	2e-6
Epochs	60 (120)	60 (150)
Batch size	32	16
1st Dense layer	768 - relu	768 - relu
Dropout	0.2	0.2
2nd Dense layer	128 - relu	64 - relu
Dropout	0.2	0.2
Output layer	1	1

Table 4.7: Model architectures and hyperparameters. The number of epochs in brackets refer to the EFEX-enriched model.

4.1.4 Experiments and results

CLC-FCE to ICNALE

We run a series of experiments starting from training EFEX on the EFCAMDAT dataset, setting VT, PAP, AG, GE, and LUW as our prediction targets and feeding only the input text. We tested EFEX on the EFCAMDAT test set, and we obtained significant results when comparing the true labels with the predicted ones in terms of PCC (see Table 4.8).

EFCAMDAT	**VT**	**GE**	**PAP**	**AG**	**LUW**
PCC	0.876	0.831	0.862	0.868	0.796

Table 4.8: EFEX performance in terms of PCC on EFCAMDAT.

Secondly, we run the scorer on ICNALE (see Table 4.9); since we do not have enough ICNALE data for proper training, we train our graders on the CLC-FCE. Considering that we test our models trained on the CLC-FCE directly on out-of-domain data without fine-tuning, we achieve overall interesting results. In this case, the performance of the EFEX-enriched model is slightly lower than the baseline when tested on the scores of the ICNALE written set but still better in terms of PCC when used for predicting the scores of the spoken set.

ICNALE	**Written**		**Spoken**	
Model	MSE	PCC	MSE	PCC
CLC baseline	0.201	0.719	0.121	0.614
+ EFEX labels	0.254	0.709	0.134	**0.625**

Table 4.9: Results on the ICNALE test dataset (MSE and PCC).

TLT-school - Written to spoken

Finally, we run our experiments on the TLT-school, training our baseline on the written training set and testing it on the spoken test set. We follow the same steps with our EFEX-enriched model, and we gain a higher performance when predicting the spoken scores both using the manual and the ASR transcriptions, as shown in Table 4.10. Additionally, we fine-tune our model on the spoken training set for 2 epochs reducing the learning rate to 2e-6, and we obtain our best performance, reaching a PCC of 0.764 on the manual transcriptions. The results on the ASR output also appear to be enhanced by fine-tuning, as we obtain a PCC of 0.642. Fine-tuning the baseline without additional features reaches a PCC of 0.741 on the manual transcriptions and 0.609 on the ASR. We find that the EFEX-enriched model achieves higher results across both metrics.

Furthermore, we continue our analysis by comparing the performance of the baseline and the EFEX-enriched model across test scores. Figure 4.4 shows the MSE variation across 4 ranges of scores, i.e., 0-3, 3-6, 6-9, and 9-12. It can be observed that the MSE is always lower for the EFEX-enriched model except in the range of scores between 0 and 3 both on the manual and ASR transcriptions, for which the EFEX-enriched model shows a modest increase of the MSE.

	TLT - Spoken			
	Man. transcr.		**ASR**	
	MSE	PCC	MSE	PCC
Baseline	0.555	0.734	0.793	0.605
+ fine-tuning	0.488	0.741	0.715	0.609
+ EFEX labels	0.468	0.759	0.688	0.638
+ fine-tuning	**0.400**	**0.764**	**0.606**	**0.642**

Table 4.10: Results on the TLT test dataset (MSE and PCC): baseline; baseline + fine-tuning; baseline + EFEX labels; baseline + EFEX labels + fine-tuning.

Such difference is probably due to the fact that, in this specific range of scores, learners' answers, in addition to having lower quality, are also shorter on average (about 11 words). As the score increases, the word average rises to 56 for scores between 9 and 12. Fewer words also means fewer and a more limited variety of errors. Therefore, EFEX might be introducing some information that is not needed for answers with lower scores. Specifically, the error distribution for the lowest range might be less informative, as can be inferred from the Frobenius norm values of the EFEX vectors for each score range shown in Table 4.11.

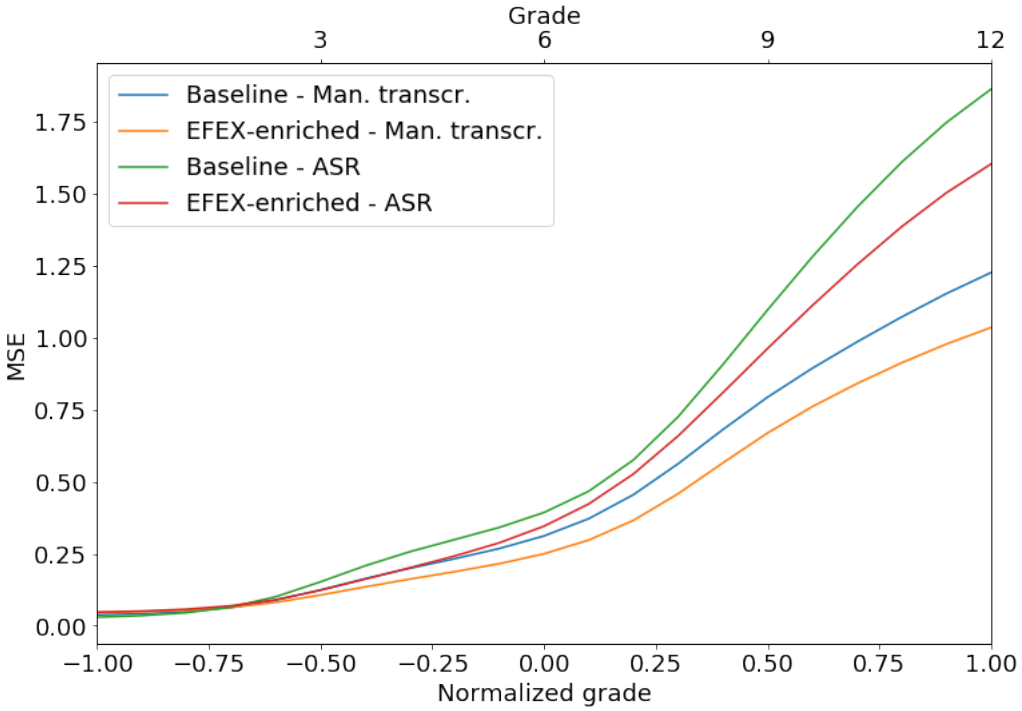


Figure 4.4: MSE variation across scores on manual transcriptions and ASR output text.

	Norm	
Score range	**Man. transcr.**	**ASR**
0-3	1.786	1.780
3-6	2.386	2.540
6-9	2.022	2.090
9-12	4.011	3.986

Table 4.11: Frobenius norm values of EFEX vectors across score ranges.

4.1.5 Conclusions

In this study, we presented a promising approach to automatic proficiency assessment of spoken responses based on the presence of errors across proficiency levels, extracted with an error feature extractor which was developed using a BERT-based architecture. Furthermore, we proposed to use models previously trained on written data in order to address the problem related to the limited availability of spoken data. First, we tried our error-based approach on some publicly available datasets, training our models on the CLC-FCE and testing them on the ICNALE Global Rating Archives. In this case, we found that our EFEX-enriched model managed to modestly improve the prediction of the dialogue scores in terms of PCC. Specifically for this experiment, one also has to consider the differences in domain and scoring metrics between the two corpora.

Subsequently, we discovered that the use of EFEX labels shows a more interesting improvement when scoring the spoken section of the TLT-school corpus after training our models on written data, suggesting that these additional features can mitigate the impact of ASR errors and some typical phenomena of the spoken modality. An example drawn from the data could be the following: *in fact when a person does a lot of movement and moves a lot and goes out in the in the nature then his his body is in more healthy.* The repetitions "in the" and "his", as well as what appears to be a wrongly inserted preposition "in", would be considered actual errors if they occurred in written productions, but not necessarily in spoken texts.

Our assumption is that BERT models, as they are trained on a large quantity of written data, already possess written grammatical knowledge and are sensitive to grammatical violations to a certain extent. Therefore, when evaluating written proficiency, they do not need to be warned with explicit indications concerning errors, but error-related features can be beneficial to understanding and decoding the typical phenomena of oral language and learning spoken and conversational grammar.

Despite such interesting results, this study still has several limitations. Considering that in

spoken responses, the grading module could take advantage of a distinction of errors made by the speaker or introduced by the ASR module (Knill et al., 2019; Lu, Gales, Knill, Manakul, Wang, & Wang, 2019), we assume that there is still room for improvement in the approaches that detect errors as additional features. In this regard, another limitation is that our proposed system extracts general information about 5 broad categories of errors, but it would be interesting to narrow down and better define error types in order to give learners and testers more specific and granular information about grammatical proficiency.

Despite its effectiveness, another limitation of this study is the introduction of a human bias, i.e., the ranking of error gravity. We could address this issue by investigating other strategies for extracting and weighing errors, e.g., an attention mechanism.

Furthermore, given that we removed spontaneous speech phenomena such as hesitations and fragments of words from the manual transcriptions for our experiments, further work could explore a combination of the approach presented in this study and the use of error-related features derived from audio recordings, such as phonological errors, as well as repetitions and other types of disfluencies, which we investigate in Study 2. Moreover, in this study, we worked on a restricted range of proficiency. Therefore, further work should also consider other CEFR levels.

Finally, we acknowledge that the presence of errors cannot be the only feature to be taken into account when assessing L2 proficiency at higher levels, but if properly weighted and balanced with other proficiency indicators, it might improve consistency and objectivity in assessment.

4.2 Study 2: Using grammatical error correction for speaking assessment

4.2.1 Introduction

Mastering grammar is a foundational aspect of L2 proficiency, as shown in Section 1.2.1, and text-based GEC has been thoroughly studied over the past decade (see Section 2.2.1). With speaking skills playing a major role in language learning, it has become increasingly important to analyse spoken grammar. As mentioned already, several previous studies have explored GED on spoken language transcriptions (Knill et al., 2019; Caines et al., 2020; Lu, Gales, Knill, Manakul, Wang, & Wang, 2019) and the combination of disfluency removal and grammar correction on spontaneous learner speech (Lu et al., 2020), while few other studies have investigated the impact of features related to grammatical accuracy (Supnithi et al., 2003; Hasan & Khaing, 2008) and complexity (Chen & Zechner, 2011; Chen & Yoon, 2012; Yoon & Bhat, 2012; Bhat et al., 2014) on automatic assessment of speaking proficiency.

The present contribution partly builds on the findings of our previous work on spoken GEC (Lu et al., 2022). Specifically, this study is divided into three interconnected main parts: in the first, we explore the task of DD; in the second, we investigate spoken GEC; in the third part, we investigate the task of proficiency assessment using a transformer-based grader fed with grammatical features obtained through spoken GEC and DD. Figure 4.5 shows a diagram representing the pipeline proposed in this study. In our experiments, we only use publicly available data for training our models, which we test on the TLT-school data (see Section 3.2.2). In this sense, this work also addresses the issue related to the scarce availability of publicly available spoken learner datasets by using information extracted from written corpora, which are typically easier to obtain.

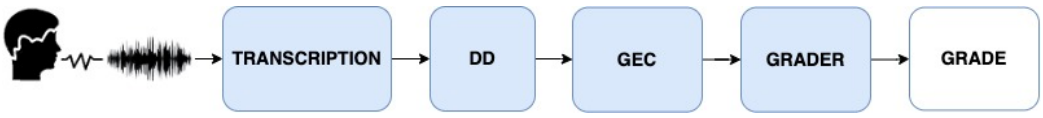


Figure 4.5: The pipeline proposed in this study.

4.2.2 Data

NICT-JLE and KIT Speaking Test Corpus

We used both the NICT-JLE and the KIT Speaking Test Corpus to train our DD module. For more information about these two corpora, the reader can refer to Section 3.1.3.

EFCAMDAT

The EFCAMDAT corpus is described in detail in Section 3.1.1. For this set of experiments, since the dataset contains noisy responses and incorrect annotations, we only kept 762,475 responses after removing punctuation and capitalisation in order to make them more similar to speech transcriptions. For our experiments on spoken GEC, we used spaCy[3] to extract sentences from the parallel responses (i.e., original versus correct), removed sentences shorter than 4 words, removed sentences containing broken XML tags and manual annotations on word limit, and finally, we excluded the parallel sentences where the token edit distance is higher than 60% of the length of the original sentence along the lines of Lo et al. (2018) in order to guarantee consistency between the original sentences and their corrected counterparts.

In addition to using this dataset with the BEA-2019 Shared Task data (see below) for training our spoken GEC system, we used it to pre-train our GEC-based grader and subsequently fine-tune it on the TLT-school data. In this case, since we are not dealing with sentences but with full responses, we kept a higher amount of data. We randomly split EFCAMDAT in a stratified fashion, using the proficiency levels as the class label, and obtained 724,351 responses for the training set and 38,124 for the development set.[4]

In other words, we used EFCAMDAT sentence-wise for spoken GEC, whereas we used it response-wise for proficiency assessment.

BEA-2019 Shared Task data

The BEA-2019 Shared Task data is a collection of text-based corpora tagged with GEC annotations, which includes the CLC-FCE, a dataset derived from Cambridge English Write & Improve, the LOCNESS, the Lang-8 Corpus of Learner English, and the NUCLE, and is described in detail in Section 3.1.1.

[3] `spacy.io`

[4] Note that we used the full range of proficiency levels from 1 to 16 (i.e., from A1 to C2) unlike in Study 1, in which we used the EFCAMDAT Cleaned Corpus (Shatz, 2020) (i.e., from A1 to C1).

Since the CLC-FCE test set has been used in previous studies (Fathullah et al., 2021; Lu et al., 2022) as a benchmark for assessing the performance of spoken GEC systems, we did the same in our experiments. Therefore, we kept it out of the training data.

As we did with EFCAMDAT, punctuation and capitalisation have been removed from all the BEA-2019 data.

Including EFCAMDAT, the data used for training the spoken GEC system amount to 2,552,825 sentences, which we randomly split into a training set of 2,527,296 and a development set of 25,529 sentences.

Disfluency detection

Corpus	Use	#Sent	#Word	%Dsf
NICT-JLE	train	27.5K	178.8K	25.6
KIT Speaking Test	train	18.8K	263.6K	27.3
TLT-GEC dev	dev	605	12.2K	16.5
TLT-GEC test	test	522	10.2K	16.8
LIN-MAN	test	3,361	38K	5.0

GEC

Corpus	Use	#Sent	#Word
EFCAMDAT	train & dev	1.4M	17.4M
BEA	train & dev	1M	11.5M
TLT-GEC test	test	522	10.2K
CLC-FCE test	test	2,681	37K

Proficiency assessment

Corpus	Use	#Response	#Word
EFCAMDAT train	train	724.3K	44.4M
EFCAMDAT dev	dev	38.1K	2.3M
TLT-school train	train	345	11.6K
TLT-school dev	dev	92	3K
TLT-school test	test	97	3.3K

Table 4.12: Corpora statistics. Note that the table reporting the statistics on the datasets used for proficiency assessment shows the number of responses (which may consist of more than one sentences) as opposed to the number of sentences reported in the tables above.

TLT-school

As in Study 1, our experiments on proficiency assessment are conducted only on the English portion of the corpus labelled as B1.

Specifically, the data we used for automatic assessment consist of 534 responses to 7 small

talk questions about everyday life situations (see Appendix A). We used 345 responses for the training set, 92 for the development set, and 97 for the test set.[5]

As regards the ASR output text of the data used for proficiency assessment, its WER is 41.13%; acoustic and language models are described in Gretter et al. (2019). As mentioned in Section 3.2.2, in addition to pre-assigned CEFR levels, the corpus was annotated with proficiency scores. The total score ranges from 0 to 12 and consists of the sum of the subscores assigned by human experts for each specific proficiency indicator (i.e., relevance, formal correctness, lexical richness and complexity, pronunciation, fluency, and communicative effectiveness). For each indicator, human raters could choose 0, 1 or 2 points. In our set of experiments, we considered the total score and the score related to formal correctness, given the specific focus on grammar of this study.

TLT-GEC

We have recently added manual annotations on disfluencies and grammatical error corrections to a part of the TLT-school data, which we refer to as TLT-GEC hereafter. In particular, 386 out of the 534 responses mentioned earlier were filtered, segmented and annotated, and they correspond to 585 sentences. Additionally, we segmented and annotated 301 responses extracted from learners' responses pre-labelled as A2, which correspond to 542 sentences. Therefore, the TLT-GEC dataset amounts to 1127 sentences for a total of 4.96 hours. We split the data into two sets, a development set of 605 sentences and a test set of 522 sentences with non-overlapping speakers.

LIN-MAN

LIN-MAN is a subset of the Linguaskill dataset described in Section 3.2.3. It consists of 833 learners from over 15 L1s, evenly distributed across CEFR proficiency levels. Manual transcriptions are segmented at the phrase level and annotated with disfluencies. We considered this dataset only for DD in order to provide a further comparison and align with previous experiments.

Table 4.12 summarises relevant information about the corpora used in our study.

[5] Our experiments described in Study 1 did not feature the development set.

4.2.3 Disfluency detection

Model and metrics

We performed DD as a sequence tagging task using BERT-based (Devlin et al., 2019) token classifier:

$$\mathbf{d}_{1:M} = \text{BERT}(w_{1:M}) \quad p(r_m|w_{1:M}) = f_d(\mathbf{d}_m)$$

where r_m is a binary tag which indicates whether word w_m is fluent or disfluent. Subsequently, all words classified as disfluencies are removed from the transcriptions. Table 4.13 considers the example previously shown in Table 2.2 and clarifies each passage once again.

Disfluent	**uhm** he see **the** the thief is catched by policeman the **la-** last night
Fluent	he see the thief is catched by policeman the last night
Corrected	he saw the thief caught by a policeman last night

Table 4.13: DD+spoken GEC. The disfluencies are indicated in bold.

Specifically, the BERT-based model consists of a BERT layer in the version provided by the HuggingFace Transformer Library (Wolf et al., 2020),[6] a dropout layer, a dense layer of 768 nodes, a dropout layer, another dense layer of 128 nodes, and finally, the output layer. The model is trained on NICT-JLE and KIT Speaking Test Corpus and uses an Adam optimiser (Kingma & Ba, 2015) with batch size 64, learning rate 1e-06, dropout rate 0.2, and negative log-likelihood as the loss.

For evaluation, we use precision, recall, and F_1 scores.

Experimental results

Table 4.14 shows the results of the DD model on the test set of TLT-GEC in terms of precision, recall and F_1 score. In our previous work (Lu et al., 2022), we used a DD model with the same architecture trained on a large non-publicly available corpus of L1 English speakers, i.e., the Switchboard corpus (Meteer et al., 1995) (see Section 3.2.3), and we tested it on LIN-MAN, a small proprietary dataset of L2 speakers. Therefore, in order to have a further comparison, we also report the results of the previous DD model on LIN-MAN in terms of F_1 score, as well as the results of the DD model used in this study. Considering that, for our model, we used a training set that amounts to approximately half of the data used in our previous experiments, we

[6] `huggingface.com/bert-base-uncased`

obtained interesting results, which are practically aligned to the ones reported in our previous study, as it also appears from the precision and recall curves shown in Figure 4.6.

	Precision	**Recall**	F_1
TLT-GEC test	80.94	83.93	82.41
LIN-MAN	-	-	76.33
LIN-MAN (Lu et al., 2022)	-	-	79.52

Table 4.14: Results of DD on the TLT-GEC test set and LIN-MAN in terms of Precision, Recall, and F_1 Score.

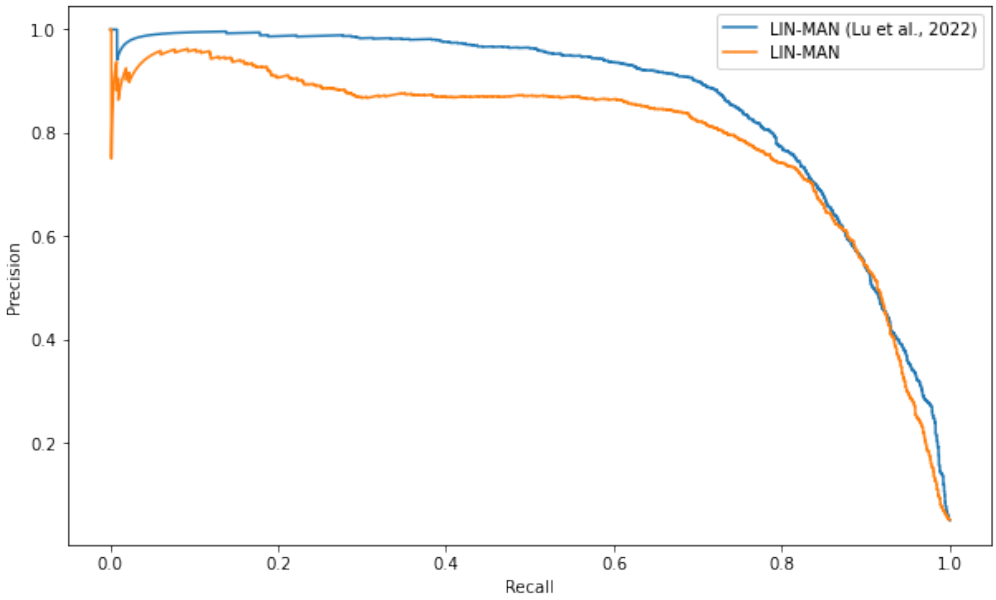


Figure 4.6: Precision and Recall curves: the model used in this study versus the model used in Lu et al. (2022).

4.2.4 GEC

Model and metrics

For the GEC model, we used a T5 model (Raffel et al., 2020) initialised from the version provided by the HuggingFace Transformer Library (Wolf et al., 2020)[7] trained on EFCAMDAT and BEA-2019 with the exclusion of the CLC-FCE test set, which we used to compare the results on

[7] `huggingface.com/t5-base`

TLT-GEC. We set the maximum sequence length to 64 using an AdamW optimiser (Loshchilov & Hutter, 2019) with learning rate 1e-5, batch size 32.

To evaluate the performance of our model, we use two common metrics for GEC, i.e., M^2 score (Dahlmeier & Ng, 2012) and GLEU (Napoles et al., 2015). The former computes the F-score of edits over the optimal phrasal alignment between the hypothesis and the reference sentences, whereas the latter is inspired by BLEU (Papineni et al., 2002) and captures grammatical corrections as well as fluency rewrites.

Experimental results

In Table 4.15, we report the results of the spoken GEC system on the TLT-GEC test set in terms of M^2 and GLEU. For further comparison, we also report the results of our model on the CLC-FCE test, and we compare them to the results of the GEC model described in our previous study.

	GLEU	$\boldsymbol{M^2}$
CLC-FCE test	70.05	57.86
CLC-FCE test (Lu et al., 2022)	-	56.60
TLT-GEC test(dsf)	35.58	43.51
TLT-GEC test(flt)	66.19	57.26
TLT-GEC test(autoflt)	58.60	50.08

Table 4.15: Results of GEC on CLC-FCE test set and TLT-GEC test set in terms of M^2 and GLEU (**dsf** = transcriptions with disfluencies; **flt** = transcriptions with disfluencies manually removed; **autoflt** = transcriptions with disfluencies automatically removed).

Considering the performance on the CLC-FCE test set, it can be observed that our proposed model performs moderately better than the model from our previous study. Similarly to what we observed for DD, these results are quite remarkable, given that we used only publicly available data, whereas our previous study employed the entire CLC corpus (see Section 3.2.1) in addition to the BEA-2019 data.

For completeness, we report the results on the TLT-GEC test set considering the performance of the GEC model on the transcription with disfluencies (dsf), with disfluencies manually removed (flt), and with disfluencies automatically removed (autoflt). As expected, there is a remarkable improvement both in terms of GLEU and M^2 when disfluencies are removed from the transcriptions.

Error analysis

Before moving on to proficiency assessment, we run the spoken GEC system for inference on the transcriptions of the TLT-school data, and we pass the original responses and the automatically corrected ones through ERRANT in order to obtain GEC edit labels. Similarly to what we did in Study 1, we compute the KL-Divergence between the distribution of the 38 resulting ERRANT edit label counts across the scores related to formal correctness in all the TLT-school data. The labels are converted into a smoothed distribution by applying add-one smoothing, and the symmetric KL-Divergence is calculated. Therefore, for error type t_i for score L_k:

$$\mathrm{P}(t_\mathrm{i}|L_\mathrm{k}) = \frac{\mathrm{cnt}(t_\mathrm{i}, L_\mathrm{k}) + 1}{\sum_{j=1}^{N}(\mathrm{cnt}(t_\mathrm{i}, L_\mathrm{k}) + 1)}$$

where $\mathrm{cnt}(t_\mathrm{i}, L_\mathrm{k})$ is the number of occurrences for a given error type at a given score.

The symmetric KL Divergence is then calculated across formal correctness scores:

$$\mathrm{KL}(L_\mathrm{k}|L_\mathrm{l}) = \left(\sum_{i=1}^{N}\mathrm{P}(t_\mathrm{i}|L_\mathrm{k})\log\left(\frac{\mathrm{P}(t_\mathrm{i}|L_\mathrm{k})}{\mathrm{P}(t_\mathrm{i}|L_\mathrm{l})}\right)\right) + \left(\sum_{i=1}^{N}\mathrm{P}(t_\mathrm{i}|L_\mathrm{l})\log\left(\frac{\mathrm{P}(t_\mathrm{i}|L_\mathrm{l})}{\mathrm{P}(t_\mathrm{i}|L_\mathrm{k})}\right)\right)$$

Table 4.16 reports the symmetric KL-Divergence between distributions of counts from all the 38 ERRANT edit labels across formal correctness scores. In light of the evident differences in the distributions of grammatical errors across scores, it appears that we can consider errors as criterial features of linguistic proficiency.

SCORE	**0**	**1**	**2**
0	0.0	0.567	1.262
1	0.567	0.0	0.376
2	1.262	0.376	0.0

Table 4.16: Symmetric KL Divergence between distributions of counts from all 38 ERRANT edit labels in the TLT-school data across formal correctness scores.

4.2.5 Proficiency assessment

Models and metrics

GEC-based grader At the end of our pipeline, there is a grading system which leverages grammatical features extracted through the spoken GEC system, which is used for inference on the TLT-school data. Subsequently, the original transcriptions (after disfluency removal) and

the respective automatically corrected versions are passed through ERRANT in order to obtain sequences of GEC edit labels containing grammatical information. Some examples of labels are `R:VERB:FORM`, which indicates an incorrect verb form, and `R:VERB:SVA`, which indicates an error in subject-verb agreement. These labels contain information about grammatical accuracy and complexity in that they highlight the presence of an error and indicate the POS of the incorrect word. Sequences of GEC edit labels are finally fed into a transformer-based grader, which predicts proficiency scores. As mentioned in the introduction, we conducted our experiments both on holistic proficiency scores and on scores related to formal correctness. Figure 4.7 provides a detailed representation of the pipeline using an example drawn from the data.

The GEC-based grader is first pre-trained on predicting the CEFR levels of EFCAMDAT using sequences of GEC edit labels obtained after feeding the original responses of this corpus and the respective manually corrected ones into ERRANT. Subsequently, the grader is fine-tuned on the TLT-school data on predicting proficiency scores from 0 to 12. Specifically, it consists of an embedding layer with size 128, a transformer block with hidden layer size 128 and 8 heads, a stack of three dense layers of 128 nodes, and finally, the output layer. The training uses an Adam optimiser with batch size set to 512 and learning rate to 1e-6, and MSE as the loss function. When the grader is fine-tuned on the TLT-school data, the transformer block is frozen.

BERT-based grader For comparison, we use a BERT-based baseline which is trained directly on the TLT-school training set. We use the version provided by the HuggingFace Transformer Library (Wolf et al., 2020).[8] The grader takes a sequence of token embeddings, i.e., of the answers provided by the learners (containing disfluencies) as inputs. Each token is transformed into a vector representation and then passed to BERT's encoder layer. We use the [CLS] token state and feed it to a regression head, which consists of a Dense layer of 768 units, a Dropout layer, a Dense layer of 128 units, another Dropout layer, and finally, the output layer. We train the model using an Adam optimiser with learning rate set to 2e-5, batch size 256, dropout rate 0.2, and MSE as the loss function. The BERT layer is frozen.

Combinations In addition to investigating the performance of the GEC-based and the BERT-based graders, we also explore two different combinations.

The first (*shallow*) is performed by means of an Ordinary Least Squares (OLS) multiple linear regression model using the two graders' predictions as predictors and setting the reference score

[8] `huggingface.com/bert-base-uncased`

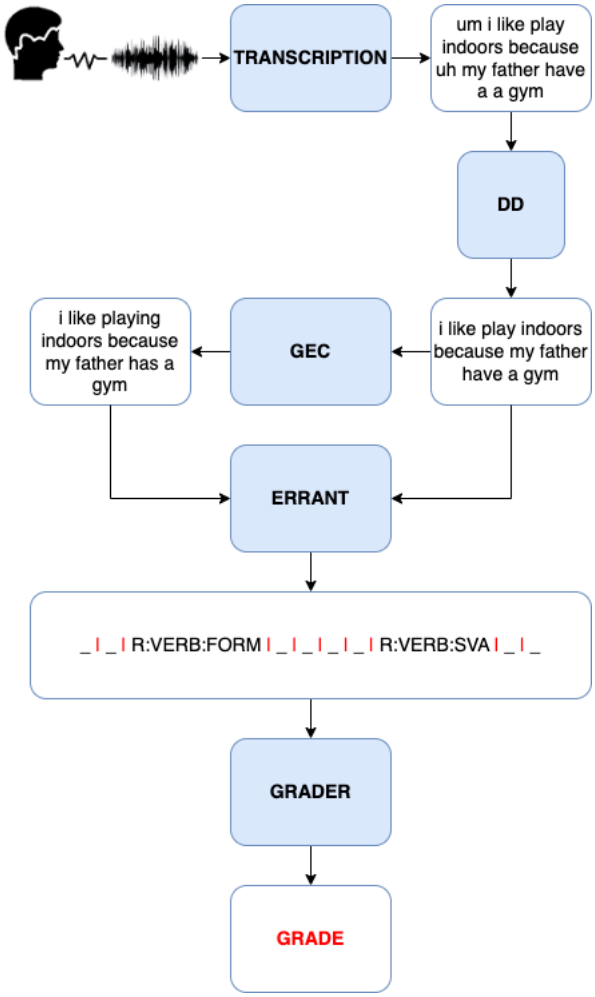


Figure 4.7: Detailed diagram of the proposed pipeline.

$\hat{y}^{(i)}$ as the target:

$$\hat{y}^{(i)} = \beta_0 + \beta_{g_c}\hat{y}^{(i)}_{g_c} + \beta_{b_t}\hat{y}^{(i)}_{b_t} + \epsilon$$

where β_0 represents the intercept, β_{g_c} is the coefficient for the prediction of the GEC-based grader $\hat{y}_{g_c}$, β_{b_t} is the coefficient for the prediction of the BERT-based grader $\hat{y}_{b_t}$, and ϵ is the model's residual. The linear model is trained on the development set.

The second combination (*deep*) is performed by concatenating the embeddings obtained with the two graders and mounting them on a small network. In particular, the BERT embeddings are fed to a Dense layer of 768 units, whilst the embeddings of the GEC-based grader are fed to a Dense layer of 128 units. The two layers are concatenated and passed through the head

of the network, which consists of a Dropout layer, a Dense layer of 128 units, another Dropout layer, and finally, the output layer. The model is trained on the TLT-school training set using an Adam optimiser with learning rate 2e-5, batch size 256, dropout rate 0.2, and MSE as the loss function. We keep the BERT layer and the transformer block of the GEC-based grader frozen.

The performance of the grading systems is evaluated using MSE, PCC, and Spearman's rank correlation coefficient (SRC).

Experimental results

Manual transcriptions We begin our series of experiments from predicting the holistic scores and the scores related to formal correctness of the TLT-school test set using manual transcriptions. Table 4.17 shows the results of the two proposed graders and their combinations in terms of PCC, SRC, and MSE.

	Manual transcriptions					
	Holistic			**Formal correctness**		
	PCC	SRC	MSE	PCC	SRC	MSE
GEC	0.849	0.801	5.388	0.745	0.721	0.324
BERT	0.847	0.840	6.120	0.765	0.778	0.315
GEC + BERT (*shallow*)	0.885	0.871	3.984	0.785	0.796	0.280
GEC + BERT (*deep*)	**0.898**	**0.883**	**3.943**	**0.809**	**0.813**	**0.261**

Table 4.17: Results on the TLT-school test set of the GEC-based (**GEC**), the BERT-based (**BERT**), and their combinations on the task of predicting the holistic score and the score related to formal correctness using manual transcriptions.

It can be observed that the results of both graders are aligned on both tasks, but the combinations of the two — especially the *deep* combination — bring significant improvements across all metrics. The β coefficients of the *shallow* combinations shown in Table 4.18 suggest that both graders are contributing almost evenly to the prediction of holistic scores and that the BERT-based grader seems to affect the linear model more than the GEC-based grader for the prediction of the analytic score of formal correctness, but in both cases, there appears to be a certain degree of complementarity. In this regard, Figure 4.8 illustrates the MSE variation across the analytic scores related to formal correctness applying Gaussian kernel smoothing with sigma set to 0.5. The performance of the GEC-based grader in terms of MSE seems to reflect the "inverted U-patterns" highlighted by Hawkins & Buttery (2010) in that the grader appears to predict middle scores (1) more easily than lower (0) and higher (2) scores. On the other hand, the BERT-based grader has a low MSE for high proficiency scores.

	***shallow* - Manual transcriptions**		
	β_{g_c}	β_{b_t}	β_0
Holistic	0.49	0.61	-1.43
Formal correctness	0.30	0.62	0.00

Table 4.18: β coefficients of the GEC-based grader (β_{g_c}) and the BERT-based grader (β_{b_t}), and intercept (β_0) in the *shallow* combination for the holistic and formal correctness scores (manual transcriptions).

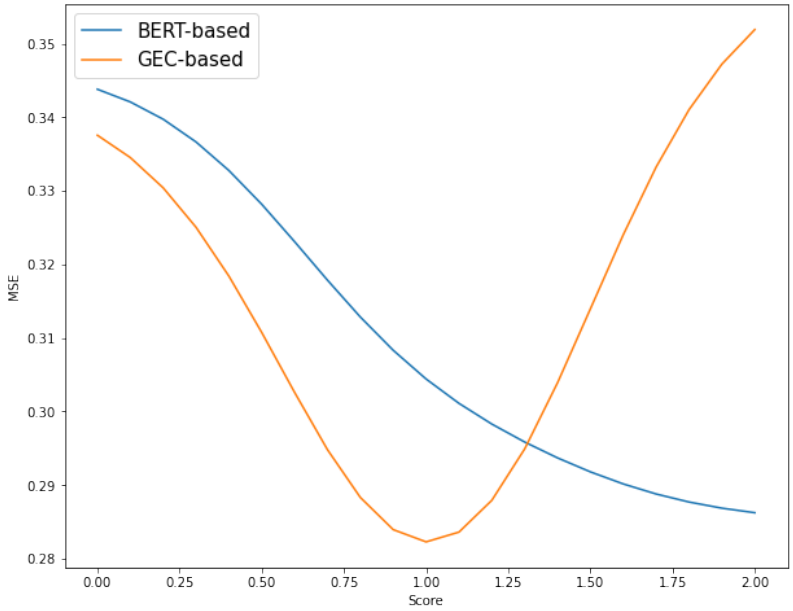


Figure 4.8: MSE variation of the two graders across formal correctness scores (manual transcriptions).

Finally, it should be noted that, although the GEC-based grader does not outperform the BERT-based grader, it should ensure a higher degree of explainability, as it is based on well-defined features such as grammatical violations, whereas the BERT-based grader provides less easily interpretable results.

ASR transcriptions We continue our experiments using the ASR transcriptions of the TLT-school data. Table 4.19 illustrates the results of the GEC-based, the BERT-based graders, and their combinations in terms of PCC, SRC, and MSE.

Unlike the results on manual transcriptions, the GEC-based grader has a better performance than the BERT-based grader on both the task of predicting the holistic score and the task of predicting the formal correctness score. It seems that the significant WER (see Section 4.2.2) is

	ASR transcriptions					
	Holistic			Formal correctness		
	PCC	SRC	MSE	PCC	SRC	MSE
GEC	0.789	0.751	7.436	**0.639**	0.641	**0.431**
BERT	0.697	0.622	10.137	0.586	0.568	0.477
GEC + BERT (*shallow*)	0.797	0.755	7.720	0.633	0.642	0.475
GEC + BERT (*deep*)	**0.817**	**0.775**	**6.279**	0.628	0.640	0.440

Table 4.19: Results on the TLT-school test set of the GEC-based (**GEC**), the BERT-based (**BERT**), and their combinations on the task of predicting the holistic score and the score related to formal correctness using ASR transcriptions.

a double-edged sword: on the one hand, BERT seems to have difficulties in extracting efficient representations from poorly transcribed speech; on the other hand, the GEC-based grader also corrects errors introduced by the ASR module, which are typically caused by pronunciation issues, and leverages them as additional features.

However, this should not lead one to think that the GEC-based grader only depends on errors introduced by ASR errors. A closer look at the categories of errors reveals that the 10 most common ERRANT edit labels are the same for manual and ASR transcriptions except for one (see Figure 4.9). Therefore, there seems to be a certain uniformity between the two systems in terms of grammatical errors. This aspect is highly relevant since it highlights the consistency of grammatical accuracy as a feature to assess L2 proficiency in a fully automated pipeline.

As shown in Figure 4.9, ERRANT also labels error types as `OTHER` when edits do not fall under any other category. A large part of errors labelled as `OTHER` are paraphrases. Their number, especially as regards `U:OTHER`, increases when we use ASR transcriptions, and this is ascribable to ASR errors, of which the spoken GEC system tries to make sense by rearranging words and syntax.

With respect to the edit label `R:SPELL`, most of these errors are caused by code-switching, which is a characteristic aspect of the TLT-school data. For example, L3 German code-switched words such as *interessant* are corrected into *interesting* and labelled as spelling errors, and this also regards ASR transcriptions (Gretter et al., 2019).

In addition to issues related to ASR transcriptions, the presence of such code-switched words might constitute a further problem for the BERT-based grader. We are aware that the limited amount of training data is also a potential issue since BERT-based grading systems are generally efficient when trained on a large quantity of data, even when compared to pronunciation-based graders, as shown in Raina et al. (2020) and in the next studies of this thesis. However, another

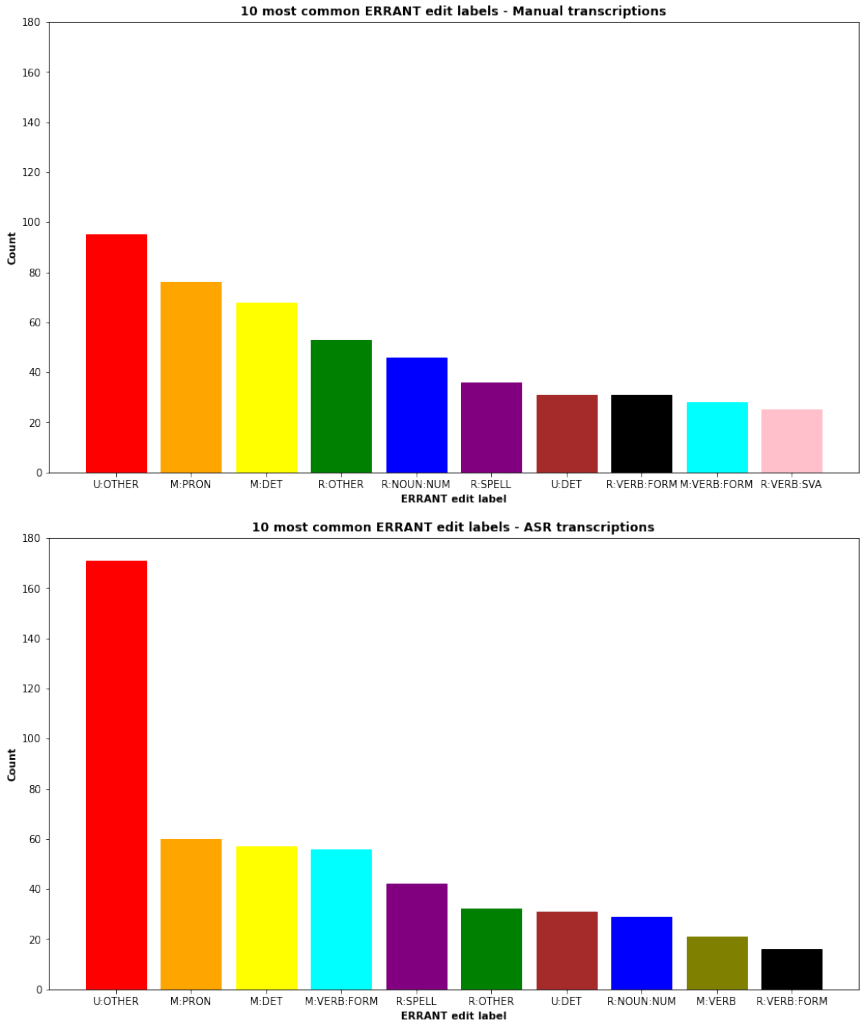


Figure 4.9: Bar charts showing the 10 most common ERRANT edit labels on manual trancriptions (above) and ASR transcriptions (below).

potentially strong point of our GEC-based approach could be its application in scenarios with a limited amount of learner speech data, which are more the rule than the exception in language learning contexts.

Finally, as expected, for the task of predicting holistic scores, the combinations of the two approaches — especially the *deep* combination — bring a modest improvement, whilst, for the task of predicting scores related to formal correctness, the combinations do not bring any further improvement. For completeness, we report the β coefficients of the two graders in the *shallow* combination in Table 4.20. It appears that the BERT-based grader plays a marginal role in both

	***shallow* - ASR transcriptions**		
	$\boldsymbol{\beta_{g_c}}$	$\boldsymbol{\beta_{b_t}}$	$\boldsymbol{\beta_0}$
Holistic	0.80	0.10	0.30
Formal correctness	0.69	-0.04	0.27

Table 4.20: β coefficients of the GEC-based grader (β_{g_c}) and the BERT-based grader (β_{b_t}), and intercept (β_0) in the *shallow* combination for the holistic and formal correctness scores (ASR transcriptions).

tasks.

4.2.6 Conclusions

In this study, we have proposed an approach to L2 proficiency assessment and feedback using grammatical features, and we have illustrated its complete pipeline using only publicly available data for training our modules.

In particular, first, we investigated the impact of DD on spoken GEC. Our DD model achieved results that are aligned with our previous studies, and we found that disfluency removal has a positive impact on GEC.

The second module of our cascaded framework is a spoken GEC system. In this case, we also obtained results that are aligned with previous studies.

In the final part of the pipeline, original answers and automatically corrected answers were passed through ERRANT in order to obtain sequences of GEC edit labels containing information about grammatical accuracy and complexity. These sequences were subsequently fed into a transformer-based model to predict holistic proficiency scores and formal correctness scores.

We compared this grading system to a BERT-based grader and found that the two systems have similar performances when using manual transcriptions. Furthermore, we investigated two types of combinations: a linear regression model fed with the predictions of each grader and a concatenation of the embeddings of the two graders. We found that both combinations — especially the latter — bring significant improvements for both the tasks of predicting holistic proficiency scores and formal correctness scores. A potential concern with the BERT-based grader is that it might not be fully valid alone since its results are not interpretable to provide feedback to a learner. In addition to boosting the assessment performance, the combinations with the GEC-based grader enhance validity and explainability since this is based on clearly defined features.

On the contrary, when using ASR transcriptions, the BERT-based grader obtains lower re-

sults than the GEC-based grader, most likely due to the relatively high WER. For this reason, the GEC-based grader probably leverages certain GEC edit labels, which serve as proxies for pronunciation issues as additional features, although actual grammatical errors do still play a major role.

Further work will consider the application of state-of-the-art end-to-end ASR systems, which should give lower WER and further improve the spoken GEC performances. Specifically for the task of proficiency assessment, the integration of features derived from DD into a grading system could also be explored.

In the next chapter, we consider the GEC-based architecture introduced in this study and apply it to a larger amount of spoken data and a framework which also comprises other aspects of proficiency.

Chapter 5

View-specific assessment

Up to this point, we have mostly focused on grammatical aspects of L2 proficiency and their implications on assessment. In this chapter, we expand our investigation to different facets of proficiency. In particular, our next study addresses a common issue in the field of language assessment: learners' proficiency is typically assessed holistically. Therefore, providing interpretable scores and informative feedback to learners through individual analytic viewpoints of proficiency is still a significant challenge. We investigate whether view-specific systems can be trained when only holistic scores are available. To enable this process, view-specific networks are defined where both their inputs and topology are adapted to focus on specific facets of proficiency. We demonstrate that it is possible to train such systems on holistic scores such that they yield view-specific scores at evaluation time. View-specific networks are designed in this way for pronunciation, rhythm, text, grammatical complexity, and grammatical accuracy.[1]

This study was presented at Interspeech 2022 and UK Speech 2022, and some of its parts have been recently presented by Dr. Kate Knill at the International Symposium on Chinese Spoken Language Processing (ISCSLP 2022) in her keynote speech.[2] The study can be found in Bannò, Balusu, et al. (2022), but additional pieces of analysis are presented in the next paragraphs.

[1] Note that the GEC-based grader used in Study 2 was introduced for the first time in this work. Therefore, the first part of Study 3 mainly focuses on the analysis of grammatical aspects.

[2] `mi.eng.cam.ac.uk/~mjfg/ALTA/presentations/Knill_ISCSLP_Keynote_2022.pdf`

5.1 Study 3: View-specific assessment

5.1.1 Introduction

As shown in Section 1.1, established standards such as the CEFR are recognised throughout the world as effective measures for grading the proficiency of L2 speakers. The CEFR scales are organised according to 'can-do' descriptors of language proficiency outcomes in relation to communicative competence. As a result, these guidelines expect graders to grade proficiency by means of holistic assessments rather than individual aspects. Nonetheless, it has been demonstrated that such holistic judgements do have a modularisable structure, which can be divisible into single aspects of proficiency, such as pronunciation, rhythm, vocabulary, and grammar, each of which is assigned a score that strongly correlates with the holistic grade. One of the first studies on the relationships between holistic and analytic proficiency was conducted by Adams (1980). The impact of five specific aspects (i.e., accent, comprehension, fluency, grammar, and vocabulary) on global proficiency was investigated considering an oral interview. The author reported that fluency, vocabulary, comprehension, and grammar influenced the holistic assessment the most at certain proficiency levels. Similarly, in a study cited several times in Chapter 1, Iwashita et al. (2008) explored the interconnections and relationships between holistic scores and measures related to five facets of proficiency, i.e., vocabulary, pronunciation, fluency, grammatical accuracy, and grammatical complexity. They found that all the considered measures were related to holistic assessment, but especially the measures concerning vocabulary and fluency showed the most significant correlations. In the study by De Jong et al. (2012), the role of various measures related to vocabulary, grammar, pronunciation, and linguistic processing skills in the assessment of functional adequacy of speaking was investigated. It was found that all linguistic skills, except two articulation measures, could explain 76% of the variance. A recent publication by Jeon & In'nami (2022) features various meta-analyses on the connections between the four L2 proficiency skills (i.e., reading, writing, listening, and speaking). In particular, the two contributions by Koizumi et al. (2022) and Jeon et al. (2022) investigated the role of internal (i.e., various measures of fluency, grammatical accuracy and complexity, vocabulary, pronunciation, delivery, content, and coherence) and external (i.e., L2 vocabulary knowledge, L2 grammar knowledge, working memory, reading comprehension, listening comprehension, L2 writing, language aptitude, metacognition, and anxiety) correlates of L2 speaking, respectively. The authors reported that internal features showed a strong correlation to L2 speaking overall and that the strength of the correlations changed depending on each feature. For example, fluency, grammar,

vocabulary, and pronunciation showed strong and highly significant correlations. As regards external features, it was found that variables related to L2 knowledge, as well as those concerning other proficiency skills, were strongly correlated with L2 speaking.

Considering the multifaceted nature of language proficiency, automatic grading can be a valuable resource a fortiori, as it has been suggested that it might be used to make consistent assessments of specific linguistic phenomena, whilst human grading would tend to perform better on more global aspects, as shown in Enright & Quinlan (2010) for written and in Loukina et al. (2015) for spoken proficiency.

Furthermore, some CALL applications distinguish between different views of proficiency during teaching, with different systems used to separately teach specific linguistic skills, such as pronunciation (Thomson, 2011), prosody (Hardison, 2004), and vocabulary (Groot, 2000). As a result, the ability to analytically assess learners' progress according to each of these views should be useful for assessment and feedback and in order to inform further teaching in an adaptive fashion.

In automatic assessment of L2 speaking proficiency, input sequential data from a learner is used to predict a holistic grade (holistic grading) and/or a grade representing proficiency with respect to a specific view (single-view grading). The input may consist, as needed, of acoustic features, recognised words, phones and/or time-alignment information, or other information, such as fundamental frequency, extracted directly from the audio signal or from ASR transcriptions. Most approaches in the literature extract sets of hand-crafted features to capture views, including fluency (Strik & Cucchiarini, 1999), pronunciation (Chen, Evanini, & Sun, 2010), prosody (Coutinho et al., 2016), and text complexity (Bhat & Yoon, 2015), which are then fed into graders, trained with human-annotated single-view scores, to predict single-view scores. Since CEFR descriptors do not provide precise information about the operationalisation of analytic scores, annotated data containing such human-annotated single-view scores are typically hard to obtain and are likely to suffer from inconsistency and incoherence between and within human evaluators. A similar approach can be used for holistic grading by concatenating multiple view-specific hand-crafted features targeting more than one aspect of proficiency in order to produce holistic feature sets, which are then passed through graders in order to predict holistic grades, as shown in Müller et al. (2009), Crossley & McNamara (2013), Wang et al. (2018), and Liu et al. (2020), with the grader trained on human-assigned holistic scores. The efficacy of hand-crafted features for either view-specific or holistic grading heavily relies on their specific underlying assumptions, and they risk discarding potentially salient information about

proficiency. This issue for holistic grading has been addressed by replacing hand-crafted features with automatically derived features for holistic grading prediction, either through an end-to-end system (Chen et al., 2018) or in multiple stages (Takai et al., 2020; Cheng et al., 2020). However, neither can be used for multi-view assessment.

This study investigates whether view-specific systems can be trained when only holistic scores for a test-taker are available.

5.1.2 View-specific training

As previously discussed, for most spoken language assessment training datasets only overall holistic scores are available. Thus, the training dataset comprises $\mathcal{D} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}$ where $\boldsymbol{x}^{(i)}$ is the set of features, or sequence of features, extracted from the audio and ASR system, and $y^{(i)}$ the associate reference score. This section motivates how this training data can be used to train view-specific models.

The assessment process can be divided into two distinct stages, where initially the features $\boldsymbol{x}$ are mapped to view-specific features $\boldsymbol{v}$, and then fed into the score-prediction network. Thus, for a particular view

$$\hat{y}_{\mathrm{v}}^{(i)} = \mathcal{F}_{\mathrm{v}}(\boldsymbol{x}^{(i)}) = f_{\mathrm{v}}(\mathbf{g}_{\mathrm{v}}(\boldsymbol{x}^{(i)})) = f_{\mathrm{v}}(\boldsymbol{v}^{(i)}) \tag{5.1}$$

where the desired training data comprises $\mathcal{D}_{\mathrm{v}} = \{\boldsymbol{x}^{(i)}, y_{\mathrm{v}}^{(i)}\}$. Unfortunately, in our case study, as in many cases, there are no view-specific reference grades, $y_{\mathrm{v}}^{(i)}$, associated with each of the training observations, $\boldsymbol{x}^{(i)}$, but only overall holistic grades, $y^{(i)}$. To address this problem, the form of the feature extractor $\mathbf{g}_{\mathrm{v}}(\boldsymbol{x}^{(i)})$ is constrained so that only information about a specific view is contained within $\boldsymbol{v}^{(i)}$ (see Figure 5.1).

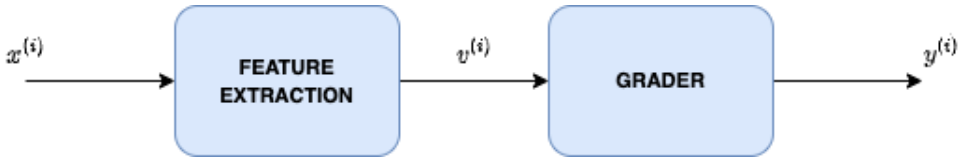


Figure 5.1: View-specific training.

For example, if only information about the text spoken is in $\boldsymbol{v}^{(i)}$, irrespective of the pronunciation of the words,[3] then the same feature vector $\boldsymbol{v}$ can be obtained from the different values

[3]There will be some influence of pronunciation on the performance of the ASR system and the respective confidence scores.

of $\boldsymbol{x}$.

Training the model parameters, $\boldsymbol{\theta}$, on the holistic training data, $\mathcal{D}$, aims to minimise the loss $\mathcal{L}(\boldsymbol{\theta})$:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \mathcal{L}(y^{(i)}, \mathcal{F}_{\mathrm{v}}(\boldsymbol{x}^{(i)})) = \sum_{j=1}^{\tilde{N}} \sum_{i \in \mathcal{S}^{(j)}} \mathcal{L}(y^{(i)}, f_{\mathrm{v}}(\boldsymbol{v}^{(j)})) \tag{5.2}$$

where $\mathcal{S}^{(j)}$ is the set of samples such that $\boldsymbol{g}_{\mathrm{v}}(\boldsymbol{x}^{(i)}) \approx \boldsymbol{v}^{(j)}$ and $\tilde{N}$ is the number of distinct values of $\boldsymbol{v}$. In this work, a least-squares cost function, $\mathcal{L}(y, \hat{y}_{\mathrm{v}})$, is used. When training the model it is not necessary for the loss function to be 'correct' provided that the gradients for training the model parameters are suitable. Thus,

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &\propto \sum_{j=1}^{\tilde{N}} \sum_{i \in \mathcal{S}^{(j)}} \left(f_{\mathrm{v}}(\boldsymbol{v}^{(j)}) - y^{(i)} \right) \frac{\partial f_{\mathrm{v}}(\boldsymbol{v}^{(j)})}{\partial \boldsymbol{\theta}} \\ &= \sum_{j=1}^{\tilde{N}} |\mathcal{S}^{(j)}| \left(f_{\mathrm{v}}(\boldsymbol{v}^{(j)}) - \frac{\sum_{i \in \mathcal{S}^{(j)}} y^{(i)}}{|\mathcal{S}^{(j)}|} \right) \frac{\partial f_{\mathrm{v}}(\boldsymbol{v}^{(j)})}{\partial \boldsymbol{\theta}} \end{aligned} \tag{5.3}$$

Thus, the gradient, and associated minima, will be consistent with training against view-specific training data $\mathcal{D}_{\mathrm{v}}$ provided

$$y_{\mathrm{v}}^{(j)} \approx \frac{\sum_{i \in \mathcal{S}^{(j)}} y^{(i)}}{|\mathcal{S}^{(j)}|} \tag{5.4}$$

Here it is assumed that each view-specific score contributes to the overall holistic score. By averaging over samples with similar view-specific features, $\boldsymbol{v}$, the resulting scores should be biased to the view-specific grades even if (5.4) is not exactly satisfied.

In this analysis, the precise concept of how the set $\mathcal{S}^{(j)}$ is derived has not been strictly specified. Assuming that there is sufficient data and the $\mathbf{g}_{\mathrm{v}}()$ is a smooth function, the standard training, the LHS expression in (5.2), can be run. The model implicitly smooths the view-specific predictions.

5.1.3 Single-view graders

In the present study, we implement 5 grading models for as many views of proficiency, namely pronunciation, rhythm, text, grammatical accuracy, and grammatical complexity. For all graders,

an ensemble of 10 models was trained.[4]

Pronunciation

The pronunciation model is described in detail in the study by Kyriakopoulos et al. (2018) (see also Section 2.2.2). Sequences of acoustic observations corresponding to phone instances are projected to fixed-length phone instance representations, with those corresponding to a specific phone label attended over to obtain an overall representation for that phone. Subsequently, Euclidean distances between phone representations are passed through a feed-forward layer in order to predict the score. The objective is for information from the observation vectors to only be preserved insofar as it characterises the way the speaker pronounced each phone compared to the pronunciation of the other phones.

Rhythm

The rhythm grader is implemented as described in the work by Kyriakopoulos et al. (2019) (see also Section 2.2.2). In this case, the grader is constrained in a way that the input only consists of durations of phones and silences, grouped into consonant and inter-consonant intervals. In this way, the grader can only leverage duration patterns for scoring.

Text

For the text grader, presented in Raina et al. (2020), we used BERT to extract word embeddings, followed by a multi-head self-attention mechanism. The output of this process is subsequently fed into a feed-forward network. The parameters of the pre-trained BERT model[5] were also fine-tuned.

Grammatical accuracy

The grader based on GEC edit sequences — which we refer to as *es* (edit sequence) hereafter — is a transformer-based model that takes GEC edit sequences as inputs in the same fashion as the GEC-based grader introduced in Study 2 (see Section 4.2). Prior to the grader, a GEC model is run on the ASR transcriptions after removing hesitations and partial words. Both corrected

[4] Note that the graders for text, grammatical accuracy, and grammatical complexity consist, in turn, of multiple graders trained on the scores of the 5 parts that compose the exam, whereas the pronunciation and rhythm graders have been trained on the overall scores of the exam.

[5] `huggingface.co/bert-base-uncased`

and original ASR texts are passed through ERRANT (Bryant et al., 2017) to obtain the GEC edit sequences.

Grammatical complexity

The grader based on POS tag sequences — *pos* (part-of-speech) herefter — has the same transformer-based architecture as the es grader but takes POS tag sequences as inputs. These sequences are generated with spaCy.[6] Figure 5.2 shows an example drawn from the data of text, GEC edit, and POS sequences. As can be seen, although the GEC edit sequences contain some information about grammatical complexity, this is the aspect that characterises the pos grader the most, all the more if we consider that all tokens — correct and incorrect — are labelled with their respective part of speech. On the other hand, only the tokens marked as incorrect have an informative label in the es grader.

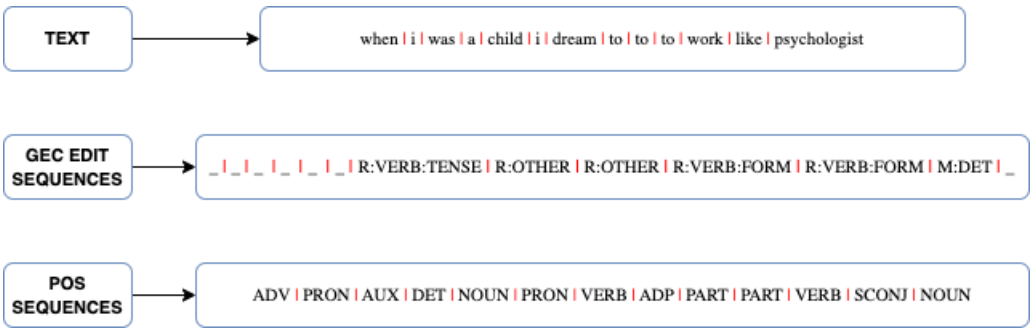


Figure 5.2: Example of text, GEC edit, and POS sequence.

In addition to a comparative analysis of the single-view graders, we investigate a possible combination by means of an OLS multiple linear regression model using the 5 graders' predictions $\hat{y}_{\mathrm{v}}^{(i)}$ as predictors and setting the reference holistic score $\hat{y}^{(i)}$ as target:

$$\hat{y}^{(i)} = \beta_0 + \beta_{\mathrm{p_r}}\hat{y}_{\mathrm{p_r}}^{(i)} + \beta_{\mathrm{r_y}}\hat{y}_{\mathrm{r_y}}^{(i)} + ... + \beta_{\mathrm{p_s}}\hat{y}_{\mathrm{p_s}}^{(i)} + \epsilon$$

where β_0 represents the intercept and β_{v} is the coefficient for a specific view prediction $\hat{y}_{\mathrm{v}}$ and ϵ is the model's residual (see Table 5.4 for notation). The linear model is trained on the development set. The performance of the single-view graders is compared against a baseline assessment system, which is a Deep Density Network (DDN) trained on a set of hand-crafted features related to different views of proficiency (Malinin et al., 2017). These features include grade-

[6] `spacy.io`

dependent language model and word level statistics, statistics of phone duration, statistics to capture rhythm, fluency metrics, and fundamental frequency statistics. As for the other graders, the baseline predictions are the result of an ensemble of 10 models. Further information about the features employed and about the ensemble approach can be found in Wang et al. (2018) and Wu et al. (2020).

5.1.4 Data and experimental setup

The data used in our experiments are obtained from candidate responses to the spoken components of the Linguaskill examinations for L2 learners of English, provided by Cambridge English Language Assessment (Ludlow, 2020) (see Appendix B for examples of question prompts). Each speaker is graded on a scale ranging from 1 to 6 based on the CEFR. Non-overlapping datasets of 31475 and 1033 speakers are used respectively as the training and development/calibration set. For evaluation, we consider two test sets, LinGen and LinBus, of 1049 and 712 speakers, respectively. Further details about the dataset can be found in Section 3.2.3.

The first step before passing the data through each automatic assessment system is recognising the text being spoken and, for the pronunciation and rhythm grading systems, aligning the audio to a sequence of phones. Both tasks are performed using an ASR system which is very similar to that in Lu, Gales, Knill, Manakul, Wang, & Wang (2019), i.e., a lattice-free maximum mutual information (LFMMI) factorised time-delay neural network (TDNN-F) system trained on approximately 500 hours of L2 English data, mostly from Linguaskill Business (BULATS) exams with over 50 L1s included. A succeeding word recurrent neural network language model (su-RNNLM) trained on 25.6M in-domain words is used for rescoring. The average WER is $\sim$20%.

In order to generate the automatically corrected versions from the original ASR texts, we use the transformer-based GEC system described in Lu et al. (2022). As mentioned in Section 4.2, it is trained on the CLC (Nicholls, 2003) (see Section 3.2.1) and BEA-2019 data (Bryant et al., 2019) (see Section 3.1.1). It is a base-sized model (Vaswani et al., 2017) with 512D hidden states, 6 encoder, and 6 decoder layers. The vocabulary is derived from CLC and Switchboard (Meteer et al., 1995) (see Section 3.2.3). Model parameters are averaged over 5 best checkpoints, and greedy decoding is used. We train the model using the Adam optimiser (Kingma & Ba, 2015) with batch size 256, dropout 0.2, and learning rate 1e-3. The GEC edit sequences are derived from ERRANT run on the original and automatically corrected ASR hypotheses. These sequences are fed into our es model which consists of an embedding layer with size 128, a transformer-block

with hidden layer size 128 and 8 heads, a dense layer of 128 nodes, and finally the output layer. To train the system, we use the Adam optimiser with batch size set to 32 and learning rate to 2e-6. The pos grader model has the same architecture.

The performance of each grading system is evaluated using root-mean-square error (RMSE), whilst further comparisons also include PCC, SRC, and the percentage of the predicted scores that are equal to or lie within 0.5 (i.e., within half a grade) of the actual score (% ≤ 0.5).

Error analysis

Similarly to what we did in Study 2, we investigate the impact of GEC on ASR transcriptions. Compared to our previous study, in this work we use a much larger dataset which includes the full range of proficiency levels from A1 to C2.

A subset of about 3000 sentences was manually transcribed, annotated with disfluencies and grammatical error corrections. Therefore, we used it to evaluate the performance of the GEC model on ASR transcriptions. Table 5.1 shows the performance of the GEC system on manual transcriptions (MAN+GEC) and ASR transcriptions (ASR+GEC) in terms of M^2. We also consider its performance on ASR transcriptions with disfluencies automatically removed for further comparison.

	M^2
MAN+GEC	37.47
ASR+GEC	17.11
ASR+DD+GEC	19.00

Table 5.1: Comparison of the performance of the GEC model on manual and ASR transcriptions in terms of M^2.

As can be observed, DD has only a small impact on GEC in this case. Therefore, we decided not to consider using this module in our pipeline for this set of experiments. These results might not seem encouraging if our main and only goal were GEC, but since our primary aim is proficiency assessment, we are mostly interested in the distribution of information related to GEC edit labels. Figure 5.3 reports the 5 most common ERRANT edit labels across three systems: manual transcriptions manually corrected, manual transcriptions automatically corrected, and ASR transcriptions automatically corrected.

The 5 most common ERRANT edit labels are the same across the three systems, and this means that there is a certain consistency with respect to this type of information, even when

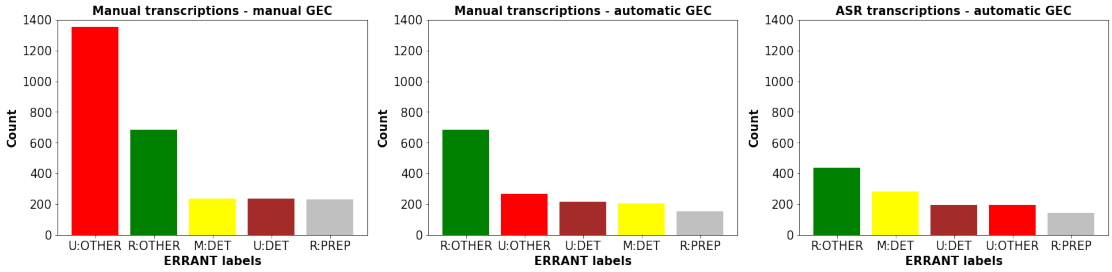


Figure 5.3: 5 most common ERRANT edit labels across three systems: manual transcriptions manually corrected, manual transcriptions automatically corrected, and ASR transcriptions automatically corrected.

using a fully automated pipeline.

As we did in Study 2, we also investigate whether the distributions of edit labels change across proficiency levels. Therefore, we take the ERRANT labels and convert them into a smoothed distribution, by applying add-one smoothing, and we compute the symmetric KL-Divergence. Therefore, for edit label t_i for level L_k:

$$\mathrm{P}(t_\mathrm{i}|L_\mathrm{k}) = \frac{\mathrm{cnt}(t_\mathrm{i}, L_\mathrm{k}) + 1}{\sum_{j=1}^{N}(\mathrm{cnt}(t_\mathrm{i}, L_\mathrm{k}) + 1)}$$

where $\mathrm{cnt}(t_\mathrm{i}, L_\mathrm{k})$ is the number of occurrences for a given label at a given level.

The symmetric KL Divergence is calculated across proficiency levels:

$$\mathrm{KL}(L_\mathrm{k}|L_\mathrm{l}) = \left(\sum_{i=1}^{N} \mathrm{P}(t_\mathrm{i}|L_\mathrm{k}) \log\left(\frac{\mathrm{P}(t_\mathrm{i}|L_\mathrm{k})}{\mathrm{P}(t_\mathrm{i}|L_\mathrm{l})}\right)\right) + \left(\sum_{i=1}^{N} \mathrm{P}(t_\mathrm{i}|L_\mathrm{l}) \log\left(\frac{\mathrm{P}(t_\mathrm{i}|L_\mathrm{l})}{\mathrm{P}(t_\mathrm{i}|L_\mathrm{k})}\right)\right)$$

In Table 5.2, we report the symmetric KL-Divergence between distributions of counts from all the ERRANT edit labels in manually annotated subset across CEFR proficiency levels.

	A1	A2	B1	B2	C
A1	0.0	0.419	0.459	0.570	0.556
A2	0.419	0.0	0.066	0.102	0.105
B1	0.459	0.066	0.0	0.061	0.073
B2	0.570	0.102	0.061	0.0	0.053
C	0.556	0.105	0.073	0.053	0.0

Table 5.2: Symmetric KL Divergence between distributions of counts from all ERRANT edit labels in the manually annotated subset across proficiency levels.

We also compute it considering the ASR transcriptions of the full training set and report the

results in Table 5.3.

	A1	A2	B1	B2	C
A1	0.0	0.067	0.362	0.808	1.307
A2	0.067	0.0	0.139	0.472	0.869
B1	0.362	0.139	0.0	0.103	0.327
B2	0.808	0.472	0.103	0.0	0.074
C	1.307	0.869	0.327	0.074	0.0

Table 5.3: Symmetric KL Divergence between distributions of counts from all ERRANT edit labels in the full training set (ASR transcriptions) across proficiency levels.

Similarly to what we observed in Study 2, we can see that there are differences in the distributions of GEC edit labels across proficiency levels also in this case. These results corroborate the hypothesis that grammatical errors can be considered criterial features of linguistic proficiency and can be used in an automatic assessment system.

Once we finish error analysis, we can shift our focus back to the main objective of this study, i.e., view-specific proficiency assessment.

5.1.5 Experimental results and analysis

Table 5.4 shows the performance of the 5 single-view graders and the baseline in terms of RMSE, considering both the individual models and the ensembles. As can be observed, the ensemble approach brings a significant improvement on all the grading systems, including the baseline.

Model	LinGen		LinBus	
	Indiv.	Ens.	Indiv.	Ens.
baseline	$0.578_{\pm 0.011}$	0.412	$0.522_{\pm 0.009}$	0.406
pron ($\mathtt{p_r}$)	$0.455_{\pm 0.004}$	0.452	$0.454_{\pm 0.003}$	0.451
rhythm ($\mathtt{r_y}$)	$0.571_{\pm 0.036}$	0.508	$0.551_{\pm 0.037}$	0.490
text ($\mathtt{t_x}$)	$0.402_{\pm 0.005}$	0.400	$0.409_{\pm 0.007}$	0.409
es ($\mathtt{e_s}$)	$0.547_{\pm 0.001}$	0.547	$0.497_{\pm 0.001}$	0.495
pos ($\mathtt{p_s}$)	$0.550_{\pm 0.001}$	0.550	$0.499_{\pm 0.003}$	0.497

Table 5.4: Performance of the single-view graders and baseline in terms of RMSE. Individual models VS ensembles.

To explore the differences between and the complementarity of each single-view grader, we only consider LinBus. In Table 5.5, we report the performance of various combinations of the single-view graders through the OLS multiple linear regression model introduced in Section 5.1.3.

We report the respective β coefficient for each component. It can be observed that the combination of all the graders improves on the performance of their individual component graders, and this result is consistent with the single-view graders extracting information which is complementary to each other. In particular, among the 5 graders, the text grader affects the linear model the most, as can be inferred from its high β coefficient and from the drop in performance in the combination that does not include it. Based on the β coefficients, the pronunciation and rhythm graders always contribute equally to the linear model, but the presence of the first appears to have a more positive impact on the overall performance. The es grader seems to have a relatively smaller impact, except when the combinations exclude the pos or the text graders. We continue our analysis focusing on the performance of each grader across proficiency levels.

Combination	$\beta_{\mathtt{p_r}}$	$\beta_{\mathtt{r_y}}$	$\beta_{\mathtt{t_x}}$	$\beta_{\mathtt{e_s}}$	$\beta_{\mathtt{p_s}}$	RMSE
$\mathtt{p_r r_y t_x e_s p_s}$	0.14	0.14	1.30	-0.05	-0.39	0.386
$\mathtt{p_r r_y t_x e_s}$	0.14	0.14	1.30	-0.31	—	0.405
$\mathtt{p_r r_y t_x \ p_s}$	0.14	0.14	1.30	—	-0.47	0.384
$\mathtt{p_r r_y \ e_s p_s}$	0.45	0.45	—	0.28	-0.05	0.432
$\mathtt{p_r \ t_x e_s p_s}$	0.29	—	1.30	-0.05	-0.39	0.385
$\mathtt{r_y t_x e_s p_s}$	—	0.29	1.30	-0.05	-0.39	0.392

Table 5.5: RMSE and β coefficients of linear regression model with different combinations.

Figure 5.4 shows the RMSE variation of the 5 graders across the 5 proficiency levels. The es and pos graders follow very similar trends as expected since ERRANT labels are based on POS tags. In particular, as already observed in Study 2, the case of es is also consistent with "inverted U-patterns" in written proficiency (Hawkins & Buttery, 2010), i.e., errors increase after B1 and then decline again by C2. In this regard, it is interesting to note that there is a correspondence between oral and written proficiency when it comes to grammatical accuracy. Compared to the other graders, the pronunciation grader has the lowest RMSE on the lowest grade (1), which gradually decreases until grade 4 and then rises again after grade 5. This is also consistent with intelligibility being a key aspect of lower levels of proficiency. For example, the descriptors of CEFR level A1 are mainly concerned with intelligibility (Council of Europe, 2001, 2020). On the other hand, the rhythm grader shows its best performance for grade 5, and this is consistent with the findings shown in Taylor (1981), in which English speech rhythm is described as one of the most difficult aspects for learners to acquire. Finally, the text grader shows the lowest RMSE, in both absolute and relative terms, for the middle grades (3-4).

Furthermore, we investigate the relationships between single-view graders through a repeated

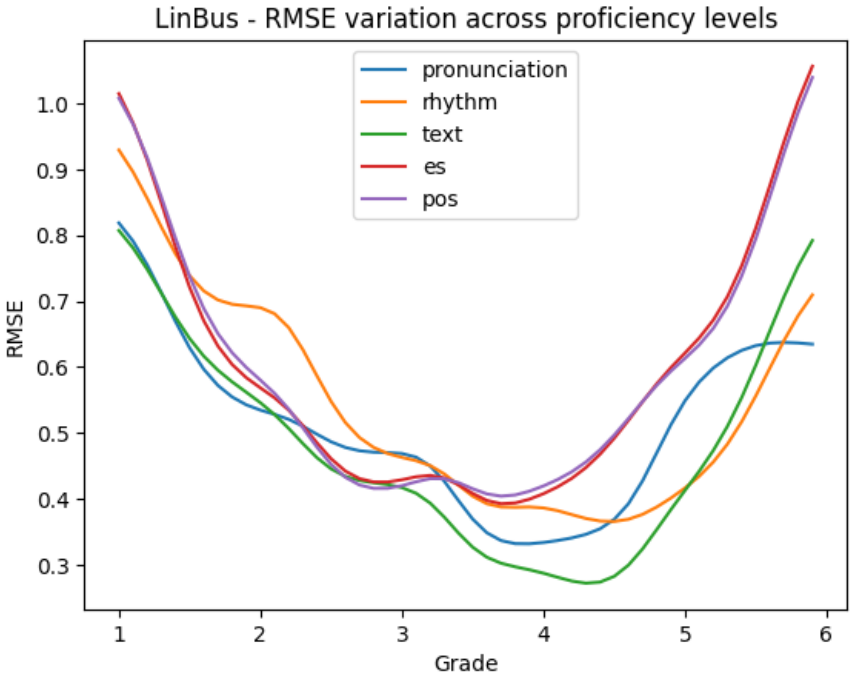


Figure 5.4: RMSE variation across proficiency levels.

measures design. Arguably, the most well-known repeated measures design is repeated measures analysis of variance (rANOVA), but since our data violate both the normality and sphericity assumptions required for rANOVA, we must perform the Friedman test (Friedman, 1937), which is the non-parametric equivalent of rANOVA and determines whether there are any statistically significant differences in ranks between the distributions of multiple paired groups. As we obtain a significant p-value, we find that there are significant differences among the graders.

In order to determine exactly which graders are significantly different, we perform post-hoc multiple comparisons using the Nemenyi test (Nemenyi, 1963). We report the test results in Figure 5.5. All paired comparisons, even those with the reference score, show significant differences (p-value<0.05), with the exception of the pairs es-pos and text-rhythm. As regards the first pair, we have already commented on the almost overlapping trends shown in Figure 5.4. As regards the latter, we might argue that the non-significant p-value should reflect the analogous trends of the RMSE variation curves followed by the text grader and the rhythm grader, despite a considerable gap between them.

Finally, in Table 5.6, we report a comparison of the baseline, our best-performing single-view model, i.e., the text grader, and the linear regression model considering the evaluation metrics

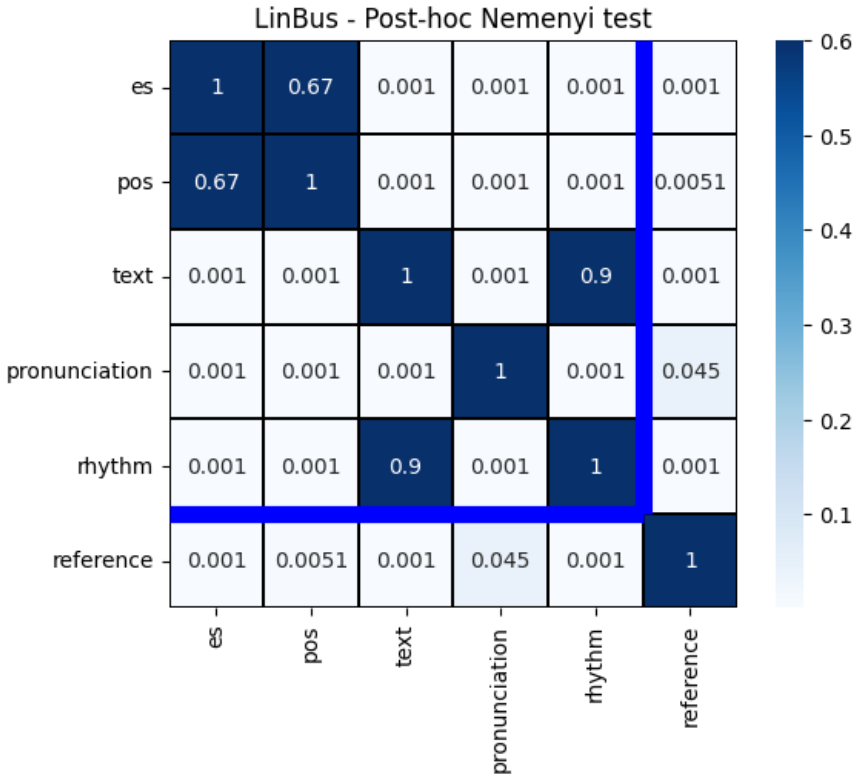


Figure 5.5: Heatmap of the results of the post-hoc Nemenyi test.

mentioned in Section 5.1.4. The combination of the single-view graders outperforms both the baseline and the text grading system across all metrics.

Model	PCC	SRC	RMSE	%≤0.5
baseline	0.910	0.915	0.406	79.1
text ($\mathtt{t_x}$)	0.920	0.925	0.409	78.9
$\mathtt{p_r r_y t_x e_s p_s}$	0.920	0.926	0.386	80.5

Table 5.6: Comparison of the performance of the baseline, text grader, and linear regression model.

5.1.6 Conclusions

In order for CALL and automatic spoken language assessment systems to give learners interpretable and explainable scores and informative feedback on their speaking ability, specific facets of their proficiency should be assessed, but for many real-world tasks, analytic scores on specific

aspects of proficiency are not available or often suffer from inconsistencies and cannot be used for training automatic systems.

This study considers whether view-specific systems can be trained when only holistic scores are available. Single-view graders are developed for views of pronunciation, rhythm, text, grammatical accuracy, and grammatical complexity. The predictions made by these graders are seen to be complementary to all the others for the task of predicting holistic grades. Moreover, the relationships between the single-view predictions are consistent with what would be expected for the respective views they are assessing. For example, the pronunciation model achieves better results on low grades, whereas the rhythm grader performs better on high grades, as shown in Figure 5.4.

Furthermore, we investigate a combination of the 5 graders by means of a multiple linear regression model, and we find that it generally improves on the performance of each single-view grader. Since the single-view scores are also made available, this multi-view system makes the holistic score significantly more interpretable by enabling useful feedback to learners who need specific indications on how to improve their speaking skills.

However, this point also raises some questions about the implementation of the assessment of the communicative approach to speaking proficiency in automatic systems since our study only considered some of its underlying competences. While linguistic competence (see Section 1.2.1) was well represented by the pronunciation (for segmental pronunciation), rhythm (for suprasegmental pronunciation), es and pos (for grammatical accuracy and complexity), and — to a certain extent — by BERT (for vocabulary), the sociolinguistic (see Section 1.2.2) and pragmatic (see Section 1.2.3) competences were not specifically investigated. The BERT-based grader might leverage aspects related to idiomaticity (sociolinguistic competence) and coherence and cohesion (pragmatic competence), but, as stated several times in this thesis, its results are not fully explainable in this sense.[7] In order to provide more informative feedback about communicative competence, future work should integrate the assessment of all its underlying competences.

Further work should also be undertaken in order to improve the performance of spoken grammatical error annotation since current systems are generally designed for written texts and are not ideal for speech. With respect to annotations, another element that should be investigated is the integration of syntactic information in addition to strictly grammatical information. We

[7] As regards idiomaticity, in Section 2.3, we have already mentioned the use of BERT for metaphor detection (e.g., in Mao et al. (2019)). With respect to coherence and cohesion, we remind the reader that one of the training strategies behind BERT is next sentence prediction. Therefore, it should possess this type of knowledge.

also plan to include other types of combinations, considering both shallow and deep combination methods.

Chapter 6

SSL-based assessment

In this chapter, we investigate the use of SSL speech representations — specifically wav2vec 2.0 — for proficiency assessment. In this way, features are extracted directly from the audio signal in a self-supervised fashion.

In Study 4, we investigate the use of wav2vec2-based graders on ICNALE and TLT-school, considering both holistic and analytic scores. We compare their performance to BERT-based graders, and we explore potential combinations of the two graders.

In Study 5, we conduct similar experiments on the Linguaskill multi-part exam, comparing the wav2vec2-based graders to BERT-based and handcrafted-features-based graders. We also investigate potential combinations of the three approaches.

Study 4 has been recently presented at the 2022 IEEE Spoken Language Technology Workshop (SLT 2022) and can be found in Bannò & Matassoni (2023), whereas Study 5 has been submitted to Interspeech 2023 and its pre-print version can be found in Bannò, Knill, et al. (2022).

6.1 Study 4: Speaking assessment using wav2vec 2.0 (Part 1)

6.1.1 Introduction

As said in Section 5.1.1, most approaches in the literature concerning automatic speaking assessment extract sets of hand-crafted features with respect to specific aspects of proficiency which

are then fed into grading systems to predict analytic view-specific scores targeting those specific aspects (Strik & Cucchiarini, 1999; Chen, Evanini, & Sun, 2010; Coutinho et al., 2016; Bhat & Yoon, 2015). Similarly, multiple hand-crafted features targeting more than one aspect of proficiency can be concatenated to create holistic feature sets, which are then fed to graders to predict holistic proficiency scores (Müller et al., 2009; Crossley & McNamara, 2013; Wang et al., 2018; Liu et al., 2020). However, we have already mentioned that the efficacy of such hand-crafted features for grading either individual aspects or overall proficiency heavily relies on their specific underlying assumptions, and they risk discarding potentially salient information about speaking proficiency. For holistic grading, this limitation has been tackled by substituting hand-crafted features with automatically derived features (Chen et al., 2018; Takai et al., 2020; Cheng et al., 2020). Other studies have used grading systems that are trained on holistic grades but are defined with both their inputs and topology adapted to focus exclusively on specific aspects of proficiency, such as pronunciation (Kyriakopoulos et al., 2018), rhythm (Kyriakopoulos et al., 2019), and text (Wang et al., 2021; Raina et al., 2020). In these cases, when scoring holistic proficiency, a possible limitation might be the missing information about aspects of proficiency that are not featured in the input data fed to the grading system, although, in Study 3 (see Section 5.1), we have shown that it is possible to combine multiple grading systems targeting different aspects of proficiency. This issue is especially true for systems using ASR transcriptions for at least two reasons: first, they suffer from a certain WER and may not faithfully represent a learner's performance; secondly, although transcriptions might preserve some information about pronunciation (e.g., in the ASR confidence scores), they do not yield any information about other crucial aspects of a learner's performance, such as suprasegmental aspects (e.g., rhythm, intonation, or prosody). Instead, transcriptions remain an essential resource for highly specific tasks in CALL applications, such as spoken GEC and feedback, as we showed in Study 2 (see Section 4.2) and in Lu et al. (2022).

In this study, to address these issues and limitations, we propose an SSL-based approach using wav2vec 2.0. As we mentioned in Section 2.2.2, recent studies have demonstrated the efficacy of SSL in multiple downstream tasks, such as ASR, emotion recognition, keyword spotting, speaker identification, and speaker diarisation. These studies applied contextual representations by means of pre-trained models. Specifically, it has been shown that such models are capable of capturing a wide range of speech-related features and linguistic information, such as audio, fluency, suprasegmental pronunciation, and even syntactic and semantic text-based features for L1, L2, read and spontaneous speech (Singla et al., 2022). In the context of CALL, SSL has

been applied to mispronunciation detection and diagnosis (Peng et al., 2021; Wu et al., 2021; Xu et al., 2021) and automatic pronunciation assessment (Kim et al., 2022), but, to the best of our knowledge, it has not been investigated for the assessment of holistic spoken proficiency nor other specific aspects of proficiency, such as formal correctness, communicative effectiveness, lexical richness and complexity, and relevance, before this study.

In this work, we first test the efficacy of wav2vec 2.0 for the task of predicting the holistic proficiency level of L2 English learners' responses included in the ICNALE monologues. Subsequently, we do the same on the TLT-school data, which also contain annotations related to individual aspects of proficiency that we attempt to predict with specific graders. The baseline system employed for comparison is a BERT-based grader fed with transcriptions. We use only manual transcriptions for our experiments on ICNALE, whereas we use both manual and ASR transcriptions for our experiments on TLT-school. Specifically, the manual transcriptions also include hesitations and truncated words, which serve as proxies for pronunciation and fluency.

6.1.2 Data

ICNALE

To test our approach, we consider ICNALE, a publicly available dataset including written and spoken answers of English learners ranging from A2 to B2 of the CEFR and partially of L1 speakers. We described the corpus in detail in Section 3.1.2.

For this set of experiments, we only considered the monologues, i.e., 4332 answers lasting between 36 and 69 seconds in which learners are required to express their opinion about the following two statements:

- *It is important for college students to have a part-time job.*
- *Smoking should be completely banned at all the restaurants in the country.*

The available metadata include the manual transcriptions of the learners' responses, personal data about learners' education history, and their assigned CEFR levels. We split the data into a training set of 3898 answers, and a development set and a test set containing 217 answers each. For the experiments on this dataset, proficiency assessment is implemented as a classification task with five classes: A2, B1_1, B1_2, B2, and L1 speakers (see Table 6.1). To the best of our knowledge, the ICNALE monologues have only been used in the study by Zhou et al. (2019), but in that study, the responses to the two statements reported above were considered and evaluated

independently, so no comparison is possible. The experiments described in Study 1, instead, only include a section of essays and dialogues, but no monologues.

	Train	**Dev**	**Test**	**Total**
A2	299	16	17	332
B1_1	792	44	44	880
B1_2	1681	94	93	1868
B2	586	33	33	652
L1	540	30	30	600
Total	3898	217	217	4332

Table 6.1: Number of answers for each CEFR proficiency level in ICNALE.

TLT-school

We have described the TLT-school data in Section 3.2.2. Note that for this set of experiments, we only kept the short answers of the B1 section. Therefore, the data considered in this set of experiments is composed of 494 responses.

Annotations related to spontaneous speech phenomena (e.g., hesitations and truncated words) were not eliminated from the manual transcriptions in order not to lose any possibly existing information about strictly speech-related aspects, such as pronunciation and fluency, although we acknowledge that they cannot substitute the role of actual speech phenomena completely.

With respect to the ASR transcriptions, its WER is 41.13%. As mentioned already, acoustic and language models are described in Gretter et al. (2019), which also includes further details about the training data used for ASR development.

As mentioned in Section 3.2.2, the total score ranges from 0 to 12 and consists of the sum of the analytic subscores for each specific proficiency indicator assigned by the human raters (i.e., relevance, formal correctness, lexical richness and complexity, pronunciation, fluency, and communicative effectiveness). For each indicator, human raters could choose 0, 1 or 2 points.

For this dataset, we treated proficiency assessment as a regression task when predicting both the holistic score from 0 to 12 and the analytic subscores ranging from 0 to 2.

Before starting our experiments, we investigated the relationships between these subscores with a repeated measures design to verify that they are effectively targeting different aspects of proficiency. Since both the sphericity and normality assumptions required for rANOVA were not met, we performed the Friedman test (Friedman, 1937). As we obtained a significant p-value, we found significant differences among the subscores. To identify exactly which pairs of subscores are

significantly different, we ran post-hoc multiple comparisons using the Nemenyi test (Nemenyi, 1963) (see Figure 6.1). All paired comparisons show significant differences (p-value<0.05), with the exception of the pairs formal correctness-lexical richness and complexity, formal correctness-pronunciation, pronunciation-communicative effectiveness, and fluency-communicative effectiveness. The absence of significant differences between the subscores related to formal correctness and those related to lexical richness and complexity seem to be consistent with the fact that: a) a poorer and simpler lexis should contain fewer errors and b) in some cases, the human evaluator may have confused and overlapped the two indicators due to the presence of lexical errors, which can be linked either to poor formal correctness or bad use of vocabulary.

Similarly, pronunciation and correctness do not show significant differences because pronunciation errors might have been incorporated in one or the other indicator.

Finally, the score targeting communicative effectiveness intersects almost by definition with those related to pronunciation and fluency. This has been especially true in recent years since pronunciation tends to be assessed in terms of general goals such as intelligibility (Levis, 2018) and communicative effectiveness (Pennington & Rogerson-Revell, 2019a) rather than closeness to L1 English (see Section 1.2.1). Apart from all these considerations, we also have to consider that the halo effect (Myford & Wolfe, 2003) might bias the scores to a certain degree.

We split the data into three sets: a training set of 322, a development set of 85, and a test set of 87 samples.

	Train	**Dev**	**Test**	**Total**
0-3	74	14	27	115
3-6	73	20	17	110
6-9	77	20	11	108
9-12	98	31	32	161
Total	322	85	87	494

Table 6.2: Number of answers for each score range in TLT-school.

6.1.3 Model architectures

wav2vec2-based graders

Wav2vec 2.0 encodes the speech audio signal through a multilayer convolutional neural network (CNN). After encoding, masking is applied to spans of the resulting latent representations, which are fed into a transformer in order to provide contextualised representations. Gumbel softmax is

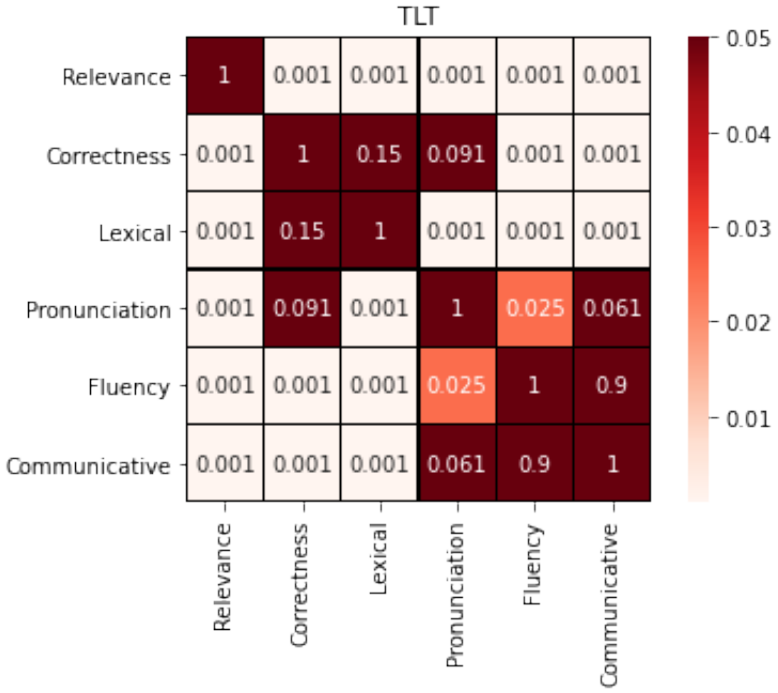


Figure 6.1: Heatmap of the results of the post-hoc Nemenyi test on the analytic subscores of the TLT-school dataset.

used to compute the contrastive loss on which the model is trained, and speech representations are learned from this training. For the models used in our experiments, we initialised the configuration and processor from a version provided by the HuggingFace Transformer Library (Wolf et al., 2020).[1] After the learners' responses are fed into the model, wav2vec 2.0 provides contextualised representations. To handle representations of various audio lengths, we use a mean pooling method to concatenate 3D representations into 2D representations. Next, these are projected to a Dense layer of 768 units, a Dropout layer and, finally, through an output layer. We tried different architectures and hyperparameters, and, finally, we chose those described in the following paragraphs.

ICNALE For our experiments on the ICNALE data, the task is multi-class classification. Therefore, the output layer has 5 units and softmax as the activation function. The grading system employs cross entropy as the loss function. The training uses the AdamW optimiser (Loshchilov & Hutter, 2019) with batch size 4, gradient accumulation step 2, dropout 0.2, and

[1] `huggingface.co/patrickvonplaten/wav2vec2-base`

learning rate 1e-5. The grader is trained for 8 epochs.

TLT-school - holistic score For the TLT-school data, assessment is treated as a regression task. Therefore the output layer has 1 unit and a linear activation function. The loss function is MSE. The grader is trained for 12 epochs using AdamW optimiser with batch size 4, gradient accumulation step 2, dropout 0.2, and learning rate 5e-5.

TLT-school - analytic subscores We trained 6 different graders for each individual aspect of proficiency. The batch size and gradient accumulation steps are the same as the holistic grader, whereas the other hyperparameters are reported in Table 6.3.

	Epochs	**Learning rate**	**Dropout**
Relevance	13	5e-6	0.1
Correctness	19	2e-6	0.1
Lexical	12	4e-6	0.1
Pronunciation	8	1e-5	-
Fluency	6	8e-6	0.1
Communicative	10	1e-5	0.1

Table 6.3: Hyperparameters of the individual wav2vec2-based graders.

As has been mentioned earlier, the first part of wav2vec 2.0 consists of a stack of CNN layers that are employed to obtain acoustically meaningful — but contextually independent — features from the raw speech signal. This part of the model has been sufficiently trained during pre-training and does not need to be fine-tuned. For this reason, we kept the parameters of the feature extractor frozen for all our experiments.

BERT-based graders

The baseline systems used for comparison are BERT-based graders in the version provided by the HuggingFace Transformer Library (Wolf et al., 2020).[2] They are fed with a sequence of token embeddings, i.e., of the responses provided by the learners as inputs. Each token is transformed into a vector representation and then passed to BERT's encoder layer. We use the [CLS] token state and feed it to a classification or regression head, depending on the nature of the task. Similarly to what we did with our wav2vec2-based models, we kept the BERT layer frozen. We tried various hyperparameters and architectures, and we opted for the ones described in the following paragraphs.

[2] `huggingface.co/bert-base-uncased`

ICNALE The classification head consists of a stack of three Dense layers of 768 units, another stack of three Dense layers of 128 units, and an output layer of 5 units with softmax as the activation function. The model is trained for 600 epochs with batch size set to 256 using Adam optimiser (Kingma & Ba, 2015) with learning rate set to 5e-5 and cross entropy as the loss function. The maximum sequence length is set to 256.

TLT-school - holistic score The regression head has the same intermediate layers as the model used on the ICNALE data and an output layer of 1 unit with a linear activation function. The model is trained for 800 epochs on the manual transcriptions and for 150 epochs on the ASR transcriptions. The batch size is set to 256 and the maximum sequence length is 64. The training employs Adam optimiser with learning rate 2e-5 and dropout 0.2.

TLT-school - analytic subscores Similarly to what we did with the wav2vec2-based graders, we trained a BERT-based grader for each specific aspect of proficiency. The architecture and hyperparameters of the individual graders are the same as the holistic grader, with the exception of the dropout rate of the fluency grader, which is set to 0.4.

For evaluating the performance of the graders trained on ICNALE, we use accuracy and weighted F_1 score, whereas both the holistic and analytic graders trained on TLT-school are evaluated using PCC, SRC, and MSE.

6.1.4 Experiments and results

Results on ICNALE

We began our experiments with the classification task on the ICNALE data. First, we trained and evaluated the BERT-based baseline grader. Secondly, we tried the wav2vec2-based grader. Table 6.4 reports the results of the performance of the two grading systems in terms of accuracy and weighted F_1 score on the ICNALE test set. Figure 6.2 reports the confusion matrices of each CEFR proficiency level for the two grading systems. As can be observed, the wav2vec-based grading system significantly outperforms the BERT-based grader across all proficiency levels. In particular, it performs best on B1_2 and on the class of L1 speakers. While the reason for these results on the latter may be attributed to a clear gap between L1 and low/mid levels of L2 English, such as the ones featured in the considered dataset, the performance on the former class is probably due to the greater amount of training data available compared to the other

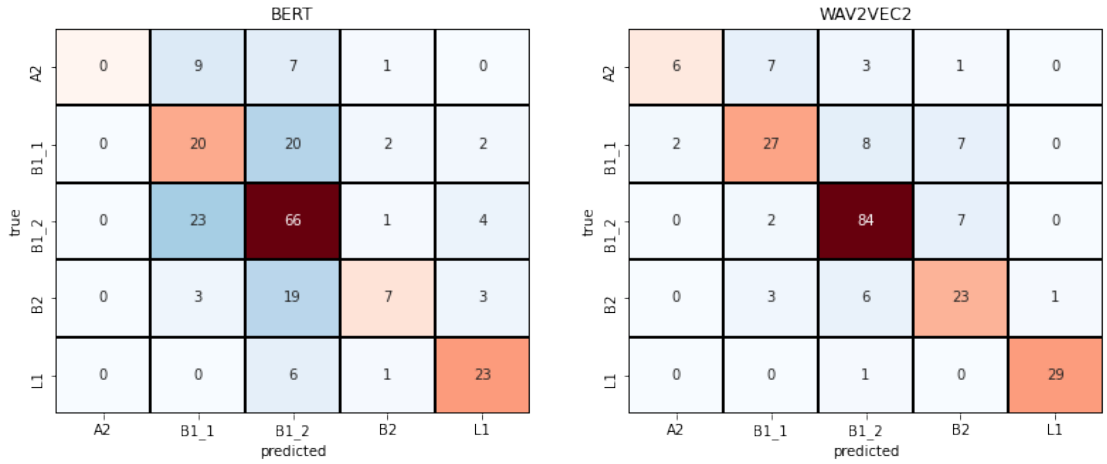


Figure 6.2: Confusion matrices of CEFR proficiency levels for the two grading systems (predicted labels on X-axis, true labels on Y-axis) on the ICNALE test set.

CEFR proficiency levels. Similarly but conversely, this can be inferred from the relatively worse performance on A2, which is the least represented CEFR proficiency level (see Table 6.1).

	Accuracy (%)	**Weighted F_1**
BERT	53.45	0.50
wav2vec2	77.88	0.77

Table 6.4: Results on the ICNALE test set of the BERT-based and wav2vec2-based graders in terms of accuracy and weighted F_1 score.

Results on TLT-school

We continued our experiments on TLT-school, beginning from the holistic graders. In this case, we compared the performance of the two approaches using both manual and ASR transcriptions. Table 6.5 shows the results on the TLT-school test set. It appears that the wav2vec2-based grading system significantly outperforms the BERT baseline both on the manual and ASR transcriptions across all considered metrics.

Subsequently, in order to verify the impact of our two approaches on individual aspects of proficiency, we trained our models on the analytic subscores of each proficiency indicator in the TLT-school test set: relevance, formal correctness, lexical richness and complexity, pronunciation, fluency, and communicative effectiveness. For this part of our experiments, we did not consider the ASR transcriptions, but we only focused on the upper bound provided by the manual

	PCC	**SRC**	**MSE**
BERT-ASR	0.749	0.743	9.877
BERT-manual	0.857	0.863	6.110
wav2vec2	0.927	0.933	2.297

Table 6.5: Results on TLT-school test set (holistic score) of the BERT-based grader (manual and ASR transcriptions) and the wav2vec2-based grader in terms of PCC, SRC and MSE.

transcriptions. The results are reported in Figure 6.3. As can be observed, the wav2vec2-based approach shows significantly better performances than the BERT-based approach at a global level.

In particular, as expected, the more exquisitely speech-related analytic subscores (i.e., pronunciation, fluency, and communicative effectiveness) are best predicted using wav2vec 2.0, despite the presence of proxies of fluency and pronunciation in the manual transcriptions. Furthermore, the wav2vec2-based graders considerably outperform the BERT-based baselines for the task of predicting the formal correctness subscore and the relevance subscore. Since the question prompt is not fed into the grading systems, the fairly good results on the latter might be due to the fact that the test set can essentially be considered a subset of the training set. In this way, the graders recognise whether a response is relevant or irrelevant based on a comparison with other responses labelled as relevant or irrelevant.

The only subscore on which the predictions of both approaches appear fairly aligned is the one related to lexical richness and complexity, despite the BERT baseline performance being lower in terms of PCC. This is quite an expectable result, given that BERT is trained on a large quantity of textual data, and lexical richness and complexity are competences that should be typically constructed and evaluated starting from text, be it written or spoken.

Considering these results, we continued our analysis focusing on this specific indicator. In particular, we wanted to analyse the performance of the wav2vec2-based and BERT-based grading systems across scores to understand whether these two approaches show different performances across proficiency and, therefore, could be complementary to each other. Figure 6.4 shows the MSE variation across scores applying Gaussian kernel smoothing with sigma set to 0.5.

We found that the BERT-based model has a moderately better performance in the central scores, whereas the wav2vec2-based approach shows a significantly lower MSE for the lowest and highest scores. To verify their complementarity, we combined them using: a) a *shallow* combination: we calculated the simple average of the scores predicted by each grader; b) a *deep*

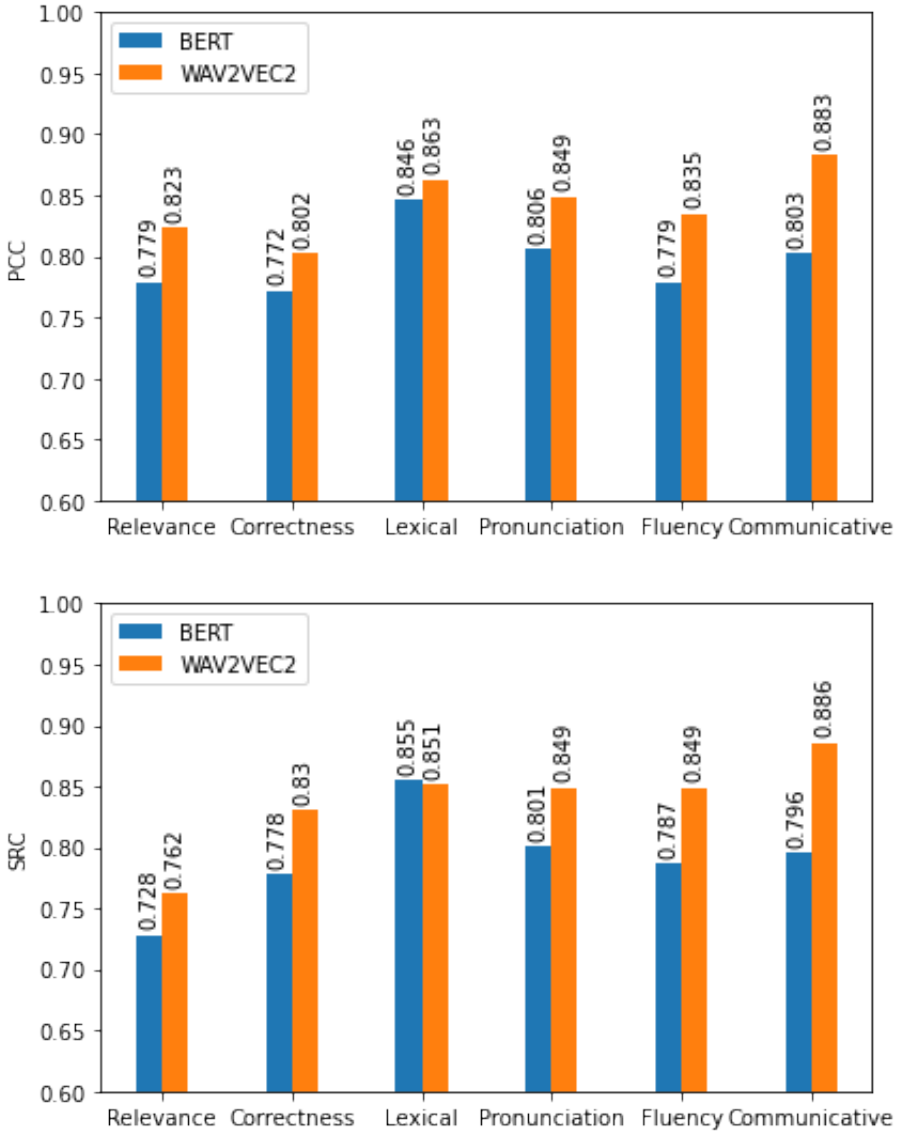


Figure 6.3: Comparison of the BERT-based and wav2vec2-based individual graders in terms of PCC and SRC.

combination: we concatenated the two hidden representations in BERT and wav2vec 2.0, and we fed them through a small network consisting of a Dense layer of 16 units, a Dropout layer with dropout rate set to 0.5, and a final output layer. The resulting network was trained for 3000 epochs with learning rate 5e-5 and batch size 512. In both cases, an interesting improvement across all considered metrics is observed, as shown in Table 6.6.

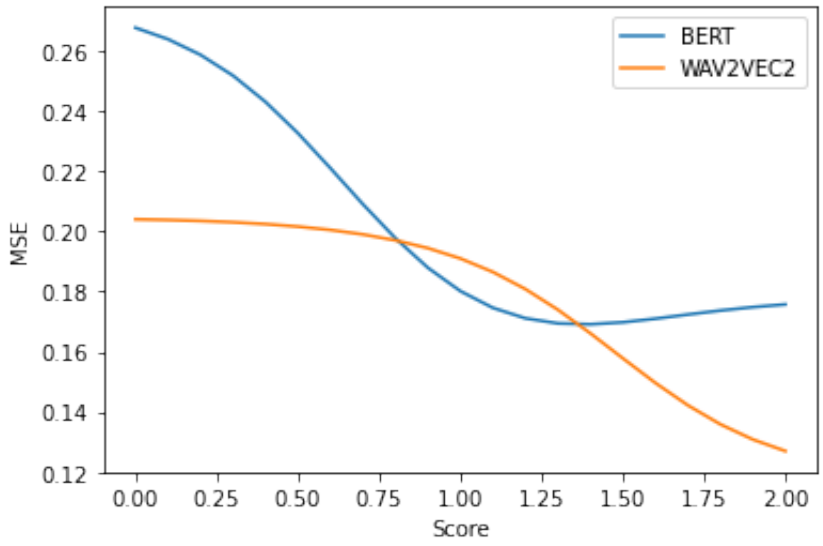


Figure 6.4: MSE variation of the wav2vec2-based and BERT-based (manual transcription) graders across scores for lexical richness and complexity.

	PCC	**SRC**	**MSE**
BERT	0.846	0.855	0.217
wav2vec2	0.863	0.851	0.178
BERT+wav2vec2 (*shallow*)	0.885	0.876	0.164
BERT+wav2vec2 (*deep*)	0.883	0.864	0.166

Table 6.6: Results on TLT-school test set (lexical richness and complexity) of the BERT-based (manual transcriptions), the wav2vec2-based graders, and their combinations in terms of PCC, SRC and MSE.

6.1.5 Conclusions

Transcriptions of L2 speaking tests are not generally easy to obtain, and even when they are made available using ASR systems, they generally contain errors and do not provide information about strictly speech-related aspects of proficiency, such as intonation, rhythm, or prosody. This study

considers whether it is possible to use wav2vec 2.0 representations to assess L2 spoken English proficiency both holistically and analytically, even when a small quantity of data is available. First, we found that this approach significantly outperforms the BERT baseline system trained on manual transcriptions of the ICNALE dataset in the task of CEFR level classification. Secondly, we investigated the use of wav2vec 2.0 for a regression task on the B1 section of TLT-school targeting holistic scores. In this case, we also achieved significant improvements on the BERT baseline trained on ASR and manual transcriptions. Finally, we tested this approach on subscores related to specific facets of proficiency (i.e., relevance, formal correctness, lexical richness and complexity, pronunciation, fluency, and communicative effectiveness) using manual transcriptions only, and we found that the wav2vec2-based grading systems significantly outperform the BERT-based baselines across all proficiency indicators. For lexical richness and complexity, i.e., the only subscore on which the two strategies showed similar results, we found that two types of combination of the two models bring an interesting improvement, thus suggesting a certain degree of complementarity. With respect to analytic proficiency specifically, further work should be undertaken in order to better understand and explain the reasons of the promising performance of the wav2vec2-based grading systems on the prediction of scores related to individual aspects of proficiency, especially to the ones that are not strictly related to speech.

A limitation of this study is that we only compared the wav2vec2-based grader to a BERT-based grader, but we did not consider a grader based on hand-crafted features for further comparison. We will address this issue in Study 5 (see Section 6.2).

With respect to combinations, specific types of combinations should be investigated according to each aspect of proficiency, e.g., a concatenation of the question prompt and learner's answer for the subscore related to relevance along the lines of Qian et al. (2018) and Wang et al. (2021).

Finally, in this work, we have focused on assessment at the response level, but we should explore whether it is possible to extend this approach in order to give feedback about specific parts of an answer, e.g., by analysing the local attention representations.

6.2 Study 5: Speaking assessment using wav2vec 2.0 (Part 2)

6.2.1 Introduction

This study extends the initial analysis conducted in Study 4 (see Section 6.1) on the SSL-based approach to L2 proficiency assessment. Our previous study had two limitations: a) the relatively small amount of data used in the experiments and b) the comparison with a BERT-based baseline only, which, despite being fed with manual transcriptions containing false starts, hesitations, and fragments of words, did not consider purely acoustic features, thus potentially missing strictly speech-related aspects of proficiency. To address these limitations, in this study, we conduct our experiments of proficiency assessment using a large amount of L2 learner data and comparing the performance of a wav2vec2-based grader to two types of grading systems: a BERT-based grader and a standard grader fed with a set of hand-crafted features related to various aspects of proficiency (see Figure 6.5). In addition to this, we test the effectiveness of wav2vec 2.0 on a multi-part examination predicting both the overall grades and the individual grades of each part of the exam. Furthermore, we investigate various combinations between the wav2vec2-based, the BERT-based, and the standard graders.

6.2.2 Data

The data are obtained from candidate responses to the spoken parts of the Linguaskill exams for L2 learners of English, provided by Cambridge English Language Assessment (Ludlow, 2020). Further details can be found in Section 3.2.3, but, for the sake of clarity, the distinctive features of each of the five parts of the exam are briefly illustrated again: in Part 1, the candidates should answer eight personal questions, of which the first two are not graded, and the answers last about 10 or 20 seconds; Part 2 includes a reading aloud activity consisting of eight sentences of 10 seconds each; Part 3 and Part 4 test the candidates' ability to deliver a long turn, speaking for up to one minute; finally, in Part 5, test-takers should give their opinions in form of responses of about 20 seconds to five questions related to a given topic. The reader can refer to Appendix B for examples of questions prompts. Each part of the examination contributes 20% to the speaking exam and is scored on a scale from 1 to 6, in compliance with CEFR standards. Therefore, the overall grade is computed as the average of the grades assigned to the five parts.

The ASR system is very similar to that in Lu, Gales, Knill, Manakul, Wang, & Wang (2019),

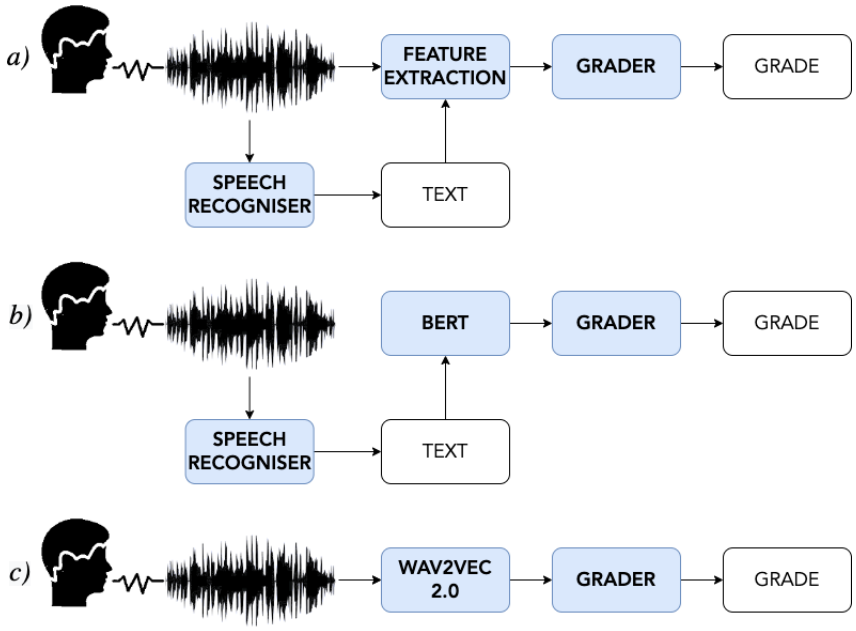


Figure 6.5: The three systems considered in this study: a) standard grader, b) BERT-based grader, and c) wav2vec2-based grader.

i.e., a TDNN-F LFMMI system trained on approximately 500 hours of L2 English data, mostly from Linguaskill Business (BULATS) exams with over 50 L1s included. A su-RNNLM trained on 25.6M in-domain words is used for rescoring. The average WER is ∼20%.

6.2.3 Model architectures

wav2vec2-based graders

For our experiments, we initialised the wav2vec 2.0 model configuration and processor from a version provided by HuggingFace (Wolf et al., 2020).[3] The learners' answers are fed into the model and wav2vec 2.0 provides contextualised representations. To handle representations of various lengths, we employ a mean pooling method to concatenate 3D representations into 2D representations, which are finally projected to a regression head, similarly to what we did in Study 4. Since we trained a grading system for each part of the exam, after exploring various architectures, for Part 1 and Part 5, we used a regression head composed of a layer of 768 units, a Dropout layer, and the output layer, whereas, for Part 2, Part 3 and Part 4, we used a deeper architecture, composed of a stack of three layers of 768 units, a Dropout layer, a layer of 128

[3] `huggingface.co/patrickvonplaten/wav2vec2-base`

units, and finally, the output layer. The graders use MSE as the loss function. We use the AdamW optimiser (Loshchilov & Hutter, 2019), and hyperparameters vary depending on each part. For Part 1, we used batch size 16, gradient accumulation step 2, dropout rate 0.1, and learning rate 5e-5, and we trained the grader for 2 epochs. For Part 2, we used batch size 16, gradient accumulation step 2, dropout rate 0.5, and learning rate 1e-6, and we trained the grader for 3 epochs. For Part 3 and Part 4, we use the same hyperparameters: we set batch size to 8, gradient accumulation step to 4, dropout rate to 0.5, and learning rate to 1e-5, and we trained the grader for 2 epochs. Finally, the grader for Part 5 has batch size set to 8, gradient accumulation step to 2, dropout rate to 0.1, and learning rate to 5e-5, and we trained it for 1 epoch.

As we did in our experiments in Study 4, we froze the parameters of the feature extractor (see Section 6.1.3).

BERT-based graders

For comparison, we use the text grading system presented in Raina et al. (2020) and already employed in Study 3 (see Section 5.1.3). For this grader, BERT was used to extract word embeddings, followed by a multi-head self-attention mechanism. The output of this process is subsequently fed into a feed-forward network. The parameters of the pre-trained BERT model[4] were also fine-tuned.

Standard graders

We also compare our SSL-based approach to a standard grading system, i.e., the one we used in Study 3 (see Section 4.2). It is a DDN trained on a set of hand-crafted features designed to cover all the different aspects of proficiency and is described in Malinin et al. (2017). These features include grade-dependent language model and word level statistics, statistics of phone duration, statistics to capture rhythm, fluency metrics, and fundamental frequency statistics. Further details about the features employed can be found in Wang et al. (2018).

As in Study 3, for all graders, their predictions are the result of ensembles. Further information about the ensemble approach can be found in Wu et al. (2020). Systems are calibrated and in a final set of experiments combined using linear combination:

$$\hat{y}^{(n)} = \beta_0 + \sum_{p \in \mathcal{P}} \beta_p \hat{y}_p^{(n)}$$

[4] `huggingface.co/bert-base-uncased`

where $\mathcal{P}$ is the set of parts to combine, which may come from multiple systems, and β_p are the coefficients associated with the parts. For the baseline submission performance, the values of $\beta_p, p > 0$ are all constrained to be the same, and equal to 0.2, to provide simple averaging consistent with the combination of operational examiner scores. When unequal weighting is used, OLS estimation using the development/calibration set is used to find the values of β_p.

For the evaluation of the grading systems at the per-part level, we use RMSE. Further comparisons also include PCC, SRC, and the percentage of the predicted scores that are equal to or lie within 0.5 (i.e., within half a grade) (% ≤ 0.5), and within 1.0 (i.e., one grade) (% ≤ 1.0) of the actual score.

6.2.4 Experimental results and analysis

Part-Level Performance

We begin our series of experiments from grading each of the five parts of the exam. For this part of our analysis, we only consider LinGen. Table 6.7 reports the results of the three grading systems in terms of RMSE.

Model	**P1**	**P2**	**P3**	**P4**	**P5**
std ($\mathtt{s_d}$)	0.625	0.662	0.671	0.686	0.633
BERT ($\mathtt{b_t}$)	0.628	0.683	0.681	0.694	0.629
wav2vec2 ($\mathtt{w_v}$)	0.601	0.827	0.845	0.845	0.674

Table 6.7: RMSE results on the five parts of the LinGen exam.

It can be observed that the performance of the wav2vec2-based grader varies across the parts of the exam, with close or better RMSE to the other two grading systems for Parts 1 and 5 and lower performance on Part 2, Part 3, and Part 5. This seems to be due to the nature of the responses required for different parts. As mentioned in Section 6.2.2, Parts 1 and 5 include several short spontaneous answers, whilst Parts 3 and 4 also comprise spontaneous speech but learners are required to elaborate a single longer and more complex response in each case. The lower performance of the wav2vec2-based grading system may be due to our use of a mean pooling method, which may be giving too compressed a representation for longer utterances.

Part 2, by contrast, is similar in length to Part 1 but consists of read-aloud responses. This part mainly targets pronunciation and fluency skills at the expense of content-related aspects of proficiency. Therefore, the wav2vec2-based grader might have been expected to do well. Its higher RMSE might be due to the absence of information related to the reference text read

aloud by the test-takers, which is present in the other two grading systems since they leverage transcriptions. It is remarkable that the standard grader, which covers all aspects of a candidate's speech, performs the best, with the BERT-based grader, which cannot measure pronunciation or prosody, slightly behind.

Submission-Level Performance

In the second part of our experiments, we focus on the overall grades of the Linguaskill exam, i.e., the average of the grades assigned to the five parts. Table 6.8 shows the results of the three grading systems both on LinGen and LinBus.

			LinGen		
Model	**PCC**	**SRC**	**RMSE**	**%≤0.5**	**%≤1.0**
$\mathtt{s_d}$	0.932	0.937	0.383	81.5	98.6
$\mathtt{b_t}$	0.929	0.934	0.395	80.3	98.5
$\mathtt{w_v}$	0.908	0.931	0.455	73.3	97.3
			LinBus		
Model	**PCC**	**SRC**	**RMSE**	**%≤0.5**	**%≤1.0**
$\mathtt{s_d}$	0.911	0.918	0.416	76.5	98.3
$\mathtt{b_t}$	0.920	0.925	0.398	80.1	99.2
$\mathtt{w_v}$	0.893	0.911	0.446	72.1	97.9

Table 6.8: Submission-level performance on LinGen and LinBus.

They all achieve interesting results across all metrics, with the standard grader and the BERT-based grader performing moderately better than the wav2vec2-based grading system. As regards the BERT-based grader in particular, we have already observed analogous trends in Raina et al. (2020) and in Study 3 (see Section 5.1), and this fact is quite significant, since it highlights the importance of content-related aspects of speaking proficiency, which is far from being a mere surrogate of elements related to fluency and pronunciation. Furthermore, it appears that the standard grader outperforms the other two grading systems on LinGen, but has a slightly worse performance than the BERT-based grader on LinBus. This might be ascribable to the different language models, since the first test set contains questions on General English, whereas the latter includes questions on Business English, which is typically more specific and complex.

Figure 6.6 shows that the wav2vec2-based graders can differentiate between lower levels of proficiency, but we can clearly see that it is not able to discriminate between the highest levels, as its maximum prediction is 4.6, i.e., between grades B2 and C1 (the plots for LinBus show

a similar trend). Our hypothesis is that higher-level assessment tends to be more dependent on message construction (what is said) rather than message realisation (how it is said), and wav2vec 2.0 should not have actual knowledge of words, although, as mentioned in Section 2.2.2 and Section 6.1.1, it has been demonstrated that SSL speech representations are able to encode a great amount of high-level linguistic features, including aspects of semantics and syntax (Singla et al., 2022).

Combinations

As a preliminary step, we investigated combinations of the grading systems by computing their simple average and using a multiple linear regression model fit with the submission-level predictions, but they did not provide significant gains. Therefore, we investigated the application of a multiple linear regression model using the per-part predictions as predictors for each individual grading system and for four combinations of them. The results on LinGen and LinBus are reported in Table 6.10. The combinations show performances that are aligned to or better than the individual models across all metrics, with the combination including all three grading systems achieving the best results. This combination also overcomes the issue of the wav2vec2-based grader related to scoring higher levels, as can be seen in Figure 6.7. Table 6.9 reports the β coefficients of the individual models described in Table 6.10 and the combination of all three. In the standard grading system, as well as in the BERT-based grading system, Parts 1, 2 and 5 affect the linear model the most, whilst in the wav2vec2-based grading system Part 1 and 5 are the most influential. In their combination, it appears that the highest β coefficients are Parts 1 and 5 of the wav2vec2-based grader and Part 2 of the BERT-based and the standard graders. These values seem to confirm the RMSE results shown in Table 6.7.

Model		**P1**	**P2**	**P3**	**P4**	**P5**	β_0
$\mathtt{s}_\mathtt{d}^{\otimes}$		0.23	0.25	0.14	0.15	0.22	-0.11
$\mathtt{b}_\mathtt{t}^{\otimes}$		0.20	0.26	0.13	0.17	0.23	-0.13
$\mathtt{w}_\mathtt{v}^{\otimes}$		0.29	0.05	0.01	0.01	0.45	0.76
	$\mathtt{s}_\mathtt{d}$	-0.01	0.12	0.06	0.01	-0.04	
$\mathtt{s}_\mathtt{d}\otimes\mathtt{b}_\mathtt{t}\otimes\mathtt{w}_\mathtt{v}$	$\mathtt{b}_\mathtt{t}$	0.06	0.16	0.05	0.09	0.09	0.20
	$\mathtt{w}_\mathtt{v}$	0.20	-0.08	-0.02	0.02	0.20	

Table 6.9: β coefficients of per-part linear regression model for the standard ($\mathtt{s}_\mathtt{d}^{\otimes}$), BERT ($\mathtt{b}_\mathtt{t}^{\otimes}$), wav2vec2 ($\mathtt{w}_\mathtt{v}^{\otimes}$), and combination ($\mathtt{s}_\mathtt{d}\otimes\mathtt{b}_\mathtt{t}\otimes\mathtt{w}_\mathtt{v}$) estimated on the calibration data.

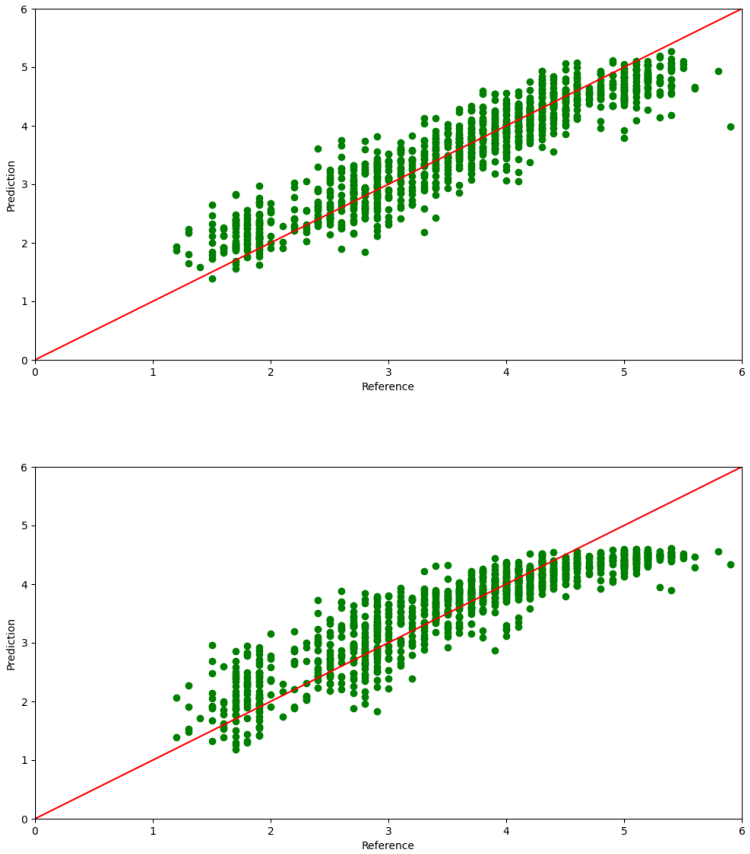


Figure 6.6: Reference vs predicted scores for standard (above) and wav2vec2-based (below) graders on LinGen.

per-part combination $\mathtt{s_d}\otimes\mathtt{b_t}\otimes\mathtt{w_v}$

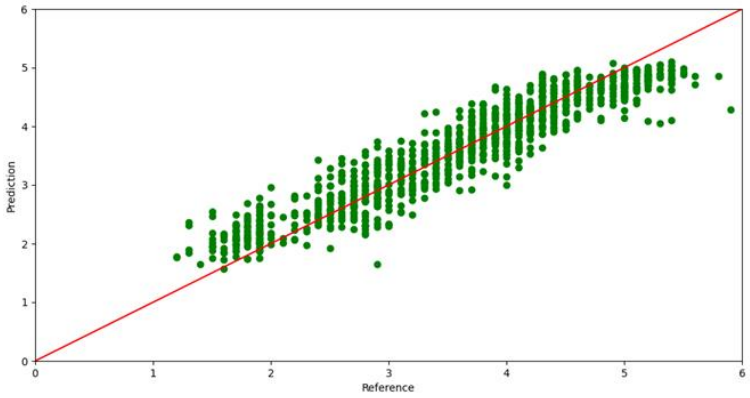


Figure 6.7: Reference vs predicted scores for combined system ($\mathtt{s_d}\otimes\mathtt{b_t}\otimes\mathtt{w_v}$) on LinGen.

LinGen					
Model	**PCC**	**SRC**	**RMSE**	**%≤0.5**	**%≤1.0**
$\mathtt{s}_\mathtt{d}^{\otimes}$	0.932	0.937	0.382	82.3	98.7
$\mathtt{b}_\mathtt{t}^{\otimes}$	0.930	0.935	0.393	80.3	98.6
$\mathtt{w}_\mathtt{v}^{\otimes}$	0.933	0.937	0.393	79.7	99.0
$\mathtt{s}_\mathtt{d}\otimes\mathtt{w}_\mathtt{v}$	0.941	0.945	0.363	84.5	99.3
$\mathtt{s}_\mathtt{d}\otimes\mathtt{b}_\mathtt{t}$	0.936	0.940	0.373	81.9	98.8
$\mathtt{b}_\mathtt{t}\otimes\mathtt{w}_\mathtt{v}$	0.943	0.947	0.359	84.3	99.2
$\mathtt{s}_\mathtt{d}\otimes\mathtt{b}_\mathtt{t}\otimes\mathtt{w}_\mathtt{v}$	0.943	0.947	0.356	85.0	99.1

LinBus					
Model	**PCC**	**SRC**	**RMSE**	**%≤0.5**	**%≤1.0**
$\mathtt{s}_\mathtt{d}^{\otimes}$	0.912	0.920	0.415	77.0	99.0
$\mathtt{b}_\mathtt{t}^{\otimes}$	0.920	0.924	0.400	80.1	99.0
$\mathtt{w}_\mathtt{v}^{\otimes}$	0.916	0.919	0.394	79.1	99.0
$\mathtt{s}_\mathtt{d}\otimes\mathtt{w}_\mathtt{v}$	0.925	0.928	0.378	82.0	99.4
$\mathtt{s}_\mathtt{d}\otimes\mathtt{b}_\mathtt{t}$	0.925	0.929	0.391	80.8	99.4
$\mathtt{b}_\mathtt{t}\otimes\mathtt{w}_\mathtt{v}$	0.930	0.932	0.368	82.7	99.3
$\mathtt{s}_\mathtt{d}\otimes\mathtt{b}_\mathtt{t}\otimes\mathtt{w}_\mathtt{v}$	0.931	0.933	0.366	82.5	99.4

Table 6.10: Results on overall grades on LinGen and LinBus using per-part linear regression estimated on the calibration data.

6.2.5 Conclusions

In this study, we have extended our recent novel approach to proficiency assessment using a wav2vec2-based grading system on a large L2 learner corpus. First, we compared its performance on the five parts of the Linguaskill examination to a BERT-based grader and a high-performing standard grading system fed with hand-crafted features. We found that our proposed approach appears to be sensitive to the nature of the responses and shows a good performance for parts consisting of short spontaneous answers. Secondly, we found that the three grading systems have comparable performances on overall grades, with the wav2vec2-based grader showing some difficulties in assessing higher proficiency levels. Finally, we combined the standard, the BERT-based, and the wav2vec2-based grading system through different linear combinations and we found interesting improvements.

A concern with the wav2vec2-based and BERT-based grading systems is that they are not fully valid individually since neither considers all aspects of the assessment construct and their results are not interpretable to provide informative feedback to learners, teachers, and testers. As well as boosting the assessment performance, combination with the standard grading systems

based on hand-crafted features removes these concerns.

Future works will include further analysis of the behaviour of the wav2vec2-based grading system on different types of assessment task, e.g., including conversational data. We plan to investigate other SSL approaches, i.e., recent models such as HuBERT (Hsu, Bolte, et al., 2021) or WavLM (Chen et al., 2022), as well as other types of combinations, considering both shallow and deep combination methods.

A limitation in our approach seems to be the use of mean pooling, which appears to provide too compressed representations for longer utterances. To address this issue, we plan to replace it with an attention mechanism.

Discussion and conclusions

Discussion

The main contributions of this thesis consist of the approaches to speaking assessment and GEC introduced in Chapter 4, the experiments on view-specific assessment reported in Chapter 5, and the investigation of SSL-based approaches to speaking assessment illustrated in Chapter 6. This chapter summarises and discusses the implications and limitations of the results of the previous chapters, as well as several future avenues of research suggested by the work conducted to date.

First, we consider the experiments on proficiency assessment using grammatical features. In Study 1 (see Section 4.1), we observed that the injection of features related to grammatical accuracy could improve the performance of a BERT-based grader when assessing speaking proficiency. Furthermore, in this study, we explored initial approaches that leverage information derived from written data to assess speech. The suggested explanation for the improvement after injecting grammatical features was that BERT-based models, as they are pre-trained on a large amount of written data, already possess written grammatical knowledge and are sensitive to violations of written grammar to a certain extent. As a result, when evaluating written proficiency, they do not need to be warned with explicit indications concerning errors. On the other hand, features related to grammatical proficiency can be beneficial to understanding and decoding the typical phenomena of oral language and learning spoken and conversational grammar. Two main limitations of this study were: a) that the feature extractor extracts general information about 5 broad categories of grammatical errors, but narrowing down and defining error types more specifically could provide learners, teachers, and testers with more precious details about grammatical proficiency; and b) that we introduced a human bias, i.e., the ranking of error gravity, an issue that could be addressed by investigating other strategies for extracting and weighing errors, e.g., an attention mechanism.

We addressed these two issues in Study 2 (see Section 4.2), in which we illustrated a pipeline which starts from transcriptions of learner speech and then features a module for DD, a module for GEC, and, finally, a module for proficiency assessment. First, we confirmed findings from previous studies, i.e., that disfluency removal is beneficial for GEC. In the final part of the pipeline, GEC edit sequences were fed into a transformer-based grader to predict holistic proficiency scores and formal correctness scores. We compared this grading system to a BERT-based grader and found that the two systems have similar performances when using manual transcriptions. Furthermore, we investigated two types of combinations and found that both bring significant improvements for both the tasks of predicting holistic proficiency scores and formal correctness scores. We noticed that a potential issue with the BERT-based grader is that it might not be fully valid alone since its results are not interpretable to provide feedback to a learner. In addition to boosting the assessment performance, the combinations with the GEC-based grader enhance validity and explainability since this is based on clearly interpretable grammatical features. On the other hand, when employing ASR transcriptions, the BERT-based grader obtains lower results than the GEC-based grader, most likely due to the relatively high WER. For this reason, the GEC-based grader probably leverages certain GEC edit labels, which serve as proxies for pronunciation issues as additional features, although we demonstrated that actual grammatical errors are still extracted from ASR transcriptions and do still play a major role. Further work considering the application of state-of-the-art end-to-end systems is warranted. Furthermore, the integration of features derived from the DD module into a specific grading system should also be explored.

Among other grading systems, a GEC-based grader was also employed in Study 3 (see Section 5.1), whose main focus was on view-specific assessment. In particular, this study investigated whether systems targeting specific aspects of proficiency (i.e., pronunciation, rhythm, text, grammatical accuracy, and grammatical complexity) can be trained when only holistic scores are available. We developed single-view graders whose predictions were seen to be partially different from but complementary to all the others for the task of predicting holistic grades. We observed that the combination of the 5 considered grading systems improves on the performance of each single-view grader. In short, single-view grading should help deconstruct holistic proficiency and make the holistic score significantly more interpretable, enabling useful feedback to learners with specific indications about their forte and their weakness. With respect to the use of grammatical features, we noted that current systems are generally designed for written texts and are not ideal for spoken data. Therefore, further work should be undertaken in this sense. Furthermore, in

this study, we only investigated one type of combination, so other types should be explored, considering both shallow and deep combination methods. As regards combinations, we also stressed the importance of including other views of proficiency in order to provide a richer representation of communicative competence. In particular, further work should investigate the assessment of the sociolinguistic and pragmatic competences.

Our last two studies investigated SSL-based approaches to proficiency assessment exploiting powerful speech representations to simultaneously encode acoustic and linguistic features without a specific ASR module. In particular, Study 4 (see Section 6.1) considered whether it is possible to use wav2vec 2.0 representations to assess L2 spoken English proficiency both holistically and analytically. We compared wav2vec2-based grading systems to BERT-based baselines trained on ASR and manual transcriptions and found that the former outperformed the latter on both the tasks of holistic and analytic proficiency assessment. For the score related to lexical richness and complexity, which was the only subscore on which the two strategies had a similar performance, we found that two types of combination of the two models showed an interesting improvement, thus suggesting a certain degree of complementarity. As concerns the relevance score, we observed an interesting — albeit not remarkable — performance, probably due to the fact that the test set essentially consists of a subset of the training set, but the graders did not leverage any information derived from the question prompts. Therefore, we plan to explore strategies involving the concatenation of the question prompt and the respective responses for the prediction of this specific score. A limitation of this study is that we only compared wav2vec2-based graders to BERT-based graders, which do not consider strictly speech-related features (apart from proxies of fluency and pronunciation), and we did not consider a grader based on hand-crafted features for further comparison. Another issue was the use of a relatively small quantity of training and testing data.

We addressed both these issues in Study 5 (see Section 6.2), in which we explored the use of wav2vec2-based graders on a large L2 learner corpus derived from the spoken parts of the Linguaskill multi-part exams for L2 learners of English, and we compared their performance to BERT-based graders and high-performance standard graders based on hand-crafted features. First, we found that our proposed approach appeared to be sensitive to the nature of the responses and showed a good performance for parts consisting of short spontaneous answers, probably due to the use of mean pooling, which provides too compressed representation for longer responses. To address this issue, we plan to replace it with an attention mechanism. Secondly, we found that the three grading systems have comparable performances on overall grades, with the wav2vec2-

based grading system showing some difficulties in assessing higher proficiency levels. Finally, we combined the standard, the BERT-based, and the wav2vec2-based graders by means of different linear combinations, and we found promising improvements. A potential issue with the wav2vec2-based and BERT-based graders is that they are not fully valid alone since neither considers all aspects of the assessment construct. Specifically, the former approach does not completely take into account message construction (what is said), whereas the latter is limited in measuring message realisation (how it is said), particularly with regard to pronunciation and prosody. Moreover, their results are not interpretable to provide informative feedback to learners. As well as boosting the assessment performance, combinations with the standard graders based on hand-crafted features help contain this issue. We plan to conduct further analysis of the behaviour of the wav2vec2-based grader on different types of grading tasks, e.g., including conversational data. We plan to investigate other SSL approaches, as well as other types of combinations, considering both shallow and deep methods.

Conclusions

This thesis investigated various automatic approaches to assessment and feedback of L2 English oral proficiency. The work was motivated by the increasing demand for automated spoken language assessment and feedback systems for applications in CALL, and part of it was conducted in the framework of the ALTA project.[5]

The theoretical framework and historical background of L2 speaking assessment were illustrated in Chapter 1. Chapter 2 provided a review of automatic assessment techniques and approaches from the early days to the present. In Chapter 3, we described the data used in our studies, as well as other spoken corpora. Chapter 4 included two studies which investigate GED, GEC, and the role of grammatical features in proficiency assessment. In Chapter 5, we explored view-specific assessment. In Chapter 6, we illustrated our experiments with SSL-based approaches to holistic and analytic proficiency assessment. Finally, the findings, implications, and limitations of the experimental results and future perspectives were discussed in the previous section. This section summarises the conclusions of this thesis with respect to the five research questions posed in the Introduction:

1. *How can we increase the performance of automatic speaking assessment systems based on objective elements present in the data?*

[5] `mi.eng.cam.ac.uk/~mjfg/ALTA/index.html`

In Chapter 4, we described two studies which used grammatical features in different ways to assess speaking proficiency. In line with previous studies, we showed that grammatical accuracy and complexity could be considered consistent criterial elements for assessing speaking proficiency. In particular, in Study 1, we introduced a ranking of error gravity drawn from the frequency of errors in an L2 learner corpus, whereas, in Study 2, we used a cascaded system including modules for disfluency detection, grammatical error correction, and proficiency assessment. The latter was a transformer-based grading system, whose attention mechanism should focus on the grammatical errors that have the most salience to the output. In both studies, we demonstrated that the use of grammatical features could improve the performance of automatic speaking assessment systems.

2. *How can we increase the validity and interpretability of results and provide informative feedback to learners, teachers, and testers?*

   We showed that self-supervised learning representations are extremely powerful and yield remarkable results, but we discussed the lack of explainability of their performances several times in this thesis. In this sense, not only did the use of grammatical features in Study 1 and Study 2 boost assessment performance, but it should increase the validity and interpretability of results by providing informative feedback about grammatical proficiency. In particular, the cascaded approach illustrated in Study 2 can be used to provide grammatical error corrections and granular information about the grammatical accuracy of learners' utterances.

   Furthermore, in Study 3, we showed a viable strategy to extract information about individual facets of proficiency when only holistic proficiency scores are available. In addition to assessing analytic aspects of learners' proficiency, it was shown that also the validity and performance of holistic grading are improved.

   Finally, in the final part of Study 5, we observed that the combination of grading systems based on self-supervised representations with standard graders based on hand-crafted features boosted the assessment performance and, at the same time, ensured the explainability of results.

3. *How can we assess communicative competence in speaking automatically?*

   In the first studies illustrated in this thesis, we mainly focused on aspects related to grammatical accuracy and complexity, and, although not with specific regard to proficiency

assessment, we investigated fluency aspects in Study 2, in which we described disfluency detection and removal.

Segmental and suprasegmental pronunciation, grammatical accuracy and complexity, and — to a certain extent — vocabulary were investigated in Study 3 on view-specific assessment, but we have already expressed the need to extend this study in order to include other underlying aspects of communicative competence — especially sociolinguistic and pragmatic elements — in a similar experimental framework.

In Study 4, among other experiments, we tried a wav2vec2-based grader and a BERT-based grader for the task of predicting analytic scores related to specific aspects of proficiency, namely pronunciation, fluency, formal correctness, relevance, lexical richness and complexity, and communicative effectiveness. In this case, sociolinguistic competence was not specifically featured in the human-assigned scores either.

Considering our studies and the existing literature on automatic speaking proficiency assessment, we can conclude that many aspects of communicative competence have been exhaustively analysed, but especially sociolinguistic competence should deserve more attention. This disregard has been due to construct-related issues, in the first place: in Chapter 1 and Chapter 2, we have highlighted the overlaps between linguistic competence — with respect to vocabulary — and sociolinguistic competence, and we have also seen that sociolinguistic and pragmatic competences suffer from crucial underspecifications, ambiguities, and inaccuracies. Not only are these aspects challenging for automatic systems, but they may even be misleading and problematic for human evaluators. Furthermore, the reader may have noticed that some aspects of pragmatic competence, as described in the CEFR, were not treated in detail or were incorporated into other elements, i.e., propositional precision, thematic development, flexibility, and turn-taking. It remains difficult to clearly distinguish these aspects from the two main branches of fluency and coherence and cohesion.

Another fundamental problem concerns the two principles of pronunciation assessment since an L1-like pronunciation is still the goal of various CAPT resources and applications, despite the supposed primacy of intelligibility. A promising approach to the automatic implementation of the "intelligibility principle" seems to be the use of attention-based models identifying and weighting each phone or each consonant and inter-consonant interval in the learner's utterance in relation to each of the others, which we used in Study 3,

provided that the scores used as references are assigned in compliance with this principle. In this regard, Lado's question as to *who* should judge intelligibility remains open (see note 10 in Chapter 1).

4. *How can we use written data in order to assess speaking proficiency?*

   In Chapter 2, we reported that various studies have used BERT embeddings, which are pre-trained on massive written corpora, to assess speaking proficiency, and we also investigated this approach in our studies. Specifically, in Study 1, we acknowledged that a crucial issue in the field of automatic assessment of spoken language proficiency is the lack of data specifically designed and annotated for this purpose, whereas there are numerous and large publicly available written corpora. Therefore, we demonstrated that it is feasible to leverage the information contained in written data to score speaking proficiency. Additionally, we used features related to errors extracted from a large L2 learner corpus of writings to increase assessment performance.

   In Study 2, we demonstrated that it is possible to train grammatical error correction models on publicly available written data and achieve state-of-the-art results. The information extracted from grammatical error correction was subsequently used for assessing both holistic proficiency and formal correctness scores.

5. *How can we assess speaking proficiency automatically, avoiding transcriptions?*

   We have seen that automatic transcriptions of L2 learner speech might often be inaccurate, thus producing noise and propagating it to the rest of a typical automatic speaking assessment (and feedback) pipeline. To avoid this problem, in Study 4 and Study 5, we proposed the use of self-supervised learning speech representations to assess both analytic and holistic proficiency. In particular, in Study 4, not only did we obtain promising results for the prediction of scores related to strictly speech-related aspects of proficiency, such as fluency, pronunciation, and communicative effectiveness, but we also achieved remarkable performances for the task of predicting scores related to content-related features, such as lexical richness and complexity, formal correctness, and — to a lesser extent — relevance.

This PhD project started in November 2019. In the last three and a half years, we have witnessed groundbreaking approaches to language assessment, which are closely connected to advancements and innovations in NLP and speech processing. For instance, BERT

was released in 2018 and has become a standard architecture that has taken the world of NLP by storm in recent years. Its use has been widely investigated in written and spoken assessment, and we have employed it in all the studies presented in this thesis. Furthermore, the first version of wav2vec was released in late 2019 and represented another milestone, quickly becoming the baseline for SSL speech representations. In the field of CALL, it has been mainly explored for mispronunciation detection and diagnosis, but we have shown its promising results for holistic and analytic proficiency assessment. Apparently, the next significant paradigm shift will be brought by large language models, such as GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023), which for the first time, have also attracted laypeople's attention. In her keynote at ISCSLP 2022, cited at the beginning of Chapter 5, Dr. Kate Knill also mentioned using ChatGPT[6] for GEC obtaining interesting results, although errors introduced by ASR still seemed to be problematic. A potentially interesting use of ChatGPT could also be a 'spoken version' employed as a language teacher and tester, which might engage learners in actual conversations in addition to providing assessment and feedback on their language proficiency.

[6] `openai.com/blog/chatgpt/`

References

Abe, M. (2007). Grammatical errors across proficiency levels in L2 spoken and written English. *The Economic Journal of Takasaki City University of Economics*, *49*(3), 4. Retrieved from `http://www1.tcue.ac.jp/home1/k-gakkai/ronsyuu/ronsyuukeisai/49_3.4/abema.pdf`

Adams, M. L. (1980). Five cooccurring factors in speaking proficiency. In J. R. Frith (Ed.), *Measuring spoken language proficiency* (p. 1-6). Washington, DC: Georgetown University Press.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *4*(91), 659-663. doi: 10.1111/j.1540-4781.2007.00627_4.x

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied linguistics*, *14*(2). doi: 10.1093/applin/14.2.115

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016* (pp. 715–725). doi: 10.18653/v1/P16-1068

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. P. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). Newark: International Reading Association.

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Hutson (Ed.), *Advances in reading / language research* (Vol. 2). Greenwich: JAI Press.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentais, prosody, and syllable structure. *Language learning*, *42*(4), 529–555. doi: 10.1111/j.1467-1770.1992.tb01043.x

Anderson-Hsieh, J., & Venkatagiri, H. (1994). Syllable duration and pausing in the speech of chinese esl speakers. *TESOL quarterly*, *28*(4), 807–812. doi: 10.2307/3587566

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., ... Collins, M. (2016). Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2442–2452). doi: 10.18653/v1/P16-1231

Andre-Obrecht, R. (1988). A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *36*(1), 29–40. doi: 10.1109/29.1486

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, *66*(1-2), 46–63. doi: 10.1159/000208930

Asakawa, S., Minematsu, N., Isei-Jaakkola, T., & Hirose, K. (2005). Structural representation of the non-native pronunciations. In *Proceedings of Interspeech 2005.* Retrieved from `https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2005/i05_0165.pdf`

Attali, Y. (2007). *Construct validity of e-rater® in scoring TOEFL® essays* (Tech. Rep. No. 1). ETS Research Report Series. Retrieved from `https://files.eric.ed.gov/fulltext/EJ1111570.pdf`

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, *4*(3). Retrieved from `https://ejournals.bc.edu/index.php/jtla/article/view/1650`

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Baese-Berk, M. M., Drake, S., Foster, K., Lee, D.-y., Staggs, C., & Wright, J. M. (2021). Lexical diversity, lexical sophistication, and predictability for speech in multiple listening conditions. *Frontiers in psychology*, *12*, 661415. doi: 10.3389/fpsyg.2021.661415

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceed-*

ings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020) (pp. 1–12). Retrieved from https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

Ballier, N., Gaillat, T., Simpkin, A., Stearns, B., Bouyé, M., & Zarrouk, M. (2019). A supervised learning model for the automatic assessment of language levels based on learner errors. In *EC-TEL 2019 14th European Conference on Technology Enhanced Learning* (p. 1-13). doi: 10.1007/978-3-030-29736-7_23

Bannò, S., Balusu, B., Gales, M. J. F., Knill, K. M., & Kyriakopoulos, K. (2022). View-specific assessment of L2 spoken English. In *Proceedings of Interspeech 2022* (pp. 4471–4475). doi: 10.21437/Interspeech.2022-10691

Bannò, S., Knill, K. M., Matassoni, M., Raina, V., & Gales, M. J. F. (2022). *L2 proficiency assessment using self-supervised speech representations.* ArXiv. doi: 10.48550/arXiv.2211.08849

Bannò, S., & Matassoni, M. (2022). Cross-corpora experiments of automatic proficiency assessment and error detection for spoken English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (pp. 82–91). doi: 10.18653/v1/2022.bea-1.12

Bannò, S., & Matassoni, M. (2023). Proficiency assessment of L2 spoken English using wav2vec 2.0. In *2022 IEEE Spoken Language Technology Workshop (SLT)* (p. 1088-1095). doi: 10.1109/SLT54892.2023.10023019

Bannò, S., Matassoni, M., & Simakova, S. (2021). Towards error-based strategies for automatically assessing ESL learners' proficiency. In *Collated Papers for the ALTE 7th International Conference* (p. 148-153). Retrieved from https://www.alte.org/resources/Documents/ALTE\%207th\%20International\%20Conference\%20Madrid\%20June\%202021.pdf#page=155

Barker, F. (2006). Corpora and language assessment: trends and prospects. *Research Notes*, *26*, 2-4. Retrieved from https://www.cambridgeenglish.org/Images/23145-research-notes-26.pdf

Baroni, M., Lenci, A., & Onnis, L. (2007). ISA meets Lara: An incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the Workshop*

on Cognitive Aspects of Computational Language Acquisition (pp. 49–56). Retrieved from https://aclanthology.org/W07-0607

Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, *34*(1), 1–34. Retrieved from https://www.cl.cam.ac.uk/teaching/1617/R216/entitycoherence.pdf

Baur, C., Caines, A., Chua, C., Gerlach, J., Qian, M., & Rayner, M. (2019). Overview of the 2019 Spoken CALL Shared Task. In *Proceedings of the 8th Workshop on Speech and Language Technology for Education (SLaTE 2019)* (pp. 1–5). doi: 10.21437/SLaTE.2019-1

Baur, C., Caines, A., Chua, C., Gerlach, J., Qian, M., Rayner, M., ... Wei, X. (2018). Overview of the 2018 spoken CALL shared task. In *Proceedings of Interspeech 2018* (pp. 2354–2358). doi: 10.21437/Interspeech.2018-97

Baur, C., Chua, C., Gerlach, J., Rayner, E., Russel, M., Strik, H., & Wei, X. (2017). Overview of the 2017 spoken CALL shared task. In *Proceedings of the 7th Workshop on Speech and Language Technology in Education (SLaTE 2017).* Retrieved from https://archive-ouverte.unige.ch/unige:97428/ATTACHMENT01

Becker, K., & Edalatishams, I. (2019). ELSA Speak - Accent Reduction [Review]. In *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference* (p. 434-438). Retrieved from https://www.iastatedigitalpress.com/psllt/article/id/15397/download/pdf/

Bell, R. T. (1993). *Translation and translating: Theory and practice.* London: Longman.

Bernstein, J., & Cheng, J. (2007). Logic, operation and validation of a spoken English test. In V. M. Holland & F. P. Fisher (Eds.), *Speech technologies for language learning* (pp. 174–194). New York: Routledge.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., & Weintraub, M. (1990). Automatic evaluation and training in English pronunciation. In *First International Conference on Spoken Language Processing.* Retrieved from https://www.isca-speech.org/archive_v0/archive_papers/icslp_1990/i90_1185.pdf

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, *69*, 65–78. doi: 10.1016/j.system.2017.08.004

Bhat, S., Xue, H., & Yoon, S.-Y. (2014). Shallow analysis based assessment of syntactic complexity for automated speech scoring. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1305–1315). doi: 10.3115/v1/P14-1123

Bhat, S., & Yoon, S. (2015). Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication*, *67*, 42–57. doi: 10.1016/j.specom.2014.09.005

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*(5), 639–668. doi: 10.1093/applin/amu059

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022. Retrieved from `https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf`

Bloomfield, L. (1933). *Language*. London: George Allen and Unwin.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). *Language teaching research*, *10*(3), 245–261. doi: 10.1191/1362168806lr195oa

Bøhn, H., & Hansen, T. (2017). Assessing pronunciation in an EFL context: Teachers' orientations towards nativeness and intelligibility. *Language Assessment Quarterly*, *14*(1), 54–68. doi: 10.1080/15434303.2016.1256407

Bosker, H. R., Quené, H., Sanders, T., & De Jong, N. H. (2014). The perception of fluency in native and nonnative speech. *Language Learning*, *64*(3), 579–614. doi: 10.1111/lang.12067

Broniś, O. (2016). Italian vowel paragoge in loanword adaptation. phonological analysis of the Roman variety of Standard Italian. *Italian Journal of Linguistics*, *28*(2), 25–68. Retrieved from `https://www.italian-journal-linguistics.com/app/uploads/2021/05/2_Bronis.pdf`

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL quarterly*, *22*(4), 593–606. doi: 10.2307/3587258

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on english-for-academic-purposes speaking tasks. *ETS Research Report Series*(1). doi: 10.1002/j.2333-8504.2005.tb01982.x

Brown, J. D. (2001). Six types of pragmatics tests in two different contexts. In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 301–325). Cambridge: Cambridge University Press.

Brown, J. D., & Ahn, R. C. (2011). Variables that affect the dependability of l2 pragmatics tests. *Journal of Pragmatics*, *43*(1), 198–217. doi: 10.1016/j.pragma.2010.07.026

Brown, P., & Levinson, S. (1987). *Politeness: some universals in language usage*. Cambridge: Cambridge University Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Bryant, C., Felice, M., Andersen, Ø. E., & Briscoe, T. (2019). The BEA-2019 Shared Task on Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2019)* (pp. 52–75). doi: 10.18653/v1/W19-4406

Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 793–805). doi: 10.18653/v1/P17-1074

Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., & Briscoe, T. (2022). *Grammatical error correction: A survey of the state of the art*. ArXiv. doi: 10.48550/arXiv.2211.05166

Burstein, J. (2002). The e-rater scoring engine: automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: a cross-disciplinary perspective* (pp. 113–122). New York: Routledge.

Burstein, J., Tetreault, J., & Andreyev, S. (2010). Using entity-based features to model coherence in student essays. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 681–684).

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper* (Tech. Rep.). Princeton: Educational Testing Service. Retrieved from `https://www.ets.org/Media/Research/pdf/RM-00-06.pdf`

Caines, A., Bentz, C., Graham, C., Polzehl, T., & Buttery, P. (2016). Crowdsourcing a multi-lingual speech corpus: Recording, transcription and annotation of the CrowdIS corpora. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 2145–2152). Retrieved from `https://aclanthology.org/L16-1340`

Caines, A., Bentz, C., Knill, K., Rei, M., & Buttery, P. (2020). Grammatical error detection in transcriptions of spoken English. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 2144–2162). doi: 10.18653/v1/2020.coling-main.195

Caines, A., & Buttery, P. (2020). REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (pp. 5614–5623). Retrieved from `https://aclanthology.org/2020.lrec-1.689`

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 33–42). New York: Newbury House Publishers.

Canale, M., & Swain, M. (1980). Theoretical bases for communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47. doi: 10.1093/applin/I.1.1

Capel, A. (2015). The English Vocabulary Profile. In J. Harrison & F. Barker (Eds.), *English profile in practice.* Cambridge: Cambridge University Press.

Carroll, J. B. (1961). Fundamental considerations in testing for English proficiency of foreign students. In *Testing the English proficiency of foreign students* (p. 30-40). Washington: Center for Applied Linguistics.

Carroll, J. B. (1964). *Language and thought.* Englewood Cliffs: Prentice-Hall International.

Carroll, J. B. (1968). The psychology of language testing. In A. Davies (Ed.), *Language testing symposium: A psycholinguistic approach* (p. 46-69). London: Oxford University Press.

Carroll, J. B. (1978). *An English language testing service: specifications.* London: British Council.

Celce-Murcia, M., Brinton, D. M., Goodwin, J. M., & Griner, B. (2010). *Teaching pronunciation: A course book and reference guide (2nd edition).* Cambridge: Cambridge University Press.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., ... Kurzweil, R. (2018). Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 169–174). doi: 10 .18653/v1/D18-2029

Chambers, F. (1997). What do we mean by fluency? *System*, *25*(4), 535–544. doi: 10.1016/ S0346-251X(97)00046-8

Charteris-Black, J. (2002). Second language figurative proficiency: A comparative study of malay and english. *Applied linguistics*, *23*(1), 104–133. doi: 10.1093/applin/23.1.104

Chen, A., Gussenhoven, C., & Rietveld, T. (2004). Language-specificity in the perception of paralinguistic intonational meaning. *Language and Speech*, *47*(4), 311–349. doi: 10.1177/ 00238309040470040101

Chen, L., Evanini, K., & Sun, X. (2010). Assessment of non-native speech using vowel space characteristics. In *Proceedings of the 2010 IEEE Spoken Language Technology Workshop (SLT 2010)* (p. 139-144). doi: 10.1109/SLT.2010.5700836

Chen, L., Tao, J., Ghaffarzadegan, S., & Qian, Y. (2018). End-to-end neural network based automated speech scoring. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)* (pp. 6234–6238). doi: 10.1109/ICASSP.2018.8462562

Chen, L., Tetreault, J., & Xi, X. (2010). Towards using structural events to assess non-native speech. In *Proceedings of the NAACL HLT 2010 5th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 74–79). Retrieved from `https://aclanthology.org/W10-1010`

Chen, L., & Yoon, S.-Y. (2011). Detecting structural events for assessing non-native speech. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2011)* (pp. 38–45). Retrieved from `https://aclanthology.org/W11-1405`

Chen, L., & Yoon, S.-Y. (2012). Application of structural events detected on ASR outputs for automated speaking assessment. In *13th Annual Conference of the International Speech Communication Association* (pp. 767–770). Retrieved from `https://www.isca-speech.org/archive/pdfs/interspeech_2012/chen12c_interspeech.pdf`

Chen, M., & Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th annual*

meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 722–731). Retrieved from https://aclanthology.org/P11-1073

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., ... others (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*. doi: 10.1109/JSTSP.2022.3188113

Cheng, S., Liu, Z., Li, L., Tang, Z., Wang, D., & Zheng, T. F. (2020). ASR-Free pronunciation assessment. In *Proceedings of Interspeech 2020* (pp. 3047–3051). doi: 10.21437/Interspeech .2020-2623

Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics in Language Teaching*, *V*(1), 161-170. Retrieved from https://edisciplinas.usp.br/pluginfile.php/5732715/mod_resource/content/1/Corder\%201968\%20\%281\%29\%20errors.pdf

Corder, S. P. (1981). *Error analysis and interlanguage*. Oxford: Oxford University Press.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press. Retrieved from https://rm.coe.int/1680459f97

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment - Companion volume*. Strasbourg: Council of Europe. Retrieved from https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Coutinho, E., Hönig, F., Zhang, Y., Hantke, S., Batliner, A., Nöth, E., & Schuller, B. (2016). Assessing the prosody of non-native speakers of English: Measures and feature sets. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Retrieved from https://aclanthology.org/L16-1211

Craighead, H., Caines, A., Buttery, P., & Yannakoudakis, H. (2020). Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2258–2269). doi: 10.18653/v1/2020.acl-main.206

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, *51*(1), 14–27. doi: 10.3758/s13428-018-1142-4

Crossley, S. A., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*(2), 171–192. doi: 10125/44329

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). What is lexical proficiency? some answers from computational models of speech data. *Tesol Quarterly*, *45*(1), 182–193. doi: 10.5054/tq.2010.244019

Cruz-Ferreira, M. (1987). Non-native interpretive strategies for intonational meaning: An experimental study. In A. James & J. Leather (Eds.), *Sound patterns in second language acquisition* (pp. 103–120). New York: Foris. doi: 10.1515/9783110878486

Crystal, D. (2011). *A dictionary of linguistics and phonetics (Sixth Edition)*. Oxford: John Wiley & Sons.

Cucchiarini, C., Hubers, F., & Strik, H. (2022). Learning L2 idioms in a CALL environment: the role of practice intensity, modality, and idiom properties. *Computer Assisted Language Learning*, *35*(4), 863–891. doi: 10.1080/09588221.2020.1752734

Cucchiarini, C., Neri, A., Wet, F., & Strik, H. (2007). ASR-based pronunciation training: Scoring accuracy and pedagogical effectiveness of a system for Dutch L2 learners. *Proceedings of Interspeech 2007*, 2181–2184. Retrieved from `https://www.isca-speech.org/archive/pdfs/interspeech_2007/cucchiarini07_interspeech.pdf`

Cucchiarini, C., Strik, H., & Boves, L. (1997). Automatic evaluation of Dutch pronunciation by using speech recognition technology. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings* (pp. 622–629). doi: 10.1109/ASRU.1997.659144

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, *107*(2), 989–999. doi: 10.1121/1.428279

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, *111*(6), 2862–2873. doi: 10.1121/1.1471894

Cummins, R., & Rei, M. (2018). *Neural multi-task learning in automated assessment.* ArXiv. doi: 10.48550/arXiv.1801.06830

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, *26*(2), 163-174. doi: 10.1016/S0346-251X(98)00001-3

Dahlmeier, D., & Ng, H. T. (2012). Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 568–572). Retrieved from `https://aclanthology.org/N12-1067`

Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 22–31). Retrieved from `https://aclanthology.org/W13-1703`

Dale, R., & Kilgarriff, A. (2011). Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation* (pp. 242–249). Retrieved from `https://aclanthology.org/W11-2838`

Davies, A. (2003). *The native speaker: myth and reality.* Clevedon: Multilingual Matters.

Davis, L., & Papageorgiou, S. (2021). Complementary strengths? evaluation of a hybrid human-machine scoring approach for a test of oral academic English. *Assessment in Education: Principles, Policy & Practice*, *28*(4), 437–455. doi: 10.1080/0969594X.2021.1979466

de Beaugrande, R., & Dressler, W. (1981). *Introduction to text linguistics.* London: Longman Publishing.

De Jong, N. H. (2017). Fluency in second language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 204–218). Boston: DeGruyter Mouton. doi: 10.1515/9781614513827

De Jong, N. H., Steinel, M. P., Florijn, A. F., R., S., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*(34), 5-34. doi: 10.1017/S0272263111000489

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in second language acquisition*, *19*(1), 1–16. doi: 10.1017/S0272263197001010

Derwing, T. M., & Munro, M. J. (2009). Comprehensibility as a factor in listener interaction preferences: implications for the workplace. *Canadian Modern Language Review*, *2*(66), 181-202. doi: 10.3138/cmlr.66.2.181

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*(4), 533–557. doi: 10.1017/S0272263109990015

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language learning*, *54*(4), 655–679. doi: 10.1111/j.1467-9922.2004.00282.x

de Saussure, F. (2011). *Course in general linguistics: Translated by W. Baskin. Edited by P. Meisel and H. Saussy.* New York: Columbia University Press.

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)* (pp. 4171–4186). doi: 10.18653/v1/N19-1423

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the versant for english test: A response. *Language Assessment Quarterly*, *5*(2), 160–167. doi: 10.1080/15434300801934744

Educational Testing Service. (2009). *The official guide to the TOEFL test (3rd edition).* New York: McGraw-Hill.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL quarterly*, *42*(3), 375–396. doi: 10.1002/j.1545-7249.2008.tb00137.x

Ellis, R. (1994). *The study of second language acquisition.* Oxford: Oxford University Press.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: validity, technical adequacy, and implementation* (pp. 261–287). Mahwah: Lawrence Erlbaum.

Enright, M., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, *3*(27), 317-334. doi: 10.1177/0265532210363144

Evanini, K., Mulholland, M., Ubale, R., Qian, Y., Pugh, R. A., Ramanarayanan, V., & Cahill, A. (2018). Improvements to an automated content scoring system for Spoken CALL responses: the ETS submission to the Second Spoken CALL Shared Task. In *Proc. Interspeech 2018* (pp. 2379–2383). doi: 10.21437/Interspeech.2018-2362

Fathullah, Y., Gales, M., & Malinin, A. (2021). Ensemble distillation approaches for grammatical error correction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (p. 2745-2749). doi: 10.1109/ICASSP39728.2021.9413385

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/7287.001.0001

Fillmore, C. (1979). On fluency. In C. Fillmore & W. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85–101). San Diego: Academic Press. doi: 10.1016/B978-0-12-255950-1.50012-3

Flege, J. E. (1984). The detection of French accent by American listeners. *The Journal of the Acoustical Society of America*, *76*(3), 692–707. doi: 10.1121/1.391256

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, *25*(2-3), 285–307. doi: 10.1080/01638539809545029

Fontan, L., Le Coz, M., & Detey, S. (2018). Automatically measuring L2 speech fluency without the need of ASR: A proof-of-concept study with Japanese learners of French. In *Proceedings of Interspeech 2018* (pp. 2544–2548). doi: 10.21437/Interspeech.2018-1336

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics*, *21*(3), 354-375. doi: 10.1093/applin/21.3.354

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., ... Cesari, F. (2000). The SRI EduSpeak™ system: Recognition and pronunciation scoring for language learning. In *Proceedings of InSTILL* (pp. 123–128). Retrieved from `https://www.sri.com/wp-content/uploads/2021/12/the_sri_eduspeak.pdf`

Franco, H., Neumeyer, L., Kim, Y., & Ronen, O. (1997). Automatic pronunciation scoring for language instruction. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Vol. 2, pp. 1471–1474). doi: 10.1109/ICASSP.1997.596227

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675–701.

Fries, C. C. (1945). *Teaching and learning English as a foreign language.* Ann Arbor: University of Michigan Press.

Fulcher, G. (2000). The 'communicative' legacy in language testing. *System*(28), 483-497. doi: 10.1016/S0346-251X(00)00033-6

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In *Proceedings of the Computer-based assessment of foreign language speaking skills* (p. 29-51).

Gamon, M., Chodorow, M., Leacock, C., & Tetreault, J. (2013). Grammatical error detection in automatic essay scoring and feedback. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (p. 251-266). New York: Routledge. doi: 10.4324/9780203122761

Gatbonton, E., & Trofimovich, P. (2008). The ethnic group affiliation and l2 proficiency link: Empirical evidence. *Language Awareness*, *17*(3), 229–248. doi: 10.1080/09658410802146867

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum* (pp. 240–254). Retrieved from `https://corpus.mml.cam.ac.uk/faq/SLRF2013Geertzenetal.pdf`

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379–399. doi: 10.1177/0265532210364407

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP* (Vol. 1, p. 517-520). doi: 10.1109/ICASSP.1992.225858

Grabowski, K. C. (2016). Assessing pragmatic competence. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 165–180). Boston; Berlin: De Gruyter Mouton. doi: 10.1515/9781614513827

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193–202. doi: 10.3758/BF03195564

Granger, S. (1998). The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3–18). London: Routledge. doi: 10.4324/9781315841342

Granger, S., Hung, J., & Petch-Tyson, S. (2002). *Computer learner corpora, second language acquisition, and foreign language teaching* (Vol. 6). Amsterdam; Philadelphia: John Benjamins Publishing. doi: 10.1075/lllt.6

Gretter, R., Matassoni, M., Allgaier, K., Tchistiakova, S., & Falavigna, D. (2019). Automatic assessment of spoken language proficiency of non-native children. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (p. 7435-7439). doi: 10.1109/ICASSP.2019.8683268

Gretter, R., Matassoni, M., Bannò, S., & Falavigna, D. (2020). TLT-school: a corpus of non native children speech. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020).* Retrieved from `https://aclanthology.org/2020.lrec-1.47`

Grice, M., Savino, M., & Roettger, T. B. (2018). Word final schwa is driven by intonation: The case of Bari Italian. *The Journal of the Acoustical Society of America*, *143*(4), 2474–2486. doi: 10.1121/1.5030923

Groot, P. (2000). Computer assisted second language vocabulary acquisition. *Language Learning & Technology*, *1*(4), 56-76. Retrieved from `https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/2d62539a-383d-450a-bf46-5bda8f588f7d/content`

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL quarterly*, *38*(2), 201–223. doi: 10.2307/3588378

Hamada, Y. (2019). Shadowing: What is it? how to use it. where will it go? *RELC Journal*, *50*(3), 386–393. doi: 10.1177/0033688218771380

Hardison, D. (2004). Generalization of computer assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, *1*(8), 34-52. Retrieved from https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/274d274e-5483-4ee2-b242-d828985a0fce/content

Harley, B. (1980). Interlanguage units and their relations. *Interlanguage Studies Bulletin*, *5*, 3–30. Retrieved from https://www.jstor.org/stable/43135250

Harris, D. P., & Palmer, L. A. (1986). *CELT listening form L-A, structure form S-A, vocabulary form V-A (2nd ed.)*. New York: McGraw-Hill.

Hasan, M. M., & Khaing, H. O. (2008). Learner corpus and its application to automatic level checking using machine learning algorithms. In *2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* (Vol. 1, pp. 25–28). doi: 10.1109/ECTICON.2008.4600364

Hassanali, K.-n., Liu, Y., & Solorio, T. (2012). Coherence in child language narratives: A case study of annotation and automatic prediction of coherence. In *3rd Workshop on Child, Computer and Interaction*. Retrieved from https://www.isca-speech.org/archive_v0/wocci_2012/papers/wc12_007.pdf

Hawkins, J., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, *1*(1), 1–23. doi: 10.1017/S2041536210000103

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004* (pp. 185–192). Retrieved from https://aclanthology.org/N04-1024

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, *25*(2), 282–306. doi: 10.1016/j.csl.2010.06.001

Hinkel, E. (2010). Integrating the four skills: Current and historical perspectives. In R. Kaplan (Ed.), *The Oxford Handbook of Applied Linguistics* (pp. 110–123). Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780195384253.001.0001

Hirschi, K. (2020). Duolingo - [Review]. In *Proceedings of the 11th Pronunciation in Second Language Learning and Teaching Conference* (p. 354-359). Retrieved from `https://www.iastatedigitalpress.com/psllt/article/id/15439/`

Howson, P. (2013). *The English effect*. London: British Council. Retrieved from `https://www.britishcouncil.org/sites/default/files/english-effect-report-v2.pdf`

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451-3460. doi: 10.1109/TASLP.2021.3122291

Hsu, W.-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., ... Auli, M. (2021). Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. In *Proceedings of interspeech 2021* (pp. 721–725). doi: 10.21437/Interspeech.2021-236

Huang, L.-f., Lin, Y.-l., & Gráf, T. (2023). Development of the use of discourse markers across different fluency levels of cefr: A learner corpus analysis. *Pragmatics*, *33*(1), 49–77. doi: 10.1075/prag.21016.hua

Huang, Y., Gertzen, J., Baker, R., Korhonen, A., & Alexopoulou, T. (2017). *The EF Cambridge Open Language Database (EFCAMDAT): Information for users.* Retrieved from `https://corpus.mml.cam.ac.uk/faq/EFCamDat-Intro_release2.pdf`

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics*, *23*(1), 28–54. doi: 10.1075/ijcl.16080.hua

Hudson, T. (2001). Indicators for cross-cultural pragmatic instruction: some quantitative tools. In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 283–300). Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139524797

Hudson, T., Detmer, E., & Brown, J. E. (1995). *Developing prototypic measures of cross-cultural pragmatics.* Honolulu: Second Language Teaching and Curriculum Center, University of Hawai'i at Manoa.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, *4*(91), 653-667. doi: 10.1111/j.1540-4781.2007.00627_5.x

Hunt, K. W. (1965). Grammatical structures written at three grade levels. *NCTE Research Report*(3).

Hymes, D. (1972). On communicative competence. In J. . J. Holmes (Ed.), *Sociolinguistics: Selected readings* (p. 269-293). Harmondsworth: Penguin.

Hönig, F., Batliner, A., Weilhammer, K., & Nöth, E. (2010). Automatic assessment of non-native prosody for English as L2. In *Proceedings of the 5th speech prosody international conference* (pp. 580–585). Retrieved from `https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=9e8108253b7ecc13c9fe6fe9894491e0edc29371`

Irujo, S. (1986). Don't put your leg in your mouth: Transfer in the acquisition of idioms in a second language. *Tesol Quarterly*, *20*(2), 287–304. doi: 10.2307/3586545

Isaacs, T. (2014). Assessing pronunciation. In J. Kunnan (Ed.), *The companion to language assessment* (Vol. 1, p. 140-155). Oxford: Wiley & Sons. doi: 10.1002/9781118411360.wbcla012

Isaacs, T. (2017a). Assessing speaking. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 131–146). Boston: DeGruyter Mouton. doi: 10.1515/9781614513827

Isaacs, T. (2017b). Fully automated speaking assessment: changes to proficiency testing and the role of pronunciation. In O. Kang, R. I. Thomson, & J. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 570–582). London; New York: Routledge. doi: 10.4324/9781315145006

Isaacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly*, *15*(3), 273–293. doi: 10.1080/15434303.2018.1472264

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*(3), 475–505. doi: 10.1017/S0272263112000150

Isaacs, T., & Trofimovich, P. (2016). *Second language pronunciation assessment: Interdisciplinary perspectives*. Blue Ridge Summit: Multilingual Matters. doi: 10.21832/9781783096855

Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. In *Corpora and language technologies in teaching, learning and research* (p. 3-

11). University of Strathclyde Press. Retrieved from https://www.researchgate.net/profile/Shinichiro-Ishikawa-2/publication/267226494_A_new_horizon_in_learner_corpus_studies_The_aim_of_the_ICNALE_project/links/5ad2bc49458515c60f51df6c/A-new-horizon-in-learner-corpus-studies-The-aim-of-the-ICNALE-project.pdf

Ishikawa, S. (2020). Aim of the ICNALE GRA project: Global collaboration to collect ratings of asian learners' L2 english essays and speeches from an ELF perspective. *Learner Corpus Studies in Asia and the World*, *5*, 121-144. Retrieved from https://www.researchgate.net/profile/Shinichiro-Ishikawa-2/publication/364690644_Aim_of_the_ICNALE_GRA_Project_Global_Collaboration_to_Collect_Ratings_of_Asian_Learners'_L2_English_Essays_and_Speeches_from_an_ELF_Perspective/links/635759728d4484154a30cddc/Aim-of-the-ICNALE-GRA-Project-Global-Collaboration-to-Collect-Ratings-of-Asian-Learners-L2-English-Essays-and-Speeches-from-an-ELF-Perspective.pdf

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in japanese as a foreign language. *Language Assessment Quarterly: An International Journal*, *3*(2), 151–169. doi: 10.1207/s15434311laq0302_4

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49. doi: 10.1093/applin/amm017

Iwashita, N., & Vasquez, C. (2015). *An examination of discourse competence at different proficiency levels in IELTS Speaking Part 2* (Tech. Rep. No. 5). IELTS Research Reports Online Series. Retrieved from https://www.ielts.org/for-researchers/research-reports/online-series-2015-5

Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE corpus: Exploiting the language learners' speech database for research and education. *International journal of the computer, the internet and management*, *12*(2), 119–125. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=cdd1934256938e9da1661b9d56e2b0ded5ac5b5c

Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., & Isahara, H. (2003). Automatic error detection in the japanese learners' english spoken data. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics* (pp. 145–148). doi: 10.3115/1075178.1075202

Jackendoff, R. (1995). The boundaries of the lexicon. In M. Everaert, E. J. Van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms: structural and psychological perspectives* (pp. 133–165). Hillsdale: Lawrence Erlbaum Associates. doi: 10.4324/9781315806501

James, C. (1998). *Errors in language learning and use: Exploring error analysis*. London: Routledge. doi: 10.4324/9781315842912

Jenkins, J. (2009). (Un)pleasant? (In)correct? (Un)intelligible? ELF speakers' perceptions of their accents. In A. Mauranen & E. Ranta (Eds.), *English as a lingua franca: studies and findings* (pp. 10–36). Newcastle: Cambridge Scholars Publishing.

Jeon, E. H., & In'nami, Y. (2022). *Understanding L2 proficiency: Theoretical and meta-analytic investigations*. Amsterdam; Philadelphia: John Benjamins. doi: 10.1075/bpa.13

Jeon, E. H., In'nami, Y., & Koizumi, R. (2022). L2 speaking and its external correlates: A meta-analysis. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (p. 339-367). Amsterdam; Philadelphia: John Benjamins. doi: 10.1075/bpa.13.11jeo

Jespersen, O. (1924). *The philosophy of grammar*. London: George Allen and Unwin. Retrieved from `https://archive.org/details/in.ernet.dli.2015.282299`

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 english speakers: Temporal measures and stimulated recall. *Language Learning*, *64*(4), 809–854. doi: 10.1111/lang.12084

Kamimura, K., & Takano, K. (2019). Pronunciation error detection in voice input for correct word suggestion. In *2019 International Electronics Symposium (IES)* (pp. 490–493). doi: 10.1109/ELECSYM.2019.8901539

Kang, O. (2013). Relative impact of pronunciation features on ratings of non-native speakers' oral proficiency. In *Proceedings of the 4th pronunciation in second language learning and teaching conference* (pp. 10–15). Retrieved from `https://www.iastatedigitalpress.com/psllt/article/15198/galley/13721/view/`

Kang, O., & Moran, M. (2014). Functional loads of pronunciation features in nonnative speakers' oral assessment. *Tesol Quarterly*, *48*(1), 176–187. doi: 10.1002/tesq.152

Kang, O., Thomson, R. I., & Murphy, J. J. (2017). *The Routledge Handbook of Contemporary English Pronunciation*. London; New York: Routledge. doi: 10.4324/9781315145006

Kanzawa, K., Mitsunaga, H., Edmonds, G., Hato, Y., Tsubota, Y., Mori, M., & Shimizu, Y. (2022). Development and administration of a Skype-based English speaking test in a Japanese high school. *Bulletin of Kyoto Institute of Technology*, *14*, 27–47. Retrieved from `https://repository.lib.kit.ac.jp/repo/repository/10212/2511/14_2_K.Kanzawa.pdf`

Karhila, R., Smolander, A.-R., Ylinen, S., & Kurimo, M. (2019). Transparent pronunciation scoring using articulatorily weighted phoneme edit distance. In *Proceedings of Interspeech 2019* (p. 1866-1870). doi: 10.21437/Interspeech.2019-1785

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the Acoustical Society of America*, *138*(2), 817–832. doi: 10.1121/1.4926561

Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *Modern Language Journal*, *2*(28), 136-150. doi: 10.2307/317331

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the twenty-eighth International Joint Conference on Artificial Intelligence (IJCAI-19)* (pp. 6300–6308). doi: 10.24963/ijcai.2019/879

Kecskes, I. (2007). Formulaic language in English Lingua Franca. In I. Kecskes & L. R. Horn (Eds.), *Explorations in pragmatics: Linguistic, cognitive and intercultural aspects* (pp. 191–219). Berlin; New York: Mouton de Gruyter. doi: 10.1515/9783110198843

Kelly, L. (1969). *25 centuries of language teaching: an inquiry into the science, art, and development of language teaching methodology, 500 b.c.-1969*. Rowley: Newbury House.

Kim, E., Jeon, J.-J., Seo, H., & Kim, H. (2022). Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning. In *Proceedings of Interspeech 2022* (pp. 1411–1415). doi: 10.21437/Interspeech.2022-10245

Kingma, D., & Ba, J. (2015). Adam: a method for stochastic optimization. In *International conference on learning representations.*

Knill, K. M., Gales, M. J. F., Manakul, P., & Caines, A. (2019). Automatic grammatical error detection of non-native spoken learner English. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (p. 8127-8131). doi: 10.1109/ICASSP.2019.8683080

Koizumi, R., In'nami, Y., & Jeon, E. H. (2022). L2 speaking and its internal correlates: A meta-analysis. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (p. 309-338). Amsterdam; Philadelphia: John Benjamins. doi: 10.1075/bpa.13.10koi

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, *20*(2), 148–161. doi: 10.1016/j.jslw.2011.02.001

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164. doi: 10.1016/j.system.2004.01.001

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods*, *50*(3), 1030–1046. doi: 10.3758/s13428-017-0924-4

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, *49*(4), 757–786. doi: 10.1002/tesq.194

Kyriakopoulos, K., Knill, K. M., & Gales, M. J. F. (2018). A deep learning approach to assessing non-native pronunciation of English using phone distances. In *Proceedings of Interspeech 2018* (pp. 1626–1630). doi: 10.21437/Interspeech.2018-1087

Kyriakopoulos, K., Knill, K. M., & Gales, M. J. F. (2019). A deep learning approach to automatic characterisation of rhythm in non-native English speech. In *Proceedings of Interspeech 2019* (pp. 1836–1840). doi: 10.21437/Interspeech.2019-3186

Lado, R. (1961). *Language testing: the construction and use of foreign language tests.* London: Longman.

Lambert, C., & Nakamura, S. (2019). Proficiency-related variation in syntactic complexity: A study of English L1 and L2 oral descriptive discourse. *International Journal of Applied Linguistics*, *29*(2), 248–264. doi: 10.1111/ijal.12224

Landauer, T. K. (2003). Automatic essay assessment. *Assessment in education: principles, policy and practice*, *10*(3), 295-308. doi: 10.1080/0969594032000148154

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259–284. doi: 10.1080/01638539809545028

Landauer, T. K., Laham, D., & Foltz, P. W. (2002). Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: a cross-disciplinary perspective* (pp. 87–112). New York: Routledge. doi: 10.4324/9781410606860

Laufer, B. (2000). Avoidance of idioms in a second language: The effect of L1-L2 degree of similarity. *Studia linguistica*, *54*(2), 186–196. doi: 10.1111/1467-9582.00059

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, *16*(3), 307–322. doi: 10.1093/applin/16.3.307

Lazaraton, A. (2014). Second language speaking. In M. Celce-Murcia, D. M. Brinton, & M. A. Snow (Eds.), *Teaching English as a second or foreign language (Fourth edition)* (pp. 110–123). Boston: National Geographic Learning.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning* (pp. 1188–1196).

Leacock, C., & Chodorow, M. (2002). Automated grammatical error detection. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: a cross-disciplinary perspective* (p. 195-207). Mahwah, NJ: Lawrence Erlbaum Associates. doi: 10.4324/9781410606860

Lebanon, G., Mao, Y., & Dillon, J. (2007). The locally weighted bag of words framework for document representation. *Journal of Machine Learning Research*, *8*(10). Retrieved from `http://jmlr.org/papers/v8/lebanon07a.html`

Lee, I. (2002). Teaching coherence to ESL students: A classroom inquiry. *Journal of second language writing*, *11*(2), 135–159. doi: 10.1016/S1060-3743(02)00065-6

Leech, G. (1983). *Principles of pragmatics*. London: Longman.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language learning*, *40*(3), 387–417. doi: 10.1111/j.1467-1770.1990.tb00669.x

Lennon, P. (2008). Contrastive analysis, error analysis, interlanguage. In S. Gramley & V. Gramley (Eds.), *Bielefeld Introduction to Applied Linguistics. A Course Book* (pp. 51–60). Bielefeld: Aisthesis.

Leong, C. W., Klebanov, B. B., Hamill, C., Stemle, E., Ubale, R., & Chen, X. (2020). A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the second workshop on figurative language processing* (pp. 18–29). doi: 10.18653/v1/2020.figlang-1.3

Levis, J. (2018). *Intelligibility, oral communication, and the teaching of pronunciation.* Cambridge: Cambridge University Press. doi: 10.1017/9781108241564

Levis, J. M. (1999). Intonation in theory and practice, revisited. *TESOL quarterly*, *33*(1), 37–63. doi: 10.2307/3588190

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL quarterly*, *39*(3), 369–377. doi: 10.2307/3588485

Levis, J. M. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation*, *6*(3), 310–328. doi: 10.1075/jslp.20050.lev

Levis, J. M., Derwing, T. M., & Munro, M. J. (2022). *The evolution of pronunciation teaching and research: 25 years of intelligibility, comprehensibility, and accentedness.* Amsterdam: Benjamins. doi: 10.1075/bct.121

Leńko-Szymańska, A. (2015). The English Vocabulary Profile as a benchmark for assigning levels to learner corpus data. In M. Callies & S. Götz (Eds.), *Learner corpora in language testing and assessment.* Amsterdam; Philadelphia: John Benjamins. doi: 10.1075/scl.70

Lieberman, H., Faaborg, A., Daher, W., & Espinosa, J. (2005). How to wreck a nice beach you sing calm incense. In *Proceedings of the 10th international conference on intelligent user interfaces* (pp. 278–280). doi: 10.1145/1040830.1040898

Lin, B., & Wang, L. (2021). Attention-based multi-encoder automatic pronunciation assessment. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7743–7747). doi: 10.1109/ICASSP39728.2021.9414451

Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, *15*(3), 294–309. doi: 10.1080/15434303.2018.1472265

Liu, J. (2006). *Measuring interlanguage pragmatic knowledge of EFL learners.* Frankfurt: Peter Lang. Retrieved from https://www.peterlang.com/document/1100119

Liu, Z., Xu, G., Liu, T., Fu, W., Qi, Y., Ding, W., ... others (2020). Dolphin: a spoken language proficiency assessment system for elementary education. In *Proceedings of The Web Conference 2020* (pp. 2641–2647). doi: 10.1145/3366423.3380018

Lo, Y.-C., Chen, J.-J., Yang, C., & Chang, J. (2018, August). Cool English: a grammatical error correction system based on large learner corpora. In *Proceedings of the 27th international conference on computational linguistics: System demonstrations* (pp. 82–85). Santa Fe, New Mexico: Association for Computational Linguistics. Retrieved from https://aclanthology.org/C18-2018

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *Proceedings of the international conference on learning representations.*

Loukina, A., Lopez, M., Evanini, K., Suendermann-Oeft, D., & Zechner, K. (2015). Expert and crowdsourced annotation of pronunciation errors for automatic scoring systems. In *Proceedings of Interspeech 2015* (pp. 2809–2813). Retrieved from https://www.isca-speech.org/archive/pdfs/interspeech_2015/loukina15b_interspeech.pdf

Lowie, A., M., V., & Van Dijk, M. (2018). The acquisition of L2 speaking: a dynamic perspective. In R. Alonso Alonso (Ed.), *Speaking in a second language* (pp. 105–125). Amsterdam; Philadelphia: John Benjamins Publishing Company. doi: 10.1075/aals.17

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. *TESOL quarterly*, *45*(1), 36–62. doi: 10.5054/tq.2011.240859

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208. doi: 10.1111/j.1540-4781.2011.01232_1.x

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, *34*(4), 493–511. doi: 10.1177/0265532217710675

Lu, Y., Bannò, S., & Gales, M. (2022). On assessing and developing spoken 'grammatical error correction' systems. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (pp. 51–60). doi: 10.18653/v1/2022.bea-1.9

Lu, Y., Gales, M. J. F., Knill, K. M., Manakul, P., Wang, L., & Wang, Y. (2019). Impact of ASR performance on spoken grammatical error detection. *Proceedings of Interspeech 2019*, 1876–1880. doi: 10.21437/Interspeech.2019-1706

Lu, Y., Gales, M. J. F., Knill, K. M., Manakul, P., & Wang, Y. (2019). Disfluency detection for spoken learner English. In *Proceedings of the 8th Workshop on Speech and Language Technology for Education (SLaTE)* (pp. 74–78). doi: 10.21437/SLaTE.2019-14

Lu, Y., Gales, M. J. F., & Wang, Y. (2020). Spoken language 'grammatical error correction'. *Proceedings of Interspeech 2020*, 3840–3844. doi: 10.21437/Interspeech.2020-1852

Ludlow, K. (2020). *Official quick guide to Linguaskill.* Cambridge: Cambridge University Press. Retrieved from `https://www.cambridgeenglish.org/es/Images/628063-official-quick-guide-to-linguaskil.pdf`

Lundeberg, O. (1929). Recent developments in audition-speech tests. *Modern Language Journal*, *4*(14), 193-202. doi: 10.2307/314375

Lyons, J. (1968). *Introduction to theoretical linguistics.* Cambridge University Press.

Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in second language acquisition*, *29*(4), 539–556. doi: 10.1017/S0272263107070428

Malinin, A., Ragni, A., Knill, K. M., & Gales, M. J. F. (2017). Incorporating uncertainty into deep learning for spoken language assessment. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 45–50). doi: 10.18653/v1/P17-2008

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development.* New York: Springer. doi: 10.1057/9780230511804

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, *8*(3), 243–281. doi: 10.1515/text.1.1988.8.3.243

Mao, R., Lin, C., & Guerin, F. (2019). End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3888–3898). doi: 10.18653/v1/P19-1378

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, *19*(2), 313–330. Retrieved from https://aclanthology.org/J93-2004

McCarthy, M. (1990). *Vocabulary.* Oxford: Oxford University Press.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written communication*, *27*(1), 57–86. doi: 10.1177/0741088309351547

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language testing*, *4*(2), 142–154. doi: 10.1177/026553228700400202

Menzel, W., Atwell, E., Bonaventura, P., Herron, D., Howarth, P., Morton, R., & Souter, C. (2000). The ISLE corpus of non-native spoken English. In *Proceedings of LREC 2000: Language Resources and Evaluation Conference, vol. 2* (pp. 957–964). Retrieved from http://www.lrec-conf.org/proceedings/lrec2000/pdf/313.pdf

Meteer, M., Taylor, A., MacIntyre, R., & Iyer, R. (1995). *Dysfluency Annotation Stylebook for the Switchboard Corpus* (Tech. Rep.). Linguistic Data Consortium. Retrieved from https://www.cs.brandeis.edu/~cs140b/CS140b_docs/DysfluencyGuide.pdf

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* ArXiv. doi: 10.48550/arXiv.1301.3781

Minematsu, N., Asakawa, S., & Hirose, K. (2006). Structural representation of the pronunciation and its use for CALL. In *2006 IEEE Spoken Language Technology Workshop* (pp. 126–129). doi: 10.1109/SLT.2006.326833

Mizumoto, T., Hayashibe, Y., Komachi, M., Nagata, M., & Matsumoto, Y. (2012). The effect of learner corpus size in grammatical error correction of ESL writings. In *Proceedings of COLING 2012: Posters* (pp. 863–872). Retrieved from https://aclanthology.org/C12-2084

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, *17*(1), 21–48. Retrieved from https://aclanthology.org/J91-1002

Morrow, K. (1977). *Techniques of evaluation for a notional syllabus.* London: Royal Society of Arts.

Morrow, K. (1979). Communicative language testing: Revolution or evolution? In C. . K. Johnson (Ed.), *The communicative approach to language teaching* (p. 143-157). Oxford: Oxford University Press.

Moyer, A. (2013). *Foreign accent: The phenomenon of non-native speech.* Cambridge: Cambridge University Press.

Müller, P., De Wet, F., C. Van Der Walt, & Niesler, T. (2009). Automatically assessing the oral proficiency of proficient L2 speakers. In *Proceedings of the 2nd Workshop on Speech and Language Technology for Education (SLaTE)* (pp. 29–32). Retrieved from `https://www.isca-speech.org/archive_v0/slate_2009/papers/sla9_029.pdf`

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, *45*(1), 73–97. doi: 10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, *52*(7-8), 626–637. doi: 10.1016/j.specom.2010.02.013

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part i. *Journal of Applied Measurement*, *4*(4), 386–422.

Napoles, C., Sakaguchi, K., Post, M., & Tetreault, J. (2015). Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers* (pp. 588–593). doi: 10.3115/v1/P15-2097

Nation, P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nemenyi, P. B. (1963). *Distribution-free multiple comparisons* (Unpublished doctoral dissertation). Princeton University.

Neri, A., Cucchiarini, C., & Strik, H. (2003). Automatic speech recognition for second language learning: How and why it actually works. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)* (pp. 1157–1160).

Neumeyer, L., Franco, H., Weintraub, M., & Price, P. (1996). Automatic text-independent pronunciation scoring of foreign language student speech. In *Proceeding of the 4th International*

Conference on Spoken Language Processing (ICSLP'96) (Vol. 3, pp. 1457–1460). doi: 10.1109/ICSLP.1996.607890

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task* (pp. 1–14). doi: 10.3115/v1/W14-1701

Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., & Tetreault, J. (2013). The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task* (pp. 1–12). Retrieved from `https://aclanthology.org/W13-3601`

Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the corpus linguistics 2003 conference* (pp. 572–581).

Oller, J. W. (1973). Discrete-point tests versus tests of integrative skills. In J. Oller & J. Richards (Eds.), *Focus on the learner: pragmatic perspectives for the language teacher* (pp. 184–200). Rowley: Newbury House.

Oller, J. W. (1979). *Language tests at school.* London: Longman.

Oller, J. W. (1983). A consensus for the eighties? In J. W. Oller (Ed.), *Issues in language testing research* (pp. 351–356). New York: Newbury House Publishers.

O'Loughlin, K. J. (2001). *The equivalence of direct and semi-direct speaking tests.* Cambridge: Cambridge University Press.

OpenAI. (2023). *GPT-4 Technical Report.* doi: 10.48550/arXiv.2303.08774

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics*, *24*(4), 492–518. doi: 10.1093/applin/24.4.492

O'Keeffe, A., & Mark, G. (2017). The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, *22*(4), 457–489. doi: 10.1075/ijcl.14086.oke

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238–243. Retrieved from `https://www.jstor.org/stable/20371545`

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, *14*(2), 210–225. doi: 10.1007/BF01419938

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318). doi: 10.3115/1073083.1073135

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–225). London: Longman.

Peng, L., Fu, K., Lin, B., Ke, D., & Zhan, J. (2021). A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis. In *Proceedings of interspeech 2021* (pp. 4448–4452). doi: 10.21437/Interspeech.2021-1344

Pennington, M. C., & Rogerson-Revell, P. (2019a). *English pronunciation teaching and research: Contemporary perspectives*. London: Palgrave Macmillan. doi: 10.1057/978-1-137-47677-7

Pennington, M. C., & Rogerson-Revell, P. (2019b). Using technology for pronunciation teaching, learning, and assessment. In *English pronunciation teaching and research: Contemporary perspectives* (pp. 235–286). London: Palgrave Macmillan UK. doi: 10.1057/978-1-137-47677-7

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). doi: 10.18653/v1/N18-1202

Pfingsthorn, J. (2013). *Variability in learner errors as a reflection of the CLT paradigm shift* (P. Lang, Ed.). Frankfurt am Main. doi: 10.3726/978-3-653-02772-3

Purpura, J. (2008). Assessing communicative language ability: Models and their components. In E. Shonany & N. Hornberger (Eds.), *Encyclopedia of language and education (2nd edition)* (Vol. 7, pp. 5–68). New York: Springer. doi: 10.1007/978-0-387-30424-3

Purpura, J. (2014). Assessing grammar. In J. Kunnan (Ed.), *The companion to language assessment* (Vol. 1, p. 100-124). Oxford: Wiley & Sons. doi: 10.1002/9781118411360.wbcla147

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly*, *6*(2), 113-125. doi: 10.1080/15434300902800059

Qian, M., Jancovic, M., & Russell, M. (2019). The University of Birmingham 2019 Spoken CALL Shared Task Systems: Exploring the importance of word order in text processing. In *Proceedings of the 8th Workshop on Speech and Language Technology for Education (SLaTE)* (pp. 11–15). doi: 10.21437/SLaTE.2019-3

Qian, X., Meng, H., & Soong, F. K. (2012). The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training. In *Proceedings of interspeech 2012.* doi: 10.21437/Interspeech.2012-238

Qian, Y., Lange, P., & Evanini, K. (2019). Automatic speech recognition for automated speech scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment* (pp. 61–74). New York: Routledge. doi: 10.4324/9781315165103

Qian, Y., Lange, P., Evanini, K., Pugh, R., Ubale, R., Mulholland, M., & Wang, X. (2019). Neural approaches to automated speech scoring of monologue and dialogue responses. In *Icassp 2019 - 2019 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 8112-8116). doi: 10.1109/ICASSP.2019.8683717

Qian, Y., Ubale, R., Mulholland, M., Evanini, K., & Wang, X. (2018). A prompt-aware neural network approach to content-based scoring of non-native spontaneous speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (p. 979-986). doi: 10.1109/SLT.2018.8639697

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... others (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*(140), 1–67.

Raina, V., Gales, M. J. F., & Knill, K. M. (2020). Universal adversarial attacks on spoken language assessment systems. In *Proceedings of Interspeech 2020* (pp. 3855–3859). doi: 10.21437/Interspeech.2020-1890

Ramanarayanan, V. (2020). Design and development of a human-machine dialog corpus for the automated assessment of conversational English proficiency. In *Proceedings of Interspeech 2020* (pp. 419–423). doi: 10.21437/Interspeech.2020-1988

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292. doi: 10.1016/S0010-0277(00)00101-3

Raux, A., & Kawahara, T. (2002). Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*. doi: 10.21437/ICSLP .2002-241

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press. doi: 10.1017/ CBO9780511732942

Read, J. (2013). Second language vocabulary assessment. *Language Teaching*, *46*(1), 41–52. doi: 10.1017/S0261444812000377

Riazantseva, A. (2001). Second language proficiency and pausing a study of Russian speakers of English. *Studies in Second Language Acquisition*, *23*(4), 497–526. doi: 10.1017/ S027226310100403X

Richard, J., Platt, J., & Weber, H. (1985). *Longman dictionary of applied linguistics*. London: Longman.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse processes*, *14*(4), 423–441. doi: 10.1080/01638539109544795

Roever, C. (2014). Assessing pragmatics. In J. Kunnan (Ed.), *The companion to language assessment* (Vol. 1, p. 125-139). Oxford: Wiley & Sons. doi: 10.1002/9781118411360.wbcla057

Rossiter, M. J. (2009). Perceptions of l2 fluency by native and non-native speakers of english. *Canadian Modern Language Review*, *65*(3), 395–412. doi: 10.3138/cmlr.65.3.395

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of intellimetric™essay scoring system. *The Journal of Technology, Learning and Assessment*, *4*(4). Retrieved from `https://ejournals.bc.edu/index.php/jtla/article/view/1651`

S.-W Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, Lakhotia, K., Y.Y. Lin, ... H.-Y. Lee (2021). SUPERB: Speech Processing Universal PERformance Benchmark. In *Proceedings of Interspeech 2021* (pp. 1194–1198). doi: 10.21437/Interspeech.2021-1775

Sahlgren, M. (2005). An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering* (p. 1-9).

Saito, K. (2020). Multi- or single-word units? the role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*, *70*(2), 548–588. doi: 10.1111/lang.12387

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217–240. doi: 10.1017/S0142716414000502

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., ... others (2017). English conversational telephone speech recognition by humans and machines. In *Proceedings of Interspeech 2017* (pp. 132–136). doi: 10.21437/Interspeech.2017-405

Scholfield, P. (1995). *Quantifying language.* Clevedon: Multilingual matters.

Segalowitz, N. (2010). *Cognitive bases of second language fluency.* New York: Routledge. doi: 10.4324/9780203851357

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 79–95. doi: 10.1515/iral-2016-9991

Seidlhofer, B. (2018). Standard English and the dynamics of ELF variation. In J. Jenkins, W. Baker, & M. Dewey (Eds.), *The Routledge handbook of English as a lingua franca* (pp. 85–100). London; New York: Routledge. doi: 10.4324/9781315717173

Setter, J. (2006). Speech rhythm in world Englishes: The case of Hong Kong. *Tesol Quarterly*, *40*(4), 763–782. doi: 10.2307/40264307

Shatz, I. (2020). Refining and modifying the EFCAMDAT. *International Journal of Learner Corpus Research*, *6*(2), 220-223. doi: 10.1075/ijlcr.20009.sha

Sickinger, P., & Schneider, K. P. (2014). Pragmatic competence and the cefr: Pragmatic profiling as a link between theory and language use. *Linguistica*, *54*(1), 113–127. doi: 10.4312/linguistica.54.1.113-127

Singla, Y. K., Shah, J., Chen, C., & Shah, R. R. (2022). What do audio transformers hear? probing their representations for language delivery & structure. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)* (p. 910-925). doi: 10.1109/ICDMW58026.2022.00120

Skehan, P., & Foster, P. (2007). Complexity, accuracy, fluency and lexis in task-based performance: a meta-analysis of the Ealing Research. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy, and fluency in second language use, learning, and teaching* (p. 207-226). Brussels: University of Brussels Press. doi: 10.1075/lllt.32.09fos

Sokhatskyi, V., Zvyeryeva, O., Karaulov, I., & Tkanov, D. (2019). Embedding-based system for the text part of CALL v3 shared task. In *Proceedings of the 8th Workshop on Speech and Language Technology for Education (SLaTE)* (pp. 16–19). doi: 10.21437/SLaTE.2019-4

Somasundaran, S., Burstein, J., & Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical papers* (pp. 950–961). Retrieved from `https://aclanthology.org/C14-1090`

Soricut, R., & Marcu, D. (2006). Discourse generation using utility-trained coherence models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions* (pp. 803–810). Retrieved from `https://aclanthology.org/P06-2103`

Spearman, C. E. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, *15*, 201–293. doi: 10.2307/1412107

Spolsky, B. (1995). *Measured words: the development of objective language testing.* Oxford: Oxford University Press.

Stengers, H., Boers, F., Housen, A., & Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *International review of applied linguistics in language teaching*, *49*(4), 321–343. doi: 10.1515/iral.2011.017

Stevenson, M. (2016). A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition*, *42*, 1–16. doi: 10.1016/j.compcom.2016.05.001

Strik, H., & Cucchiarini, C. (1999). Automatic assessment of second language learners' fluency. In *Proceedings of the international congress of phonetic sciences (icphs) 1999.* Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_0759.pdf

Supnithi, T., Uchimoto, K., Saiga, T., Izumi, E., Virach, S., & Isahara, H. (2003). Automatic proficiency level checking based on SST corpus. In *Proceedings of ranlp 2003* (pp. 29–33). Retrieved from https://www.researchgate.net/publication/228523712_Automatic_Proficiency_Level_Checking_based_on_SST_Corpus

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143–167. doi: 10.1017/S0272263119000421

Suzuki, S., & Kormos, J. (2022). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*, 1–27. doi: 10.1017/S0272263121000899

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal*, *105*(2), 435–463. doi: 10.1111/modl.12706

Tajiri, T., Komachi, M., & Matsumoto, Y. (2012). Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers* (pp. 198–202). Retrieved from https://aclanthology.org/P12-2039

Takai, K., Heracleous, P., Yasuda, K., & Yoneyama, A. (2020). Deep learning-based automatic pronunciation assessment for second language learners. In *Proceedings of the 22nd International Conference on Human-Computer Interaction* (pp. 338–342). Springer. doi: 10.1007/978-3-030-50729-9_48

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT journal*, *65*(1), 71–79. doi: 10.1093/elt/ccq020

Tavakoli, P., Nakatsuhara, F., & Hunter, A. (2017). *Scoring validity of the Aptis speaking test: investigating fluency across tasks and levels of proficiency* (Tech. Rep. No. 7). ARAGs Research

Reports Online. Retrieved from https://www.britishcouncil.org/sites/default/files/tavakoli_et_al_layout.pdf

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–276). Amsterdam: John Benjamins. doi: 10.1075/lllt.11.15tav

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, *70*(2), 506–547. doi: 10.1111/lang.12384

Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn treebank: an overview. *Treebanks*, 5–22. doi: 10.1007/978-94-010-0201-1_1

Taylor, D. S. (1981). Non-native speakers and the rhythm of English. *International Review of Applied Linguistics in Language Teaching*, *19*(4), 219–226. doi: 10.1515/iral.1981.19.1-4.219

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, *97*(1), 77–101. doi: 10.1111/j.1540-4781.2012.01422.x

Thomas, J. (1983). Cross-cultural pragmatic failure. *Applied linguistics*, *4*(2), 91–112. doi: 10.1093/applin/4.2.91

Thomson, R. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *Calico Journal*, *3*(28), 744-765. Retrieved from https://www.jstor.org/stable/calicojournal.28.3.744

Townshend, B., Bernstein, J., Todic, O., & Warren, E. (1998). Estimation of spoken language proficiency. In *STiLL-Speech Technology in Language Learning.* Retrieved from https://www.isca-speech.org/archive/pdfs/still_1998/townshend98_still.pdf

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302–327. doi: 10.1093/applin/amw009

Uchihara, T., & Clenton, J. (2020). Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research*, *24*(4), 540–556. doi: 10.1177/1362168818799371

Uchihara, T., Eguchi, M., Clenton, J., Kyle, K., & Saito, K. (2022). To what extent is collocation knowledge associated with oral proficiency? a corpus-based approach to word association. *Language and Speech*, *65*(2), 311–336. doi: 10.1177/00238309211013865

Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, *28*, 79–105. doi: 10.1007/s40593-017-0142-3

Vajjala, S., & Rama, T. (2018). Experiments with universal CEFR classification. In *Proceedings of 13th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 147–153). doi: 10.18653/v1/W18-0515

van Dalen, R. C., Knill, K. M., & Gales, M. J. F. (2015). Automatically grading learners' English using a gaussian process. In *Proceedings of the 6th Workshop on Speech and Language Technology for Education (SLaTE)* (pp. 7–12). Retrieved from `https://www.isca-speech.org/archive/pdfs/slate_2015/dalen15_slate.pdf`

Van Doremalen, J., Cucchiarini, C., & Strik, H. (2009). Automatic detection of vowel pronunciation errors using multiple information sources. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 580–585). doi: 10.1109/ASRU.2009.5373335

van Els, T., Bongaerts, T., Extra, G., van Os, C., & Janssen-van Dieten, A. M. (1984). *Applied linguistics and the learning and teaching of foreign languages*. London: Edward Arnold.

Van Moere, A., & Downey, R. (2017). Technology and artificial intelligence in language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 341–357). Boston: DeGruyter Mouton. doi: 10.1515/9781614513827

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., A.N. Gomez, ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008). Retrieved from `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`

Wang, X., Bruno, J., Molloy, H., Evanini, K., & Zechner, K. (2017). Discourse annotation of non-native spontaneous spoken responses using the Rhetorical Structure Theory framework. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 263–268). doi: 10.18653/v1/P17-2041

Wang, X., Evanini, K., Qian, Y., & Mulholland, M. (2021). Automated scoring of spontaneous speech from young learners of English using transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)* (p. 705-712). doi: 10.1109/SLT48900.2021.9383553

Wang, X., Evanini, K., & Zechner, K. (2013). Coherence modeling for the automated assessment of spontaneous spoken responses. In *Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 814–819). Retrieved from `https://aclanthology.org/N13-1101`

Wang, X., Gyawali, B., Bruno, J. V., Molloy, H. R., Evanini, K., & Zechner, K. (2019). Using Rhetorical Structure Theory to assess discourse coherence for non-native spontaneous speech. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019* (pp. 153–162). doi: 10.18653/v1/W19-2719

Wang, Y., Gales, M. J. F., Knill, K. M., Kyriakopoulos, K., Malinin, A., van Dalen, R. C., & Rashid, M. (2018). Towards automatic assessment of spontaneous spoken english. *Speech Communication*, *104*, 47–56. doi: 10.1016/j.specom.2018.09.002

Wang, Y., Wang, Y., Dang, K., Liu, J., & Liu, Z. (2021). A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *12*(5), 1–51. doi: 10.1145/3474840

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language teaching research*, *10*(2), 157–180. doi: 10.1191/1362168806lr190oa

Wei, S., Hu, G., Hu, Y., & Wang, R.-H. (2009). A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication*, *51*(10), 896–905. doi: 10.1016/j.specom.2009.03.004

Weinreich, U. (1953). *Languages in contact: Findings and problems*. New York: Linguistic Circle of New York.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *3*(22), 281-300. doi: 10.1191/0265532205lt309oa

Weir, C. J., Vidaković, I., & Galaczi, E. D. (2013). *Measured constructs: A history of Cambridge English examinations, 1913-2012* (Vol. 37). Cambridge: Cambridge University Press.

Wennerstrom, A. (1994). Intonational meaning in english discourse: A study of non-native speakers. *Applied linguistics*, *15*(4), 399–420. doi: 10.1093/applin/15.4.399

Wester, M., & Mayo, C. (2014). Accent rating by native and non-native listeners. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (p. 7699-7703). doi: 10.1109/ICASSP.2014.6855098

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, *20*(1), 6–10. doi: 10.3758/BF03202594

Witt, S. M. (1999). *Use of speech recognition in computer assisted language learning* (Unpublished doctoral dissertation). University of Cambridge.

Witt, S. M. (2012). Automatic error detection in pronunciation training: Where we are and where we need to go. In *International symposium on automatic detection on errors in pronunciation training* (pp. 1–8). Retrieved from `https://www.researchgate.net/publication/250306074_Automatic_Error_Detection_in_Pronunciation_Training_Where_we_are_and_where_we_need_to_go`

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, *30*(2-3), 95–108. doi: 10.1016/S0167-6393(99)00044-8

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. (2020). 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). doi: 10.18653/v1/2020.emnlp-demos.6

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511519772

Wu, M., Li, K., W.-K. Leung, & Meng, H. (2021). Transformer based end-to-end mispronunciation detection and diagnosis. In *Proceedings of Interspeech 2021* (pp. 3954–3958). doi: 10.21437/Interspeech.2021-1467

Wu, X., Knill, K. M., Gales, M. J. F., & Malinin, A. (2020). Ensemble approaches for uncertainty in spoken language assessment. In *Proceedings of Interspeech 2020* (pp. 3860–3864). doi: 10.21437/Interspeech.2020-2238

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRaterSM v1.0* (Tech. Rep. No. 2). ETS Research Report Series. doi: 10.1002/j.2333-8504.2008.tb02148.x

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., ... Zweig, G. (2017). Towards human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2410–2423. doi: 10.1109/TASLP.2017.2756440

Xu, X., Kang, Y., Cao, S., Lin, B., & Ma, L. (2021). Explore wav2vec 2.0 for Mispronunciation Detection. In *Proceedings of Interspeech 2021* (pp. 4428–4432). doi: 10.21437/Interspeech .2021-777

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education*, *15*(4), 391–412. doi: 10.1207/S15324818AME1504_04

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*. Retrieved from `https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf`

Yannakoudakis, H., Andersen, Ø. E., Geranpayeh, A., Briscoe, T., & Nicholls, D. (2018). Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, *31*(3), 251–267. doi: 10.1080/08957347.2018.1464447

Yannakoudakis, H., & Briscoe, T. (2012). Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 33–43). Retrieved from `https://aclanthology.org/W12-2004`

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: Human language technologies* (Vol. 1, pp. 180–189). Retrieved from `https://aclanthology.org/P11-1019`

Yoon, S. Y., & Bhat, S. (2012). Assessment of ESL learners' syntactic competence based on similarity measures. In *Proceedings of the 2012 Joint Conference of Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 600–608). Retrieved from `https://aclanthology.org/D12-1055`

Yoon, S.-Y., Bhat, S., & Zechner, K. (2012). Vocabulary profile as a measure of vocabulary sophistication. In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP* (pp. 180–189). Retrieved from https://aclanthology.org/W12-2021

Yoon, S.-Y., Lu, X., & Zechner, K. (2019). Features measuring vocabulary and grammar. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment* (pp. 123–137). New York: Routledge. doi: 10.4324/9781315165103-8

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied linguistics*, *31*(2), 236–259. doi: 10.1093/applin/amp024

Yu, Z., Ramanarayanan, V., Suendermann-Oeft, D., Wang, X., Zechner, K., Chen, L., ... Qian, Y. (2015). Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 338–345). doi: 10.1109/ASRU.2015.7404814

Yuan, Z., & Briscoe, T. (2016). Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 380–386). doi: 10.18653/v1/N16-1042

Zechner, K., & Bejar, I. (2006). Towards automatic scoring of non-native spontaneous speech. In *Proceedings of the Human Language Technology Conference of the NAACL, main conference* (pp. 216–223). Retrieved from https://aclanthology.org/N06-1028

Zechner, K., Bejar, I. I., & Hemat, R. (2007). Toward an understanding of the role of speech recognition in nonnative speech assessment. *ETS Research Report Series*, *2007*(1), i–76. doi: 10.1002/j.2333-8504.2007.tb02044.x

Zechner, K., Higgins, D., & Xi, X. (2007). Speechrater™: a construct-driven approach to scoring spontaneous non-native speech. In *Proceedings of the Workshop on Speech and Language Technology for Education (SLaTE)* (pp. 128–131). Retrieved from https://www.isca-speech.org/archive_open/archive_papers/slate_2007/sle7_128.pdf

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883–895. doi: 10.1016/j.specom.2009.04.009

Zhang, J., Zhang, Z., Wang, Y., Yan, Z., Song, Q., Huang, Y., ... Wang, Y. (2021). speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Proceedings of Interspeech 2021.* doi: 10.21437/Interspeech.2021-1259

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R&D Connections*(21), 1-11. Retrieved from `https://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf`

Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., & Gutierrez-Osuna, R. (2018). L2-ARCTIC: A non-native English speech corpus. In *Proceedings of Interspeech 2018* (pp. 2783–2787). doi: 10.21437/Interspeech.2018-1110

Zhou, Z., Vajjala, S., & S.V. Mirnezami. (2019). Experiments on Non-native Speech Assessment and its Consistency. In *NLP4CALL 2019* (pp. 86–92). Retrieved from `https://aclanthology.org/W19-6309`

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*(1), 69-84. doi: 10.1016/j.system.2007.11.004

Zuskin, R. D. (1993). Assessing L2 sociolinguistic competence: In search of support from pragmatic theories. *Pragmatics and Language Learning*, *4*, 166-182. Retrieved from `https://files.eric.ed.gov/fulltext/ED396556.pdf`

Appendix A
TLT-school and TLT-GEC question prompts

B1

- Answer the following questions. - Would you prefer to live in a town or in a small village? Why?
- What's your favourite TV channel? Why?
- What's your favourite book? Why?
- Is fashion important to you? Why?
- What country would you like to visit in the future? Why?
- How do you use the Internet?
- Your friend doesn't do much sport; convince him that playing sports is useful and fun and is a healthy way to spend your free time. Give some advice on how to get started and get him involved in your activity.

A2

- Answer the questions in English: it's your first day in a new school. Introduce yourself to the class in English and talk about the following topics. - Greetings - Where you come

from - Your hobbies and what you do in your free time - Your family - Favourite foods and drinks - Your friends/mates - What you did last night or last weekend - What you want to do when you grow up (what are your plans for the future) and why.

- Nowadays sport is increasingly present in the life and interests of teenagers and probably also in yours. You will be asked questions that you will answer by considering your personal habits. Attention: it is very important that you explain the reasons for your answers! - Do you play sport?
- Do you prefer outdoor or indoor sports? Why?
- Talk about your favourite sport.
- What TV programmes on sport do you watch? Why?
- Why is it important to do sport regularly?

Appendix B

Linguaskill question prompts

Linguaskill General

The following question prompts are part of the Linguaskill practice materials.[7]

Part 1

You will be asked 8 questions. Listen to each question and answer after the tone.

For questions 1-4, you will have 10 seconds to speak. For questions 5-8 you will have 20 seconds to speak.

- What's your name?
- How do you spell your family name?
- Where are you from?
- Do you work or are you a student?
- What do enjoy doing at weekends?
- Do you get many opportunities to speak English?
- What's the best thing that happened to you last week?
- Where would you like to live in the future?

[7] cambridgeenglish.org/exams-and-tests/linguaskill/information-about-the-test/practice-materials/

Part 2

You will see 8 sentences on the screen. You will have 10 seconds to read each sentence aloud after the tone.

- The library is closed for staff training until 11 am.
- Mrs Hill would like to accept the invitation.
- The bus timetable can sometimes change at short notice.
- Thank you for coming to the film club's summer event.
- How easy will it be for students to find accommodation near the university?
- After you have finished making online payments, remember to log out of your account.
- A 'UV index' reading of 11 indicates an extreme risk of harm from the sun's rays.
- On average there are twice as many applicants for undergraduate degree courses as places available.

Part 3

You will have 1 minute to talk about a topic. First, you have 40 seconds to read the task and prepare what you are going to say. You will then have 1 minute to speak. Please speak for all the time you have.

Talk about a person you know that is special to you.

You should say:

- who the person is
- how you know the person
- why the person is special to you.

Part 4

You will have 1 minute to leave a message for an English-speaking friend about some visual information. First, you have 1 minute to look at the information and prepare what you are going

to say. You will then have 1 minute to leave your message. The visual information will stay on the screen. Please speak for all the time you have.

Your English-speaking friend needs to travel to a nearby city.

This table shows the different ways your friend could travel.

Leave a message for your friend, recommending a way to travel and explaining why you think this way to travel is best.

Part 5

You will hear five questions about a topic. First, you have 40 seconds to read the task. After you hear each question, you will have 20 seconds to give your answer. Please speak for all the time you have.

A researcher is writing a report about young people's leisure time. He wants to find out your opinion about the importance of leisure time for young people.

He will ask you questions about:

- daily leisure time
- playing sports
- being alone

- joining clubs
- too much leisure time

How important is it for young people to have some leisure time every day?

In your opinion is it a good idea for all young people to play sports?

Is it better for young people to spend their leisure time alone or with other people?

Some people say that all students should join university clubs. Do you agree?

How might having too much leisure time affect how well students are doing in their courses?

Linguaskill Business

Part 1

You will be asked 8 questions. Listen to each question and answer after the tone.

For questions 1-4, you will have 10 seconds to speak. For questions 5-8 you will have 20 seconds to speak.

- What's your name?
- How do you spell your family name?
- Where are you from?
- What's your job?
- How long have you been with your present company?
- How do you use English in your work?
- What are the opportunities for promotion in your current job?
- What will you do at work next week?

Part 2

You will see 8 sentences on the screen. You will have 10 seconds to read each sentence aloud after the tone.

- The team needs sales staff who can speak more than one language.

- The 5% discount is only on orders over $10,000.
- Have the long-term goals of the company changed?
- Your account will become active on receipt of the first payment.
- Mrs Atkins called to say that she is away at a marketing conference this week.
- The R&D budget has been frozen for five years but will increase again next January.
- The organisation, which has its headquarters in Canada, has now expanded into many European countries.
- The best way to reduce distribution costs is to use our subsidiary to transport goods.

Part 3

You will have 1 minute to talk about a topic. First, you have 40 seconds to read the task and prepare what you are going to say. You will then have 1 minute to speak. Please speak for all the time you have.

Talk about a training course you have attended for your work.

You should say:

- what the course was
- why you did the course
- whether you would recommend this course.

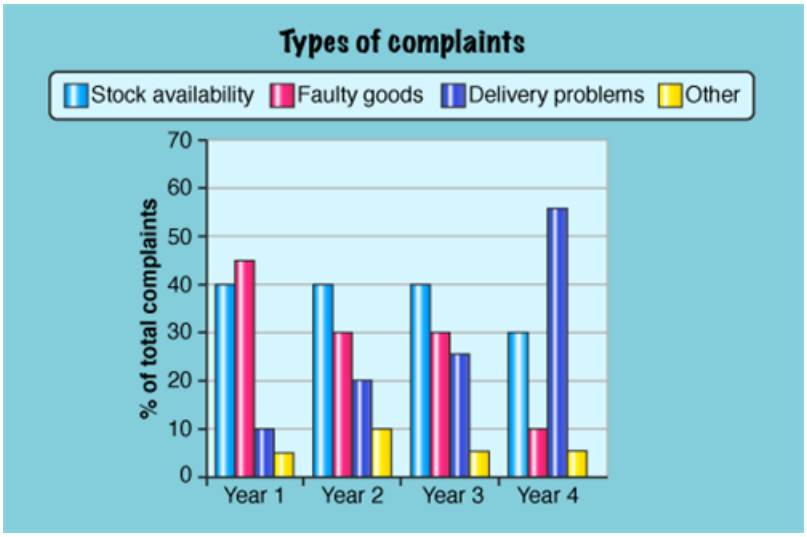
Part 4

You will have 1 minute to leave a message for an English-speaking friend about some visual information. First, you have 1 minute to look at the information and prepare what you are going to say. You will then have 1 minute to leave your message. The visual information will stay on the screen. Please speak for all the time you have.

Your manager has asked you for information about the types of complaints your company has received.

This chart shows the percentage of total complaints received during Years 1 – 4.

Look at the chart and then talk about the information, describing how the types of complaints changed during the four years.



Part 5

You will hear five questions about a topic. First, you have 40 seconds to read the task. After you hear each question, you will have 20 seconds to give your answer. Please speak for all the time you have.

A business owner is thinking about using sponsorship to publicise his company. He wants to find out your opinion about the best way to organise a sponsorship programme.

He will ask you questions about:

- benefits for companies
- who to sponsor
- length of sponsorship
- possible problems
- judging success

In your opinion, what are the benefits of companies offering sponsorship?

Would it be better to sponsor an individual or an organisation?

How long should a sponsorship programme last?

What problems could there be with a sponsorship programme?

How could a company judge whether its sponsorship has been successful?