Università di Pisa
PhD Program in Computer Science

Tracking Linguistic Abilities in Neural Language Models

Doctoral Thesis

Author:
Alessio Miaschi

Supervisors:
Felice Dell'Orletta
Anna Monreale

Pisa, May 2022

Abstract

IN the last few years, the analysis of the inner workings of state-of-the-art Neural Language Models (NLMs) has become one of the most addressed line of research in Natural Language Processing (NLP). Several techniques have been devised to obtain meaningful explanations and to understand how these models are able to capture semantic and linguistic knowledge. The goal of this thesis is to investigate whether exploiting NLP methods for studying human linguistic competence and, specifically, the process of written language evolution is it possible to understand the behaviour of state-of-the-art Neural Language Models (NLMs). First, we present an NLP-based stylometric approach for tracking the evolution of written language competence in L1 and L2 learners using a wide set of linguistically motivated features capturing stylistic aspects of a text. Then, relying on the same set of linguistic features, we propose different approaches aimed at investigating the linguistic knowledge implicitly learned by NLMs. Finally, we propose a study in order to investigate the robustness of one of the most prominent NLM, i.e. BERT, when dealing with different types of errors extracted from authentic texts written by L1 Italian learners.

Acknowledgements

THIS thesis would not have been possible without the support of many people. First, I would like to thank my supervisor, Felice Dell'Orletta, for guiding me along this journey and helping make sense of thousands of experiments and revisions. I also thanks my supervisor, Anna Monreale, and the internal committee members, Davide Bacciu and Alina Sirbu, for their precious feedback and suggestions.

I would also like to acknowledge my thesis reviewers, Afra Alishahi and Arianna Bisazza, for their careful reading of my work and for their comprehensive and extremely helpful comments.

This journey would not have been the same without the help and the support of all the members of the ItaliaNLP Lab (Institute for Computational Linguistics "A. Zampolli", ILC-CNR, Pisa). In particular, I would like to thank Giulia Venturi and Dominique Brunato, whose expertise helped me a lot to improve my work and widen my perspective and understanding of this complex and fascinating research field.

I am extremely thankful to my colleagues, Chiara and Lorenzo, for their support and to the dissemination group of AILC (Associazione Italiana di Linguistica Computazionale) for teaching me that Computational Linguistics is not just about research.

Finally, I would like to thank my parents for their support through all these years and Chiara, who always offered me love, encouragement and motivation.

List of publications

International Journals

1. Miaschi A., Sarti G., Brunato D., Dell'Orletta F., Venturi G. (2022). Probing Linguistic Knowledge in Italian Neural Language Models across Language Varieties. *Italian Journal of Computational Linguistics* (IJCoL), forthcoming.

2. Miaschi A., Brunato D., Dell'Orletta F. (2021). A NLP-based stylometric approach for tracking the evolution of L1 written language compentece. *Journal of Writing Research* (JoWR), Vol. 13, Issue 1, pages 71-105.

International Conferences/Workshops with Peer Review

1. Albertin G., Miaschi A., Brunato D. (2021). On the role of Textual Connectives in Sentence Comprehension: a new Dataset for Italian. In *Proceedings of the Eighth Italian Conference on Computational Linguistics* (CLiC-it 2021).

2. Miaschi A., Alzetta C., Brunato D., Dell'Orletta F., Venturi G. (2021). Probing Tasks Under Pressure. In *Proceedings of the Eighth Italian Conference on Computational Linguistics* (CLiC-it 2021).

3. Miaschi A., Brunato D., Dell'Orletta F., Venturi G. (2021). What Makes My Model Perplexed? A Linguistic Investigation on Neural Language Models Perplexity. In *Proceedings of the 2nd Workshop on DeeLIO* (NAACL 2021, Online).

4. Puccetti G., Miaschi A., Dell'Orletta F. (2021). How Do BERT Embeddings Organize Linguistic Knowledge? In *Proceedings of the 2nd Workshop on DeeLIO* (NAACL 2021, Online).

5. Miaschi A., Sarti G., Brunato D., Dell'Orletta F., Venturi G. (2020). Italian Transformers Under the Linguistic Lens. In *Proceedings of the Seventh Italian Conference on Computational Linguistics* (CLiC-it 2020).

6. Miaschi A., Alzetta C., Brunato D., Dell'Orletta F., Venturi G. (2020). Is Neural Language Model Perplexity Related to Readability? In *Proceedings of the Seventh Italian Conference on Computational Linguistics* (CLiC-it 2020).

7. Miaschi, A., Brunato, D., Dell'Orletta, F., and Venturi, G. (2020). Linguistic Profiling of a Neural Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

8. Miaschi, A., Davidson, S., Brunato, D., Dell'Orletta, F., Sagae, K., Sanchez-Gutierrez, C. H., and Venturi, G. (2020). Tracking the Evolution of Written Language Competence in L2 Spanish Learners. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

9. Miaschi A., Dell'Orletta F. (2020). Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP* (ACL 2020, Online).

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivations

The field of Natural Language Processing (NLP) has seen an unprecedented progress in the last years. Much of this progress is due to the replacement of traditional systems with newer and more powerful algorithms based on machine learning (ML) and, more recently, on deep learning (DL) techniques. In fact, state-of-the-art ML and DL models proved to outperform earlier systems and humans in a number of tasks, bringing the advantage of both better results and fast predictions. In the context of NLP, these techniques have been employed in several application scenarios, such as sentiment analysis [Sun et al., 2019], parsing [Wang et al., 2019] or machine translation [Bentivogli et al., 2016], and different domains [Li et al., 2019,Chalkidis et al., 2020]. This improvement, however, comes at the cost of interpretability, since complex neural models offer little transparency about their inner workings and their abilities.

To some extent, the problem of analyzing and interpreting neural networks (NNs) in NLP falls into the larger question of interpretability in machine learning, which has been the subject of much debate in recent years. Despite divergent opinions, there is a general agreement on the need for the implementation of principles for obtaining "meaningful explanations of the logic involved" when automated decision-making take place [Guidotti et al., 2018]. In fact, by relying on complex machine learning models trained on massive datasets, one of the main risk is to create and use decision systems that we do not really understand. This impacts not only ethics but also on accountability [Kroll et al., 2016], on safety [Danks and London, 2017] and on industrial

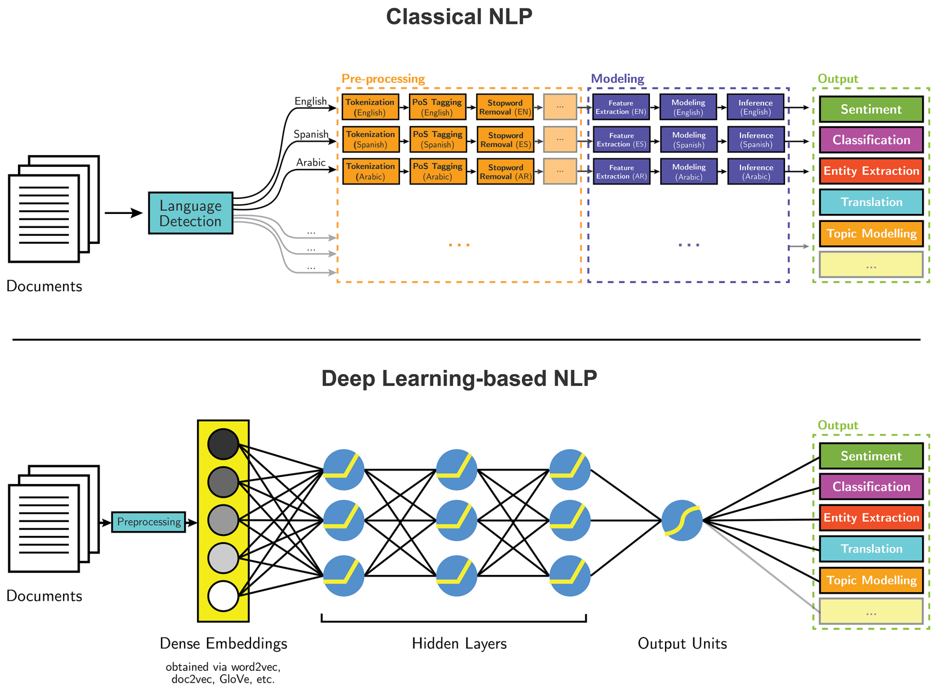


Figure 1.1: *Source* `https://s3.amazonaws.com/aylien-main/misc/blog/images/nlp-language-dependence-small.png`. *Classical vs. DL-based NLP pipeline.*

liability [Kingston, 2016]. It is therefore no coincidence that these issues have attracted the attention not only of the Artificial Intelligence (AI) research community but also of legislators[1]. All these arguments naturally apply to machine learning applications in NLP.

In the context of NLP this question needs to be understood in light of earlier work, where features automatically extracted from text, such as morphological properties, lexical classes, syntactic categories or semantic relations are used to make predictions with the so-called "feature-engineered systems". In this scenario, one could observe the importance assigned by statistical NLP models to such features in order to gain a better understanding of the model. In contrast, it is more difficult to understand what kind of linguistic knowledge is encoded in an end-to-end Neural Language Model (NLM) [Belinkov and Glass, 2019]. Consider a text classification task (see Figure 1.1). Differently from a classical NLP pipeline, a typical NN model would take as input word (or sentence) vectors from the input text and apply non-linear transformation over the vectors. This transformation process can be further repeated via recurrence or recursion of the network, before reaching the final prediction. As a consequence of this procedure, the model often lacks a good explanation of its computation and of the linguistic competence implicitly acquired during the training process. This problem has

[1] In 2018, the European Parliament adopted the General Data Protection Regulation (GDPR), which contains, among the others, clauses focused on automated decision-making. Link: `https://eur-lex.europa.eu/eli/reg/2016/679/oj`

become even more critical since the introduction of the latest state-of-the-art NLMs which, given their large size in terms of model parameters and the huge amount of textual data used for the training process, are getting more and more challenging to inspect. Moreover, it should be noted that being able to obtain insights about the inner workings of such models can provide important feedback for the future development of more efficient and meaningful models.

For all these reasons, in the last few years the analysis of NN language models is becoming more prevalent and several research questions are being raised. What happens in an end-to-end neural network model when trained on a language modeling task? What kind of linguistic knowledge is encoded within their representations? Is there a relationship between the linguistic knowledge implicitly encoded and the ability to solve a specific task? As a matter of fact, several survey papers summarising the main studies focused on the analysis of such models have been published in the last two years only (e.g. [Belinkov and Glass, 2019,Rogers et al., 2020]). This thesis falls within the context of these studies and propose different approaches for understanding the inner behaviour of NN models of language and, more specifically, recent state-of-the-art NLMs based on the Transformer architecture [Vaswani et al., 2017].

1.2 Objective and Contributions

The main objective of this thesis is to deepen the study of the inner workings of recent state-of-the-art NLMs. To this aim, we started with the following research question: can we use NLP methods developed to study human linguistic competence and, specifically, the process of written language evolution to interpret the linguistic knowledge encoded by NLMs? Starting from this hypothesis, we decided to exploit several approaches which focus on the study of the linguistic competence implicitly learned by these models rather than addressing their internal structure. In this sense, our methodology is in line with the studies that, as described by [Belinkov and Glass, 2019], seek to understand "how linguistic concepts that were common as features in NLP systems are captured in neural networks".

In the following we discuss the main contributions of the thesis.

1.2.1 An NLP approach for tracking the evolution of written language competence

We introduce an NLP-based stylometric approach for tracking the evolution of written language competence in L1 and L2 learners. It should be noted that We refer to the concept of "written language competence evolution" as the process of monitoring and understanding the properties of L1 and L2 learners language and how these properties evolve over time. The approach is based on the core assumptions of computational stylometry, i.e. formal properties of a text characterizing its style can reveal underlying traits about the author [Daelemans, 2013]. From this perspective, we argue that a ML model trained with a wide set of linguistically motivated features can provide important insights about the process of writing skills evolution. We experimentally assess our approach on two longitudinal corpora containing essays written by Italian L1 and Spanish L2 learners respectively. The results show that our set of linguistic features allows the model to automatically predict the relative order of two essays written by the same student at different course levels. More importantly, we show that our approach can be exploited to investigate the typologies of language phenomena (i.e. linguistic features) that contribute more to the prediction task and how they change according to different temporal spans.

1.2.2 Interpreting Neural Language Models Linguistic Abilities

We propose different methodologies aimed at investigating the inner workings of recent NLMs, deriving a vast amount of hints about the linguistic knowledge encoded by these models. More in detail, relying on the same set of linguistic features used in the experiments devised for tracking the evolution of written language competence, we design several experiments exploiting multiple interpretation techniques, ranging from the definition of probing tasks to the analysis of the relationship between linguistic competence and NLMs perplexity scores. The results show that Transformer-based models are able to capture a wide range of linguistic phenomena, even without being explicitly designed to learn such properties. Moreover, we show that the linguistic knowledge stored in the internal representations of one of the most popular NLM positively affects its ability to solve a classification task.

Focusing instead on the NLM sensitivity on authentic texts, we propose an extensive analysis on the behaviour of a NLM when dealing with the learner errors derived from the corpus of Italian L1 students. In particular, we provide a comprehensive investigation of how non-standard linguistic forms are encoded in a pre-trained model by inspecting its inner mechanisms from different perspectives with the aim of understanding, and to what extend, internal representations diverge when the model is exposed to incorrect and correct forms. Moreover, we study the relationship between the presence of certain typologies of linguistic errors in a sentence and the model ability to correctly encode within its internal representations a set of linguistic properties characterising that sentence.

1.3 Structure of the Thesis

The thesis is organized in 3 parts and 9 chapters, plus introduction and conclusions.

In part I we present related work. Specifically, in Chapter 2, we introduce the background topics relevant for developing the thesis. In particular, we provide an overview of the approaches developed in the last years for computing and learning representations from texts relying on ML techniques. In Chapter 3, we present the main studies and methodologies, based on data-driven and NLP approaches, to study the process of language development and second language acquisition, both in spoken and written language. In Chapter 4, we propose an overview of the works that aim to interpret and understand the inner workings of NLMs and, more specifically, the linguistic concepts implicitly learned by these models.

Part II focuses on the studies devised for tracking the evolution of written language competence. In Chapter 5, we present the approach and the experiments we designed in order to monitor the evolution of written language competence in L1 learners, while in Chapter 6 we discuss a complementary study in which we applied the same methodology to L2 learners of Spanish. In Chapter 7 we draw the conclusion and give suggestions to future research work direction.

In Part III we report the works we have conducted in order to understand the linguistic proprieties encoded by state-of-the-art NLMs. Chapter 8 focuses on the experiments we devised relying on the basis of the so-called *probing classifiers* paradigm. Chapter 9 discuss the analyses we did on the relationship between NLMs perplexity scores and grammatical generalization abilities and on the performance of these models on targeted diagnostic tests. Finally, in Chapter 10 we present an investigation aimed at understanding the robustness and the sensitivity of a NLM against non-standard forms emerging from the authentic texts we used for the experiments of Chapter 5.

In Chapter 11, we draw our conclusions, summarising the contributions and the results obtained, and presenting possible future research directions.

Part I

Background

CHAPTER 2

Representation Learning in NLP: from Feature Engineering to Neural Representation

In the last years, there has been an exponential growth in the number of machine learning (ML) algorithms in the context of NLP applications. The success of these learning algorithms relies on their capacity to understand a complex system such as language and non-linear relationships within data. Nevertheless, finding the most effective architectures and techniques for inferring textual representations from data still represents a challenge in the NLP community. In this chapter, we provide an overview of the state-of-the-art approaches developed for computing and learning textual representations from texts. Specifically, in 2.1 we focus on the conventional machine learning approaches that have been widely used until the early 2010s. In 2.2, instead, we take a closer look at the latest approaches based on deep learning techniques, which integrate feature engineering into the model fitting process by learning distributed representations as non-linear combinations of weights in a Neural Network.

2.1 Conventional Machine Learning Approaches

Until the early 2010s, ML methods developed for NLP applications were primarily based on the extraction of sample features by artificial methods [Li et al., 2020]. Therefore, the effectiveness of these methods, often defined as conventional machine learning systems, was largely due to the feature extraction process. More specifically, the typical workflow of a conventional ML system built for a text classification problem can be divided in four steps, as illustrated in Figure 2.1:

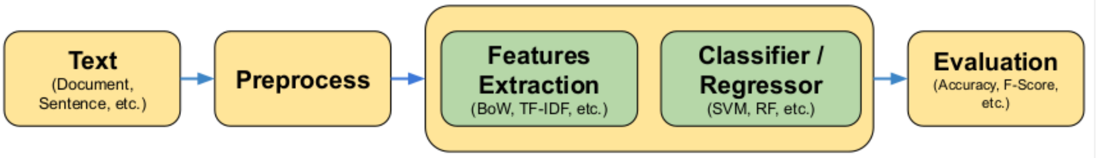


Figure 2.1: *Flowchart of a text classification task with a conventional ML system.*

- The raw textual input for training the model is preprocessed (e.g. data cleaning, tokenization, POS tagging, syntactic parsing, etc.);
- Preprocessed text is converted in text representations that can be used as input features for the ML model. These representations are typically computed relying on statistical features, such as Bag-of-Words (BOW) or n-grams, or on hand-crafted features that can be extracted with specific NLP systems, e.g. types of words and entities, semantic roles, parse trees, etc;
- The text represented according to the selected features is used as input of a classifier/regressor model;
- Finally, the classification/regression model is evaluated according to the appropriate metric (e.g. accuracy, F-score, etc).

Numerous models have been proposed, ranging from simple classification algorithms, such as logistic regression (LR) and Naïve Bayes Classifier (NBC), to tree-based classifiers [Xu et al., 2012] (e.g. Decision Tree and Random Forest) and Support Vector Machine (SVM) models [Manevitz and Yousef, 2001]. Nevertheless, as mentioned above, deciding the right features is a crucial aspect (if not the most important) of a successful ML project and this is especially true for language data, which comes in the form of a sequence of discrete symbols.

In the following subsections, we review the most common techniques to perform feature extraction on textual data. Specifically, in 2.1.1 we present some of the approaches that rely on statistical methods, while in 2.1.2 we examine the methodologies developed for inferring linguistic properties from lexical resources or automatically annotated texts.

2.1.1 Statistical Features

One of the most intuitive way to represent a word in order to be processed by a ML system (i.e. an input vector) is relying on a one-hot word representation. Formally speaking, given a vocabulary $V = w_1, w_2, ..., w_{|V|}$ we can represent a word *w* with a |V|-dimensional vector **w**, where each dimension of *w* is 0 or 1:

$$w_i = \begin{cases} 1 & if w = w_i \\ 0 & otherwise. \end{cases} \tag{2.1}$$

In other words, one-hot word representations maps each word to an index of *V*. One of the main drawbacks of this approach is that this type of representation cannot capture the relatedness among words: the difference between *house* and *home* is as much as the difference between *house* and *football*.

Moving from word to sentence (or paragraph, document) representations, a simple technique of representing such data as input features is to rely on the counts of the characters or the words within the text. A very common feature extraction approach based on this methodology is the bag-of-words approach (BOW). Roughly speaking, the BOW approach is a representation that converts arbitrary text into fixed-length vectors by simply counting how many times each word appears within the given text. When using the BOW, it is possible to improve the feature vector computation by using TF-IDF weighting [Luhn, 1957, Robertson and Jones, 1976]. Although very simple, BOW models perform very good in applications like spam filtering, text classification and information retrieval.

Besides computing word or sentence representations relying on single words, it is also possible to count consecutive word sequences of a given length. These representations are called *n*-grams. Differently from the previous approaches, *n*-grams models are more informative, since they can detect structures that go beyond individual words: e.g. *New York*, *not good* and, in case of *n*-grams consisting of more than two words, basic syntactic structures. Given their ability to exploit more complex structures, *n*-grams models are still widely used nowadays and, in some specific cases, are able to achieve results comparable to those of state-of-the-art neural models.

2.1.2 Hand-crafted Features

In certain NLP tasks, besides exploiting statistical information that can be immediately extracted from the words (or sentences) of a text, it may be useful to derive representations that can contain information regarding the syntactic or semantic structure of such texts. In fact, while the linguistic properties of a text are not directly observable from the surface of words and their order, they can be inferred from a sentence (or a document) with varying degrees of accuracy. Nowadays, there are several NLP tools developed for the prediction of Part-of-Speech (POS) tags, syntactic trees, semantic roles and other properties [Straka and Straková, 2017, Qi et al., 2020]. The predictions of these systems often serve as input features for further ML models.

Once a sentence (or a document) is automatically annotated with a linguistic annotation tool, the inferred output (i.e. the analyzed sentence, as in the example of Figure 2.2) can be used to extracted features related to, e.g.: POS tags, dependency labels, subtrees or paths that connect words within the tree, as well as properties of the paths, etc. It is important to notice, however, that although these systems are highly effective, it is acknowledged that their accuracy (especially for what concerns statistical parsers) decreases when tested against texts of a different typology from that used in training [Gildea, 2001]. For this reason, it is important to keep in mind that the introduction of such systems for the extraction of linguistic features may introduce errors in the subsequent phases of the task.

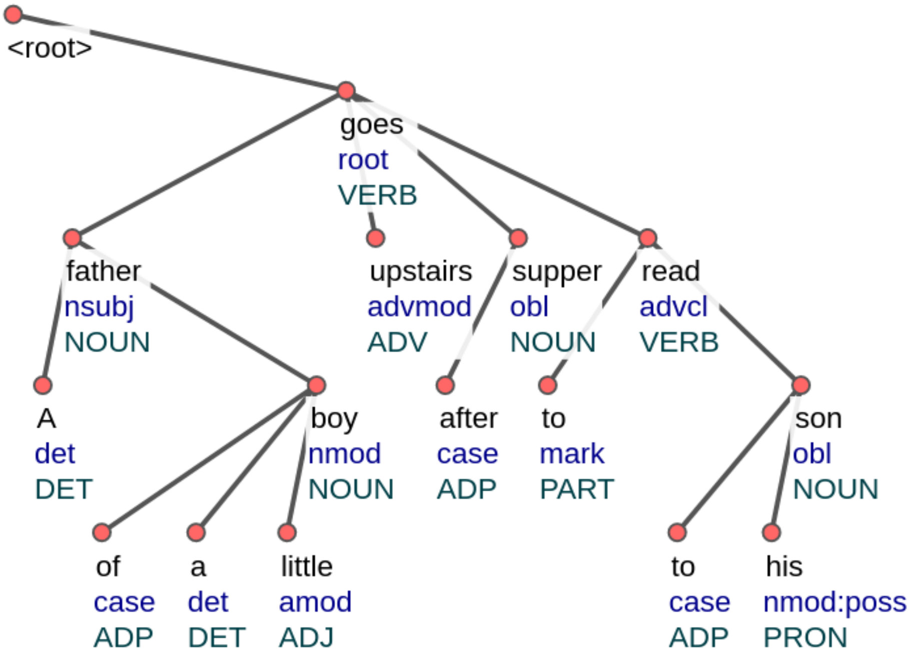


Figure 2.2: *Example of the linguistic annotation performed on the sentence "A father of a little boy goes upstairs after supper to read to his son" with the UDPipe tool [Straka and Straková, 2017].*

Linguistic Profiling

By relying on different levels of linguistic annotation, it is possible to extract a large number of features modeling lexical, grammatical and semantic phenomena that, all together, contribute to represent language variation within and across several texts. These are the prerequisite of the linguistic profiling approach, a methodology in which counts of a large number of linguistic features are used in order to detect and quantify differences and similarities across texts representative of distinct language varieties [van Halteren, 2004]. Following this approach, the linguistic structure of a text is analyzed to extract relevant features, and a representation of the text is constructed out of occurrence statistics of these features, either absolute/relative frequencies or more complex statistics. This approach is nowadays applied in different contexts and areas of research, which share the purpose of reconstructing the linguistic profile underlying linguistic productions originating in specific contexts, e.g. in socio–culturally defined demographic groups or individual author. In other terms, linguistic profiling allows the extraction of "meta-knowledge" from texts [Daelemans, 2013], i.e. what are the features and how they combine together within a specific language variety as opposed to another one of the same nature. Meta-knowledge extraction thus consists in associating the feature-based representation of texts with a functional context, or with a class of speakers and/or addressees, or with individual authors. In the last years, several studies have focused on developing profiling features capturing register, stylistic and linguistic complexity properties [Nguyen et al., 2016]. They range from studies that explored features based on morphosyntactic and syntactic structure, such as POS frequencies [Argamon et al., 2003, Otterbacher, 2010] or features based on context-free-grammar (CFG) rules [Bergsma et al., 2012], to more complex systems that allows the extraction of wide ranges of linguistic properties spanning across different levels of linguistic annotation [Brunato et al., 2020].

2.2 Neural Network Approaches

Since the 2010s, ML approaches for NLP applications has gradually changed from shallow and conventional learning models to deep learning (DL) models. Compared with the algorithms based on shallow learning, DL methods avoid designing rules and features by humans and automatically provide semantically meaningful representations for text mining. When moving from the approaches described previously to those based on DL techniques, the way in which we represent each linguistic object change from local or symbol-based representations (e.g. word scores in BOW models) to distributed representations. In distributed representation, each entity is represented by a pattern of activation distributed over multiple elements and each of them is involved in representing multiple entities [Liu et al., 2020]. The idea of distributed representation was originally inspired by the neural computation scheme of humans and, with the great success of deep learning and artificial neural networks (NNs), has become the most powerful and commonly used approach for inferring representation from textual data.

2.2.1 Neural Language Models

One of the first model developed with the purpose of learning distributed representation for words is the Neural Probabilistic Language Model (NPLM) [Bengio et al., 2003], which is based on a NN model trained to approximate the language modeling function. Language modeling (LM) is the task of predicting the joint probability of sequences of words. Formally speaking, a probabilistic language model defines the probability of a sentence $s = [w_1, w_2, ..., w_N]$ as:

$$P(s) = \prod_{i=1}^{N} P(w_i | w_1, w_2, ...w_{i-1}) \tag{2.2}$$

Traditionally, models based on *n*-grams were employed for predicting the next word in a *n*-gram sequence, following the Markov assumption that the probability of the target word only relies on the previous $n - 1$ words. Nevertheless, language models based on *n*-grams suffer from a number of limitations. First, although several smoothing techniques have been proposed to alleviate the problem of data sparsity, a *n-gram* LM still performs poorly on unseen and uncommon words. Moreover, since these models are trained on huge datasets, the number of unique words (and possible sequences) increases exponentially with the size of the vocabulary, causing again a data sparsity problem. To address this issue, [Bengio et al., 2003] proposed a Neural Language Model (NLM) that assigns a distributed vector for each word and then uses a NN architecture to predict the

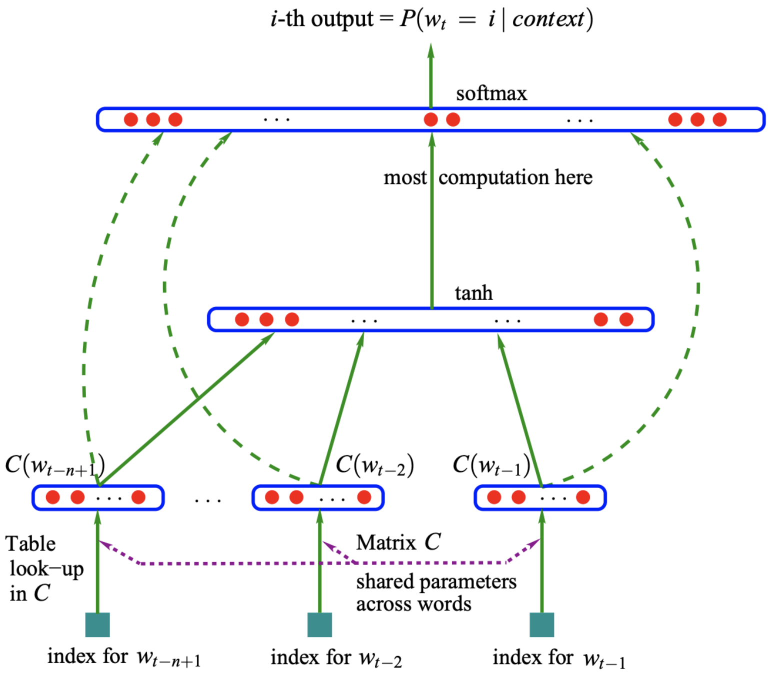


Figure 2.3: *Source [Bengio et al., 2003]. Neural architecture:* $f(i, w_{t-1}, ..., w_{t-n+1}) = g(i, C(w_{t-1}), ..., C(w_{t-n+1}))$ *where g is the neural network and* $C(i)$ *is the i-th word feature vector.*

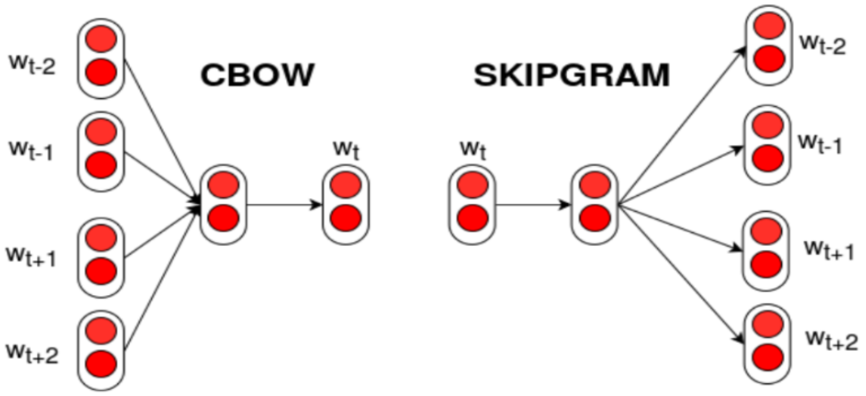


Figure 2.4: *Word2vec architectures according to CBOW (left) and Skip-gram (right) models.*

next word (See Figure 2.3). By training it trough a specific corpus, the NPLM learns how to model the joint probability of sentences and, at time the same time, returns word embeddings (i.e. low-dimensional word vectors) as learned parameters. In contrast with the previous approaches, word embeddings learned by a NLM reduce the dimensionality of categorical variables and meaningfully represent categories in the transformed space.

Influenced by NPLM, several methods that embed words into distributed representations learned by a NN have been devised: e.g. word2vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014]. Although different from each other, all of these models

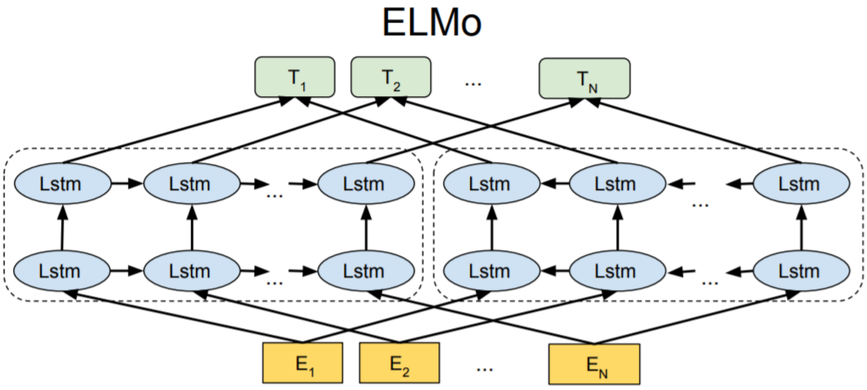


Figure 2.5: *Source [Devlin et al., 2019]. ELMo architecture.*

are particularly efficient and have been widely adopted in several NLP tasks over the last few years. For instance, the popular word2vec algorithm starts from the language modeling task and then it modifies it to produce faster results. More precisely, word2vec came as a software package[1] implementing two different context representation: CBOW and Skip-gram (Figure 2.4). CBOW predicts the center word w_i given a window of context (e.g. 5 words in the example of Figure 2.4):

$$P(w_i|w_{j(|j-i|\leq l,j\neq i)}) = Softmax\left(M\left(\sum_{|j-i|\leq l,j\neq i} w_f\right)\right) \tag{2.3}$$

where $P(w_i|w_{j(|j-i|\leq l,j\neq i)})$ is the probability of word w_i given its contexts, l is the size of training contexts, M is the weight matrix in $\mathbb{R}^{|V|\times m}$, V is the vocabulary and m is the dimension of the word vector. The CBOW model is then optimized by minimizing the sum of negative log probabilities:

$$L = -\sum_i logP\left(w_i|w_{j(|j-i|\leq l,j\neq i)}\right) \tag{2.4}$$

On the contrary, the Skip-gram model predicts the context given the center word w_i:

$$P(w_j|w_i) = Softmax(M_{w_i})(|j-i|\leq l, j\neq i) \tag{2.5}$$

where $P(w_j|w_i)$ is the probability of the context word w_j given w_i and M is the weight matrix. The loss function is then computed as:

$$L = -\sum_i \sum_{(|j-i|\leq l,j\neq i)} P(w_j|w_i) \tag{2.6}$$

Contextualized models

One of the main drawbacks of these approaches for learning word vectors is that they can only allow a single context-independent representation for each word. For instance:

- *She keeps all her money in a bank.*
- *On the more resisting bank, called a point bar, the river deposits some of its load of silt and rock as it flows by.*

In these two sentences, although the word *bank* is the same, their meanings are different. Since traditional word embeddings models (word2vec, GloVe, etc) learn a unique representation for each word, it is impossible for them to capture how the meanings of words change based on their surrounding contexts. To overcome this issue, several works proposed methods for enriching NLMs with subword information [Wieting et al., 2016, Bojanowski et al., 2017] or learning separate vectors for each word [Neelakantan et al., 2014]. More recently, other studies has also focused on developing models for learning context-dependent representations [Melamud et al., 2016, McCann et al., 2017, Peters et al., 2017].

In 2018, [Peters et al., 2018] proposed ELMo, a deep bidirectional LSTM model that can represent each word depending on the entire context in which it is used. Specifically, ELMo transforms words into low-dimensional vectors by feeding the word and its surrounding text into two-layer biLMs (see Figure 2.5). Instead of using a standard language model, ELMo utilizes a bidirectional LM to learn word representations. Formally, given sequence of *N* words $(w_1, w_2, ..., w_N)$, ELMo computes a forward LM (as in equation 2.2) and a backward LM. The backward LM is similar to the forward one, the only difference is that it reverses the input word sequence to $(w_N, w_{N-1}, ..., w_1)$ and predicts each word according to the future context:

$$P(s) = \prod_{i=1}^{N} P(w_i | w_{i+1}, w_{i+2}, ...w_N) \tag{2.7}$$

Evaluating it across a diverse set of six benchmark NLP tasks, the authors showed that ELMo obtained large performance improvement when compared with previous state-of-the-art systems. Moreover, through ablations and other controlled experiments, they also confirmed that the biLM layers efficiently encode different types of syntactic and semantic information about words in context.

2.2.2 The Transformer Model

The research on representation learning in NLP took a big leap when the Transformer model [Vaswani et al., 2017] came out. In particular, the Transformer model is based on a encoder-decoder architecture that, relying on attention mechanisms, eschews recurrence and relays entirely on an attention mechanism to draw global dependencies between input

[1] `https://code.google.com/archive/p/word2vec/`

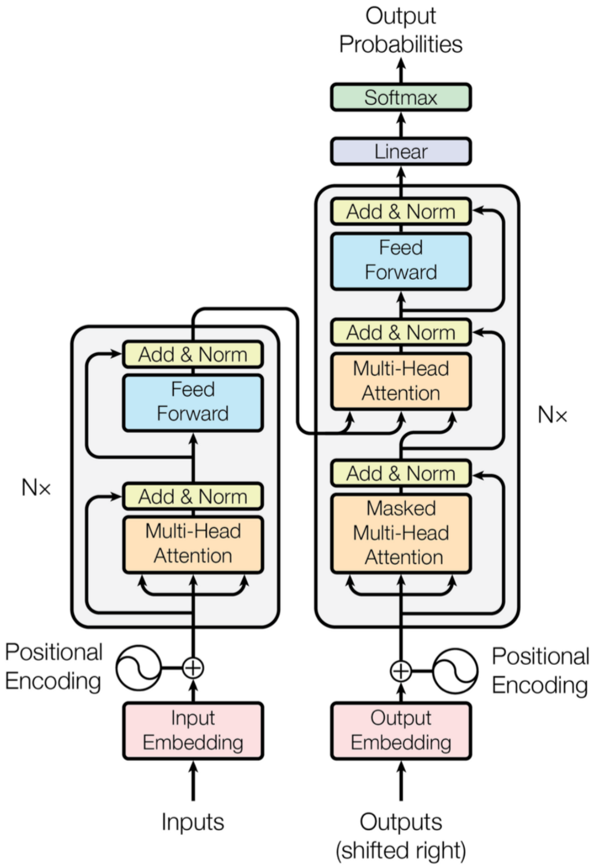


Figure 2.6: *Source [Vaswani et al., 2017]. The Transformer - model architecture.*

and output. This property let Transformers allow for significantly more parallelization at the cost of quadratic complexity in the input sequence length.

Encoder-decoder models (e.g. [Bahdanau et al., 2014, Cho et al., 2014]) encode an input sequence of symbol representations $x_1, ..., x_n$ to a sequence of continuous representations $z = (z_1, ..., z_n)$. Given z, the decoder then generates an output sequence $(y_1, ..., y_m)$ of one element at a time. At each step the model consumes the previously generated symbols as additional input when generating the next. The Transformer follows this overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and the decoder, as shown in Figure 2.6. The encoder is composed of a stack of 6 layers, each of which consists of two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The decoder is also composed of 6 identical layers but, in addition to the two sub-layers in each encoder layer, presents also a third sub-layer, which performs multi-head attention over the output of the encoder stack.

As mentioned above, each layer of the Transformer models is composed of a multi-head attention sub-layer. An attention function can be viewed as mapping a query and a set of key-value pairs to an output, where the query, keys, values and output are all

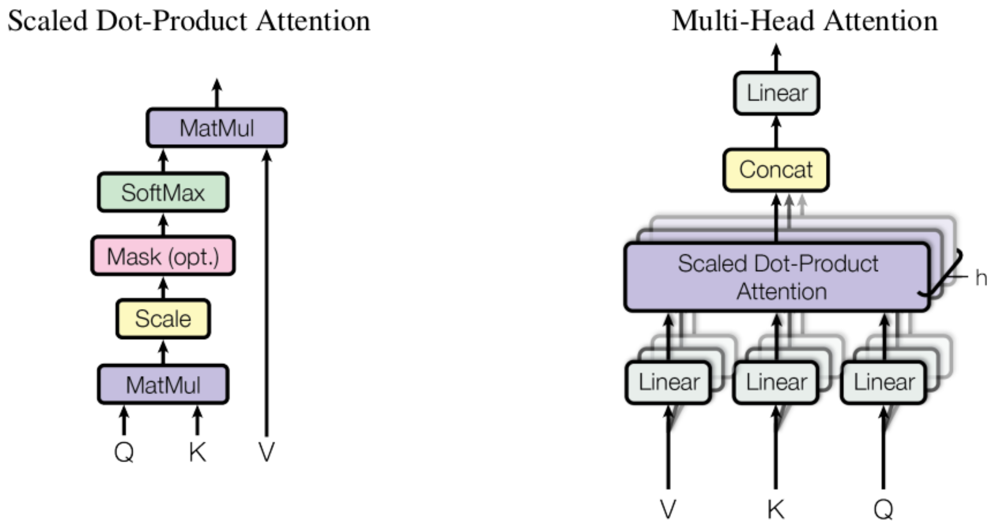


Figure 2.7: *Source [Vaswani et al., 2017]. (left) Scaled Dot-Product Attention. (right) Multi-head Attention consists of several attention layers running in parallel.*

vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. Formally speaking, given the query matrix Q, the key matrix K and the value matrix V as inputs, the output is computed as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.8}$$

where d_k is the dimension of the query matrix.

The multi-head attention sub-layer linearly projects the input hidden states H several times into the query, key and value matrices for h heads, as showed in Figure 2.7. In other words, the multi-head attention sub-layer can be formulated as:

$$Multihead(H) = [head_1, head_2, ..., head_h]W^O \tag{2.9}$$

where $head_1 = Attention(HW_i^Q, HW_i^K, HW_i^V)$ and W_i^Q, W_i^K and W_i^V are linear projections. W^O is also a linear projection for the output.

Although originally proposed to solve the problem of machine translation, given its ability of better modeling long-term dependencies, the Transformer model was successfully exploited in several works to build highly performative language models [Devlin et al., 2019, Radford et al., 2019, Sun et al., 2020b][2]. Moreover, the introduction of the Transformer architecture in the development of more powerful NLMs started a new approach in the NLP pipeline. Previously, in fact, word embeddings were simply adopted as input representations of another classification/regression model. Nowadays, it became a common practice to keep using the same NN architecture in both pre-training

[2] [Lin et al., 2021] for a comprehensive overview of the most popular Transformer models developed in the last few years.

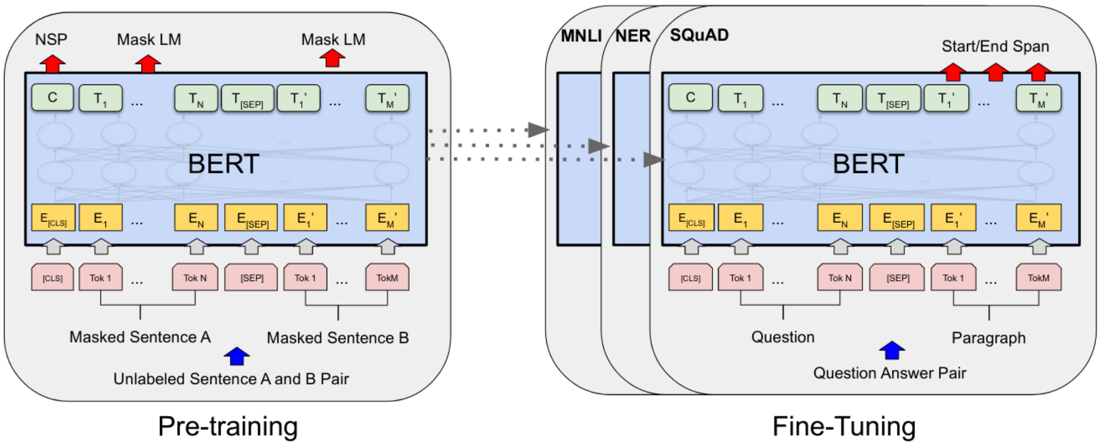


Figure 2.8: *Source [Devlin et al., 2019]. Overall pre-training and fine-tuning procedures for BERT.*

(e.g. language modeling) and fine-tuning (e.g. sentiment analysis). This approach is called *transfer learning* and is based on the key concept of transferring as much knowledge as we can from the source setting (i.e. the pre-training phase) to the target task or domain.

The BERT model

One of the most popular NLM based on the Transformer architecture is BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin et al., 2019]. Differently from the original Transformer model (or e.g. the GPT one [Radford et al., 2019]), BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context. However, since standard conditional language models can only be trained left-to-right or right-to-left (bidirectional conditioning would allow each word to indirectly "see itself"), the authors proposed a a modified version of the original LM task: Masked Language Modeling (MLM). The MLM task consists in masking some percentage of the input tokens at random, and then asking the model to predict those tokens. BERT is also pre-trained with a next sentence prediction task (NSP), i.e. the task of predicting if two sentences are consequent or not. For what concerns fine-tuning, the model is first initialized with the pre-trained parameters, and then all the parameters are fine-tuned using labeled data from the downstream tasks (e.g. NER, MNLI, CoLA, etc.). An overview of the pre-training and fine-tuning procedures are shown in Figure 2.8.

BERT became rapidly a milestone work in the field on NLP, achieving significant empirical results on several tasks, including SQUAD [Rajpurkar et al., 2016], GLUE [Wang et al., 2018], etc. Given its success, several variants of the original model for better language representations have been proposed, such as RoBERTa [Liu et al., 2019b] and XLNet [Yang et al., 2019]. Moreover, since the vast majority of works focused on how BERT and BERT-based models works on different NLP tasks and less on its workings, in the last two years there has been a growing interest in the field of study concerned with investigating the inner behaviour of these models. We will give an overview of these studies in Chapter 4.

CHAPTER 3

NLP methods for Language Development

In this chapter we take a closer look at studies in the literature which have relied on data–driven approaches, complemented with NLP–based analyses at different degrees of sophistication, to study the process of language development and second language acquisition, both in spoken and written language.

3.1 Investigating Language Development in L1 Learners

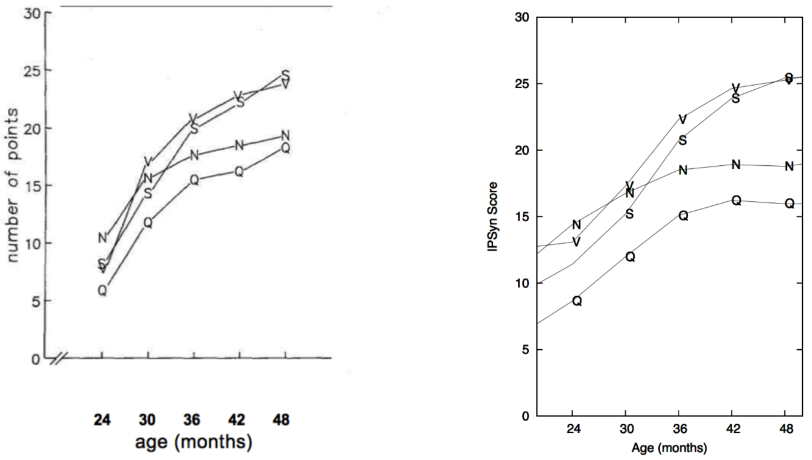
Over the last years, there has been a growing interest to exploit the potential of Natural Language Processing (NLP) tools and machine learning methods in the context of language development, with the aim of characterizing the properties of L1 learner language and how it evolves over time, across modalities and stages of acquisition. A similar concern has been paid to turn theoretical considerations into educational applications, such as *Intelligent Computer-Assisted Language Learning* (*ICALL*) systems [Granger, 2003] and tools for automatically scoring learners' writing with respect to language proficiency and writing quality. Two main ingredients stand at the core of this research: the availability of large digitalized corpora of authentic texts produced by learners, which make it possible to complement theoretical underpinnings with corpus-based evidence, and the reliability of language analyses generated by computational tools that allow quantifying and evaluating the impact of a large number of linguistically–motivated indices considered in the literature as proxies of language development.

3.1.1 Child Language Acquisition

In the context of child language acquisition, a first line of research has focused on modeling the development of syntactic abilities in preschool children using data from the CHILDES database [MacWhinney, 2000] and a variety of features derived from a semi– or fully–automatic process of linguistic annotation. The CHILDES corpus contains transcripts of spoken interactions in natural settings involving children of different ages for over 25 languages, which makes it a reference corpus for empirical research on language acquisition. Based on a subset of utterances from English speaking children (age 1-6), which were automatically annotated for syntactic dependency relations, [Sagae et al., 2005] demonstrated that the hand-crafted calculation of the Index of Productive Syntax (IPSyn)[1] [Scarborough, 1990] can be effectively automated using features extracted from the sentence parse tree, in addition to information related to Part-of-Speech (POS) tagging. The index was introduced to overcome the limit of a well-known metric of grammatical complexity, the Mean Length of Utterance (MLU), which is easy to calculate yet largely criticized for its insensitivity to capture relevant changes in child language. In this sense, the possibility of automatizing a metric like the IPSyn, which is undoubtedly much more powerful than MLU but requires a huge effort in terms of manual computation, represented a first important result towards the development of computational tools supporting research on language acquisition on a large–scale basis. In a similar vein, [Lu, 2009] proposed a heuristic–based approach to automatically assign a score of syntactic complexity to children's utterances according to a revised version of the D-Level Scale [Covington et al., 2006], a seven-step developmental level scale based on empirical observations about the emergence of increasingly more complex constructions from the child speech literature. In this case too, a corpus of utterances from CHILDES was automatically analyzed with a state–of–the–art English parser to allow the extraction of the grammatical structures contained in the reference scale.

The main lesson from these studies was that NLP techniques can be used in a reliable way to help automate the laborious computation of expressive metrics for child language development. However, a more challenging step was tackled by [Lubetich and Sagae, 2014], which proposed a completely data-driven approach to measure syntactic development without the need of previously designing the sophisticated inventory of grammatical structures associated to a given metric. In this study, a corpus of transcripts of children from 1 to 8 years was syntactically annotated and automatically assigned with its IPSyn score. Then, for each transcript, the IPSyn score was associated with a set of language-independent features extracted from text (e.g. unigrams of parts-of-speech, unigrams of syntactic dependency labels) and deliberately meant to capture information about the syntactic structure of children's sentences. The hypothesis was that if the IPSyn scores could be predicted from these generic vectors, the selected features would be at least enough informative for tracking child language development as the inventory of IPSyn structures. Experiments were performed using a Support Vector Machine (SVM)

[1] IPSyn is a sophisticated metric of child language acquisition, which scores children's utterances according to the distribution of more than 50 syntactic constructions (e.g. relative clauses, wh-questions with auxiliary inversion, propositional complements).



(a) Original IPSyn study.

(b) Automatically generated.

Figure 3.1: *Source: [Lubetich and Sagae, 2014]. Comparison between the IPSyn development curves of (a) the original IPSyn study (reproduced from [Scarborough, 1990]), and (b) the ones automatically generated.*

regression model. The results showed a high correlation between predicted and real IPSyn scores (as shown in Figure 3.1), supporting the hypothesis that simple parse tree features are as indicative of language development as a sophisticated language-dependent metrics. The authors also tested the data–driven approach on an age prediction task in which the regression model was trained to predict the age at which an unseen child transcript was produced, using the feature vector extracted from his/her other transcripts available in training. The underlying idea is that child language development could be better approached with age, rather than with a metric score, on the assumption that language acquisition (at least in typical setting) evolves monotonically over time. The results showed strong correlations (r > 0.90) between actual and predicted age for the tested children.

A similar approach was discussed in [Sahakian and Snyder, 2012], where a set of linguistic features was computed on child speech samples and used as input for a linear regression classifier in two age prediction experiments. In the first experiment, a child-specific metric was used to predict the age at which speech samples were produced. In the second one, a more language-independent developmental index was created for predicting relative temporal orderings of speech samples. In spite of the different implementations of the age prediction task, both these studies share the idea that child language development could be better approached with age, rather than with a metric score, on the assumption that language acquisition (at least in typical setting) evolves monotonically over time.

More recently, a few studies have also started to employ more sophisticated learning algorithms (i.e. neural networks) to investigate the child language acquisition process. [Sagae, 2021], for instance, showed that a fully data-driven model of language development that uses a recurrent NN encoder for utterances can track how child language utterances change over the course of language development in a way that is comparable to what is achieved using established language assessment metrics designed by experts.

3.1.2 Later Language Acquisition

The rapid and remarkable changes child language undergoes before age five justify the amount of research for the earliest stages of acquisition, which is the framework underlying all the aforementioned studies. However, under the assumption that linguistic competence keeps growing during the school years as a result of explicit literacy instruction [Karmiloff-Smith, 1986], research on "later language acquisition" has gained increased attention prompted by the awareness that "becoming a native speaker is a rapid and highly efficient process but becoming a proficient speaker takes a long time" [Berman, 2004]. Also in this scenario corpus-based approaches complemented with linguistically-informed indices (semi)-automatically extracted from text have started being applied to track the development of writing skills throughout the school years. Note that if, in the case of spoken language, the growth is tracked as a function of age, the development of writing skills is typically addressed as a function of increasing grade level, both in elementary and middle school children and in high school and college level students [Crossley et al., 2011a]. Inspired by the Multi-Dimensional Analysis (MDA) pioneered by Douglas Biber, which assumes that "linguistic features from all levels function together as underlying dimensions of variation" [Biber, 1993], [Chipere et al., 2001] applied this framework to the field of first language development during the school years. This study examined a large corpus of 899 graded essays written by school children (aged 8 to 15) with the aim of assessing the relationship between vocabulary diversity and age and level of linguistic ability. The latter was operationalized in terms of a normalized version of *type–token ratio* (TTR), to account for the effect of text length. Results showed that vocabulary diversity is in fact correlated with age and ability level, although with few exceptions involving the transitions between middle and high school grades (i.e. 11 and 14 years). With this respect, the authors recognized that vocabulary diversity is only one of the factors qualifying writing ability and that an index like TTR could attribute lower scores to essays in which pupils intentionally use repeated words not because they don't have enough lexical knowledge but to produce a more coherent discourse.

Recent developments in computational linguistics methods and machine learning techniques have granted researchers the opportunity to assess large corpora of graded essays to examine overall writing ability and its development. With the aid of the automatic tool *Coh-Metrix*[2], [Crossley et al., 2011a] enlarged the analysis to several linguistic domains and examined to what extent essays written at various grade levels can be distinguished from one another using a number of linguistic features related to lexical sophistication (i.e., word frequency, word concreteness), syntactic complexity (i.e., the

[2] *Coh-Metrix* is a computational system for computing cohesion and coherence metrics in written and spoken texts (http://cohmetrix.com).

number of modifiers per noun phrase), and cohesion (i.e., word overlap, incidence of connectives). The main findings show that high school and college writers develop different linguistic strategies as a function of grade level and that even in advanced writers, lexical and syntactic constructions continue to develop. In contrast, as the grade increases, writers tend to produce fewer cohesive devices, which is interpreted as a tendency towards a more elaborate and complex discourse composition. Similar conclusions are reported by [McNamara et al., 2010], which relied on the same tool to examine the degree to which high- and low-proficiency essays rated by experts can be predicted by linguistic indices of cohesion, syntactic complexity, the diversity of words used by the writer, and characteristics of words. The study showed that the three most predictive indices of essay quality were syntactic complexity, lexical diversity and word frequency but, interestingly, no indices of cohesion correlated with essay ratings. In a subsequent study, [McNamara et al., 2015] evaluated the use of a hierarchical classification approach to automated assessment of essays relying on features computed using three different automated tools: Coh-Metrix; the Writing Assessment Tool (WAT), i.e. a tool that includes a set of variables designed to assess the quality of written documents; Linguistic Inquiry and Word Count (LIWC), i.e. an automated word analysis tool that reports the percentage of words in a text that are in particular psychological categories [Pennebaker et al., 2007]. The use of automated tools in order to facilitate and augment formative writing assessment was also discussed in [Wilson et al., 2017]. In this study, the Coh-Metrix measures were used to demonstrate that automated tools are able to discriminate intra-individual differences in writing skills across levels of languages in ways that are meaningfully related to external measures of writing ability.

3.1.3 Assessing Writing Development in Non-English Languages

As expected, large part of empirical studies based on NLP approaches and machine learning techniques has been carried out with respect to the English language and focused on high school and college learners. However, more recently other L1s and age samples have been addressed. In this respect, [Weiss and Meurers, 2019] investigated writing development in German speaking students across elementary and secondary school. In particular, relying on the *Karlshruhe Children's Text* (KCT) corpus [Lavalley et al., 2015], a cross-sectional collection of 1,701 German texts produced by German elementary and secondary school students from first to eighth grade, the authors built classification models for early academic language development. Using a broad set of linguistically informed measures modeling text complexity and accuracy, together with error rate (see Table 3.1 for an overview) and background information on topic essay and school tracks, their best performing model was able to reach an accuracy of 72.68% in predicting the correct grades of students according to a fourth-level classification; notably, the model using only linguistically informed features, without any meta-data information, performs almost at the same level. A fine-grained analysis of the contribution of the individual features also revealed that writing acquisition in initial grades is best characterized in terms of accuracy development, while the upper

Feature Set	Size	Description
Lexical complexity	31	measures vocabulary range (lexical density and variation) and sophistication, measures of lexical relatedness; e.g., type token ratio
Discourse complexity	64	measures the use of cohesive devices such as connectives; e.g., connectives per sentence
Phrasal complexity	47	measures of phrase modification; e.g., NP modifiers per NP
Clausal complexity	27	measures of subordination or clause constituents; e.g., subordinate clauses per sentence
Morphological complexity	41	measures inflection, derivation, and composition; e.g., average compound depth per compound noun
Language Use	33	measures word frequencies based on frequency data bases; e.g., mean word frequency in Subtlex-DE [Brysbaert et al., 2011]
Human Language Processing	24	measures of cognitive load during human sentence processing, mostly based on Dependency Locality Theory [Gibson et al., 2000] e.g., average total integration cost at the finite verb
Error Rate	41	measures ratios of error types per error or word; e.g., spelling mistakes per word

Table 3.1: *Source [Weiss and Meurers, 2019]. Overview of the feature sets used in the classification experiments.*

stages of secondary school exhibit an increased linguistic complexity, in particular in the domains of lexis and syntactic complexity at the phrasal level.

Similar findings have been investigated by an analogous study by [Kerz et al., 2020] carried out on the same corpus, which still focused on the predictive role of language complexity features to tracking writing development but obtained through a sliding window technique, in order to monitor the progression of complexity within a text. In particular, after extracting a series of 'complexity contours' relying on the sliding window technique, the contours were fed into a RNN classifier to perform grade-level classification tasks. Moreover, by performing a feature ablation analysis based on an adpated version of the iterative sensitivity-based pruning algorithm proposed by [Díaz-Villanueva et al., 2010][3], the authors showed that the most important features pertain the length of production unit, lexical diversity, syntactic complexity and information density.

3.2 Investigating Language Development in L2 Learners

Several studies of L2 writing have focused on linguistic complexity as an indicator of writing development [Lu, 2011, Ortega, 2003]. However, Instead of considering the construct as being multidimensional [Norris and Ortega, 2009, Bulté and Housen, 2012] and, thus, encompassing an array of different features, most studies have selected one or two of these measures and used them as single indicators of complexity [Bulté and Housen, 2014]. This has prevented the development of much needed research that

[3] The algorithm ranks the features based on a 'sensitivity measure' and removes the least relevant variables one at a time.

associates different steps of linguistic and written development with specific sets of characteristics. This situation has also prevented the formation of an in-depth picture of how those specific aspects develop in relation to the grammatical, lexical or stylistic content taught in classes at different language course levels. Therefore, over the last few years, research on second language acquisition has benefited from the use of Natural Language Processing (NLP) technologies applied to large–scale corpora of authentic texts produced by learners. In fact, the empirical evidence acquired from L2 learner corpora, complemented with the increased reliability of linguistic features extracted by computational tools or machine learning approaches, has promoted a better understanding of learners' language properties and how they change across time and increasing proficiency level [Crossley, 2020]. A first line of research, as we already seen for what regards studies on L1 writing development, has focused on providing automatic ways of operationalizing sophisticated metrics of language development or to automatically extract linguistic features to alleviate the laborious manual computation of these metrics by experts [Lu, 2011]. A second line of research, instead, has taken the more challenging step of implementing completely data-driven approaches, which use wide set of linguistic features extracted from texts to automatically assign a learner's language production to a given developmental level and to understand which phenomena change across proficiency levels [Crossley and McNamara, 2012].

3.2.1 Linguistic Features to assess L2 Writing Development

Given the difficulty of defining a unique indicator of linguistic complexity in the context of L2 language development, a great variety of features from all linguistic levels have been used as proxies for investigating which are the properties that highly influence L2 proficiency. Numerous studies revealed that properties related to syntactic complexity, such as measures and subordination ratios, as well as more specific features pertaining to the usage of particular structures, can be considered as one of the key skills that affect L2 writing competence. For instance, [Lu, 2011] analyzed college-level English as a second language (ESL) writers' language development relying on 14 features of syntactic complexity automatically extracted from the Written English Corpus of Chinese Learners [Wen et al., 2005]. The results showed that the features that best discriminate between proficiency levels are those based on the number of complex nominals per sentence and the mean sentence length.

Features related to lexical complexity has been also investigated in the assessment of L2 written proficiency. [Kyle and Crossley, 2015], for example, introduced the Tool for Automatic Analysis of LExical Sophistication (TAALES), which computes 135 lexical indices related to word frequency, range, bigram and trigram frequency, academic language and psycholinguistic word information. In particular, TAALES was used to examine its ability to explain the variance in human judgments of lexical proficiency and speaking proficiency for L2 learners. The experiments, performed on a corpus of 180 unstructured writing samples from 10 L2 English learners at an intensive program over a 1-year period [Crossley et al., 2011b], showed that the indices computed

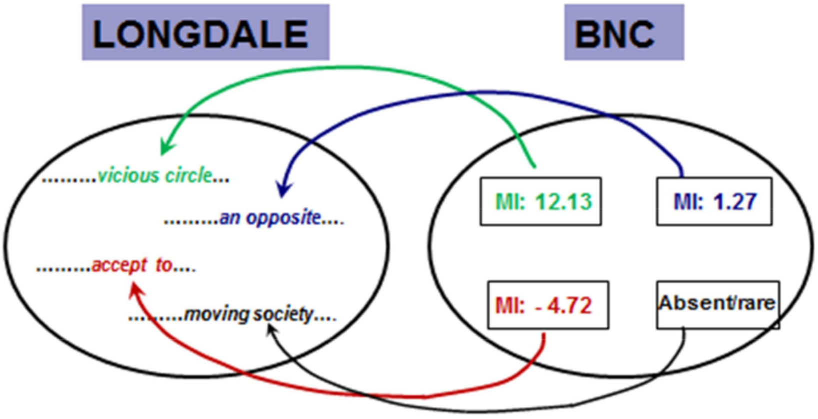


Figure 3.2: *Source: [Bestgen and Granger, 2018]. Assignment of BNC scores to LONGDALE data.*

with TAALES are able to explain 47.5% of the variance in holistic scores of lexical proficiency. Combining different metrics of syntactic and lexical complexity, accuracy and fluency, [Yoon, 2017] investigated the development of writing competence over time of ESL students. Specifically, studying narrative and argumentative essays written over the course of a 4-month semester by 37 students, the authors showed that there were limited changes over time on most features and especially for what concerns those related to accuracy.

Other metrics have been also proposed to investigate the development of L2 learners writing competence. For instance, [Bestgen and Granger, 2018] assessed learners' phraseological development proposing a methodology based on collgrams, i.e. word bigrams that have been assigned two scores (mutual information and t-score) on the basis of a large reference corpus. For their study, the authors used a subcorpus of the *Longitudinal Database of Learner English* (LONGDALE)[4] consisting of 178 essays written by 89 French-speaking English language and literature undergraduates at the University of Louvain. Collgrams scores were computed by extracting all the bigrams in the LONGDALE essays and then assigning their corresponding mutual information (MI) and t-score in the *British National Corpus*[5], as showed in Figure 3.2. Results showed that there is a general tendency for third year texts to contain fewer non-collocational bigrams and fewer high-scoring t-score collgrams, but more high-scoring MI collgrams.

3.2.2 Machine Learning Approaches

Studies that have adopted supervised ML approaches are mainly devoted to predict the degree of second language proficiency according to expert–based evaluation [Crossley and McNamara, 2012] or to model the evolution of grammatical structures' competence with respect to predefined grades, such as the Common European Framework of Reference for Languages (CEFRL) [Hancke and Meurers, 2013, Vajjala and Lõo,

[4] https://www.uclouvain.be/en-cecl-longdale.html

[5] http://www.natcorp.ox.ac.uk/corpus/

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

Table 3.2: *Source [Ballier et al., 2019]. EFCAMDAT error tagset.*

2014, Volodina et al., 2016]. These studies used numerous linguistic features in combination with a host of different classifier or regression models. [Hancke and Meurers, 2013] proposed one of the first study employing ML techniques along with essays rated with CEFR levels. In particular, they investigated which linguistic properties reliably support the classification of 1,027 professionally rated essays from the MERLIN corpus [Boyd et al., 2014][6] and comprising CEFR exams taken by German second language learners. Relying on a broad set of 3,821 features automatically extracted with NLP tools, they trained a classifier based on the Sequential Minimal Optimization (SMO) algorithm to predict five CEFR levels. When using cross-validation on all data, their system achieved 64.5% of accuracy when trained with all linguistic features. Moreover, examining the performance of individual feature groups, the authors showed that lexical and morphological features were the most important predictors of the CEFR levels. The same algorithm was also tested by [Volodina et al., 2016], where a set of 61 count-based, lexical, syntactic, morphological and semantic features were used for the classification of CEFR levels of L2 Swedish learner essays.

Another method besides classification is to define the proficiency level prediction as a regression task. [Vajjala and Lėo, 2014] trained a Linear Regression model in WEKA [Witten et al., 1999] to predict the CEFR levels assigned to the essays of the Estonian Interlanguage Corpus (EIC)[7], a corpus of texts written by learners of Estonian as a second or foreign language. The feature set used for the experiments consisted of 78 features, ranging from surface features to more complex properties related to morphological complexity and lexical variation.

In recent years, other ML models and set of linguistic features have been introduced to automatically assign proficiency levels to L2 learners. For instance, [Ballier et al., 2019] investigated the predictive power of errors in terms of levels and to identify which error types appear to be relevant features in determining proficiency levels. In particular, they

[6] https://merlin-platform.eu/index.php

[7] http://evkk.tlu.ee/?language=en

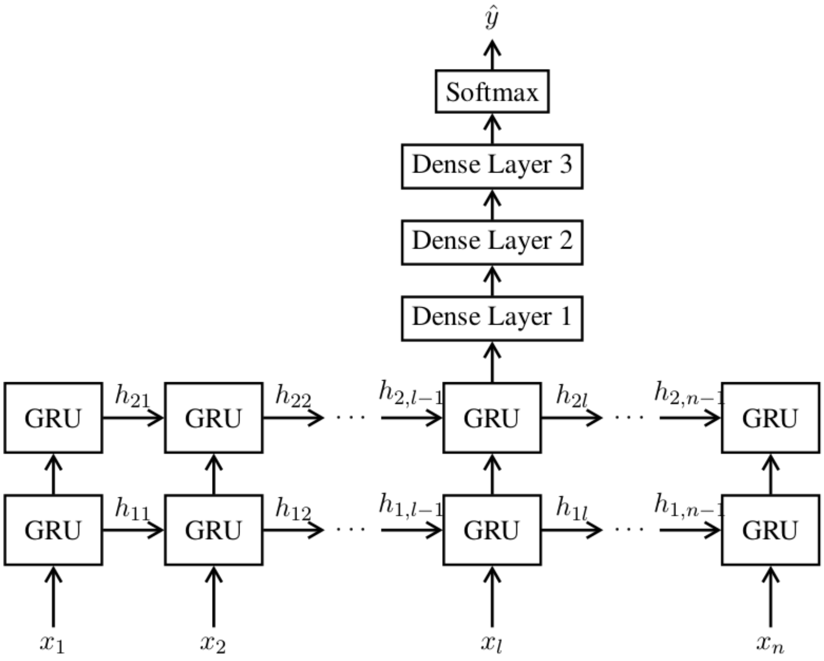


Figure 3.3: *Source: [Kerz et al., 2021]. Contour-based RNN model based on complexity contours.*

analyzed the possibility of automatically predicting CEFR levels on the basis of manually annotated errors in a subset of the EFCAMDAT corpus, an 83 million word learner corpus collected by the Cambridge University [Geertzen et al., 2013]. The errors were manually annotated on the basis of the Cambridge tagset, consisting of 24 types, as detailed in Figure 3.2. To find the optimal classifier, the authors compared the scores obtained with multinomial logistic regression, random forests, linear discriminant analysis, k-nearest neighbours, Gaussian naive Bayes, Support Vector Machine and decision tree classifier. Best results (F-Score = 0.70) were obtained using the random forest model. In addition, a second analysis based on logistic regression to investigate the relative importance of the 24 error types across learner levels revealed that mechanic errors (i.e. punctuation, inappropriate or missing spaces, capitalization issues and spelling) are significant across all CEFR levels, as well as the syntax error related to verb tense. Relying on the same corpus, [Kerz et al., 2021] proposed a methodology based on the extraction of 'complexity contours', a series of measurements of L2 proficiency obtained by a tool that implements a sliding window technique, and the classification trough recurrent neural network (RNN). Specifically, the 57 features used in their work were assessed using CoCoGen [Ströbel, 2014], with the aim of enabling a local assessment rather than a global assessment of complexity of a text. As regards the classification models, they used a RNN classifier with Gated Recurrent Unit (GRU) cells. As shown in Figure 3.3, the input of the model is a sequence $X = (x_1, x_2, ..., x_l, x_{l+1}, ..., x_n)$ where x_i is the output vector of CoCoGen for *i*th window of a document. The results showed that the inclusion of complexity contours led to an increase in overall classification accuracy of more that 9% (from 66.1% to 75.4%). Furthermore, a feature ablation experiment based on the technique already used in [Kerz et al., 2020] showed that the top 20 features that contributed most to the classification accuracy are those related to frequency n-grams pertaining to the usage of multiword sequences.

The vast majority of studies devoted to the prediction of L2 proficiency levels described approaches that work with a single language (mostly English). However, there are some exceptions. [Vajjala and Rama, 2018] performed experiments with cross-lingual and multilingual classifiers in order to verify the possibility of a 'Universal CEFR classifier'. To test their hypothesis, the authors experimented with graded texts of L2 German, Italian and Czech learners available in the MERLIN corpus. Relying on a set of domain-agnostic features (i.e. POS and dependency n-grams), as well as word and character embeddings and testing different classification models (e.g. logistic regression, random forest, MLP, SVM), they showed that average results on multilingual models are close to the ones obtained with the monolingual ones. In the context of the REPROLANG 2020 shared task on 'Language Proficiency Scoring', [Caines and Buttery, 2020] reproduced and extended [Vajjala and Rama, 2018] experiments. The results obtained in their experiments suggested that feature combination is the most robust approach to the L2 automatic proficiency scoring task, while neural network classifiers tend to achieve lower performance for text datasets of the same size.

CHAPTER 4

Interpreting Neural Networks for Language Understanding

Neural networks have rapidly become a central component in NLP systems in the last few years. Models based on neural networks have obtained improvements in accuracy and performance in various tasks, such as language modeling [Mikolov et al., 2010, Jozefowicz et al., 2016], syntactic parsing [Kiperwasser and Goldberg, 2016] or machine translation [Bahdanau et al., 2014, Sutskever et al., 2014]. As we already discussed in Chapter 2, this progress has been accompanied by new neural network architectures that rapidly replaced the traditional feature-based systems by end-to-end models that aim to map input text to some output prediction. However, the introduction of such systems has come at the cost of understanding how this NLP models work. For this reason, in the last few years, the analysis of the inner workings and, especially, of the linguistic competence learned by state-of-the-art Neural Language Models (NLMs) has become one of the most addressed line of research in NLP. Several methods have been implemented to obtain meaningful explanations and to understand how these models are able to capture syntax- and semantic-sensitive phenomena [Linzen et al., 2016, Belinkov and Glass, 2019]. These approaches range from the definition of fill-in-the-gap probes [Goldberg, 2019] and *probing tasks* that a model can only solve if it has encoded a precise linguistic phenomenon [Conneau et al., 2018, Zhang and Bowman, 2018, Hewitt and Liang, 2019], to the analysis of attention mechanism [Raganato and Tiedemann, 2018, Htut et al., 2019, Kovaleva et al., 2019] and correlations between representations [Saphra and Lopez, 2019] or between perplexity scores and grammatical generalization abilities [Hu et al., 2020].

In this chapter, we describe the studies which focused on the definition of techniques for interpreting the inner workings on NLMs and, more specifically, on the linguistic competence implicitly learned by those models.

4.1 Probing for Linguistic Competence

Several studies have focused at the knowledge encoded in NLMs by probing their ability of capturing linguistic phenomena. One of the technique employed is based on the so-called fill-in-the-gap probes, where a NLM is trained to predict the identity of masked words based on both the prefix and suffix surrounding these words. For example, given the pair of sentences s_1 = *"the game that the guard hates is bad"* and s_2 = *"the game that the guard hates are bad"*, the task consist on feeding the model with:

```
the game that the guard hates [MASK] bad
```

and compare the probabilities assigned to *is* and *are*. Previous studies based on non-Transformer models (e.g. [Linzen et al., 2016]) showed that NNs for language processing can capture a non-trivial amount of grammatical structure given targeted supervision. [Goldberg, 2019] evaluated instead a pre-trained BERT model with two of the syntactic test cases defined in [Marvin and Linzen, 2018a]: subject-verb agreement and reflexive anaphora. By masking out the single focused verb in each sentence and asking the model for its word predictions, the author showed that BERT is able of capturing syntactic regularities with scores substantially higher than reported in previous work. Relying on the same diagnostic dataset, [van Schijndel et al., 2019] showed that Transformer models (i.e. GPT and BERT) did not consistently outperform the scores achieved with a LSTM model trained on less data. In particular, they noticed that BERT's agreement accuracy tends to decrease as the subject becomes more distant from its verb. [Warstadt et al., 2019] instead investigated BERT's knowledge of negative polarity items (NPIs), i.e. words or expressions that can only appear in environments that are negative (e.g. *any* is an NPI since it can appear only in negative sentences). In particular, they used BERT Masked Language Modeling (MLM) to investigate whether the NLM is able to assign a higher probability to the token from the acceptable sentence in a minimal pair. Results showed that BERT can distinguish between acceptable and unacceptable sentences in the NPI domain. [Ettinger, 2020] proposed instead a set of diagnostics targeting a set of linguistic capacities drawn from human psycholinguistic experiments (e.g. commonsense and pragmatic inference, semantic role sensitivity, negation, etc). Applying these tests to analyze strengths and weakness of the BERT model, the study demonstrated that the model shows sensitivity to role reversal and same-category distinction and it succeeds with noun hypernyms, but it struggles with inferences and role-based event predictions.

Additional line of work deals with the use of the probing framework to test the linguistic competence implicitly learned by state-of-the-art NLMs. Such studies showed that these models are capable of implicitly encode within their representations several

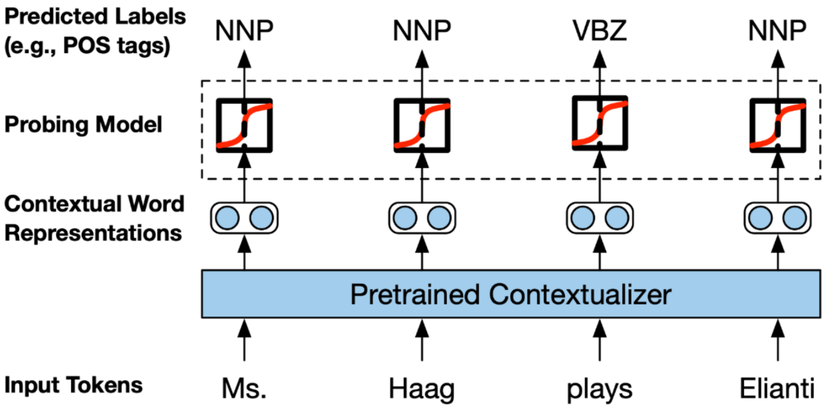

Predicted Labels (e.g., POS tags)
NNP
NNP
VBZ
NNP
Probing Model
Contextual Word Representations
Pretrained Contextualizer
Input Tokens
Ms.
Haag
plays
Elianti

Figure 4.1: *Source: [Liu et al., 2019a]. An illustration of the probing model setup used to study the linguistic knowledge within contextual word representations.*

properties related to morphological, syntactic and semantic information. The idea behind this framework is actually quite simple: using a diagnostic classifier, the *probing model* or *probe* (usually a linear classifier/regressor), that takes the output representation of a NLM as input to perform a *probing task*, e.g. predict a given language property (e.g. POS tags, as in Figure 4.1). If the probing model will predict the property correctly, then we can assume that the NLM representation somehow encodes that property. Formally speaking, let $f : x_i \rightarrow y_i$ be a neural network model mapping a corpus of input sentences $X = (x_1, \ldots, x_n)$ to a set of target labels $Y = (y_1, \ldots, y_n)$ for a learned downstream task. Assume that each sentence x_i is also labeled with some linguistic annotations z_i, reflecting the underlying properties we aim to detect. Let also $h_l(x_i)$ be the network's output at the l-th layer given the sentence x_i as input. To estimate the quality of representations h_l with respect to property z, a supervised model $g : h_l(x_i) \rightarrow z_i$ mapping representations to property values is trained. We take such model's performances as a proxy of $H(h_l(x), z)$. In information theoretic terms, the probe is trained to minimize entropy $H(z|h_l(x))$, and by doing that it maximizes mutual information between the two quantities.

[Alain and Bengio, 2016] were among the first to use linear probing classifiers as tools to evaluate the presence of task-specific information inside neural networks' layers. The approach was later extended to the field of NLP, with studies showing that NLMs learn a variety of language properties in a hierarchical manner [Belinkov et al., 2017, Blevins et al., 2018, Tenney et al., 2019b] and that their representations also support the extraction of dependency parse trees [Hewitt and Manning, 2019]. For instance, [Tenney et al., 2019a] demonstrated that, training a simple probing classifier that has access only to the per-token contextual embeddings of a BERT model, the order in which specific abstractions are encoded within the internal representations reflects the traditional hierarchy of the NLP pipeline: POS tags are processed earliest, followed by constituents, dependencies, semantic roles, and coreference. In addition, they observed that syntactic information is more localizable, with weights related to syntactic tasks

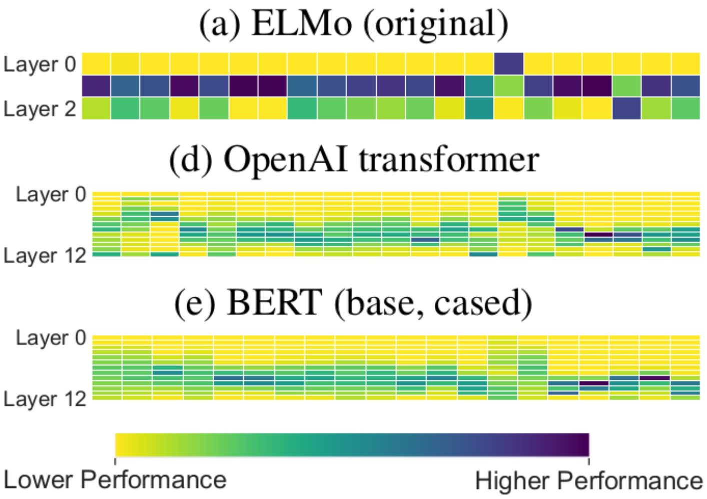


Figure 4.2: *Source: [Liu et al., 2019a]. A visualization of layerwise patterns in task performance. Each column represents a probing task, and each row represents a NLM layer.*

tending to be concentrated on a few layers, while information related to semantic tasks is generally spread across the entire network. In general, the vast majority of works based on the probing tasks paradigm applied to BERT internal representations agreed that syntactic information is most prominent in its middle layers [Hewitt and Manning, 2019, Jawahar et al., 2019, Liu et al., 2019a]. However, as pointed out by [Rogers et al., 2020] in their survey, there are some conflicting evidences about BERT's behaviour on more specific syntactic skills. For example, [Tenney et al., 2019a] and [Jawahar et al., 2019] reported that lower layers are better suited for chunking, while [Liu et al., 2019a] found that both POS-tagging and chunking obtained best probing results relying on the information extracted from middle layers representations.

Most of the recent work on probing representations have focused on BERT. Nevertheless, there are several studies that extended their analysis to other models. For instance, [Liu et al., 2019a] quantified differences in the linguistic competence and in the transferability of individual layers between three contextualized models: ELMo, OpenAI transformer and BERT. Testing the models with a suite of seventeen English probing tasks (e.g. POS tagging, syntactic constituency ancestor tagging, conjunct identification, etc.), they showed that probing models trained with NLMs representations are competitive with state-of-the-art task-specific models, but fail on task requiring fine-grained linguistic knowledge (e.g. grammatical error detection and named entity recognition). They also found that the fist layer output of LSTM-based models (ELMo) is the most transferable, while it is the middle layers for Transformers (See Figure 4.2). [Fayyaz et al., 2021] probed the representations of three NLMs with different pre-training objectives: BERT (masked language modeling), XLNet (permuted language modeling, [Yang et al., 2019]) and ELECTRA (replaced token detection, [Clark et al., 2020]). The results demonstrated that XLNet accumulates linguistic knowledge in the earlier layers than BERT, while ELECTRA linguistic competence is mainly concentrated in the final layers. Moreover, employing RSA similarity measure on fine-tuned models, they showed that XLNet is more prone to forgetting linguistic knowledge in final layers and that the changes in representations are proportional to the gain provided in a downstream task.

Despite this emerging body of work, there are still several open questions: which probing model should we use for assessing the linguistic competence of a NLM? Are probes the most effective strategy to achieve such goal? Moreover, as pointed out by [Tenney et al., 2019a], "the fact that a linguistic pattern is not observed by our probing classifier does not guarantee that it is not there, and the observation of a pattern does not tell us how it is used". These questions fostered complementary lines of research, with the specific aim of identifying new techniques to perform probing tasks or to verify their validity in extrapolating linguistic competence out of NLMs internal representations. Among these, several work investigated which model should be used as probe and which metric should be employed to measure their performances. Some studies advocated that simpler models are better suited as they solve the probing task relying more on the representations received as input rather than on sophisticated abstractions [Hewitt and Manning, 2019, Liu et al., 2019a, Hall Maudslay et al., 2020]. Conversely, [Pimentel et al., 2020b] argued that one should always chose the highest-performing probe regardless of its complexity in order to have a better estimate of the information associated to the representations. On a different note, [Voita and Titov, 2020] suggested a novel information-theoretic metric to measure the actual model effort on the task by balancing probe inner complexity and task performance. Specifically, instead of evaluating probe accuracy, they evaluated *minimum description length* (MDL) of labels given representations, i.e. the minimum number of bits needed to transmit the labels knowing the representations.

Concerning instead works facing the issue of investigating the effectiveness of the probing paradigm, [Hewitt and Liang, 2019] proposed *control tasks*, a set of tasks that associate word types with random outputs. Measuring the difference between linguistic task accuracy and control task accuracy (defined as *selectivity*), they tested whether the probing accuracy on a linguistic task truly reflects the properties of the representation. In particular, they showed that multilayer perceptron (MLP) probes achieve very low selectivity, suggesting caution in interpreting how their results reflect properties of representations. Along the same line, [Ravichander et al., 2021] tested probing tasks by creating control datasets where a property is always reported in a dataset with the same value, thus it is not discriminative for testing the information contained in the representations. Their experiments highlighted that the probe may learn a property also incidentally, thus casting doubts on the effectiveness of probing tasks.

4.2 Analysis of Attention Mechanism

Another body of work is focused on the analysis of attention mechanism in contextualized NLMs and, specifically, on how linguistic knowledge is directly encoded in model's self-attention heads. [Clark et al., 2019] found that the majority of BERT's

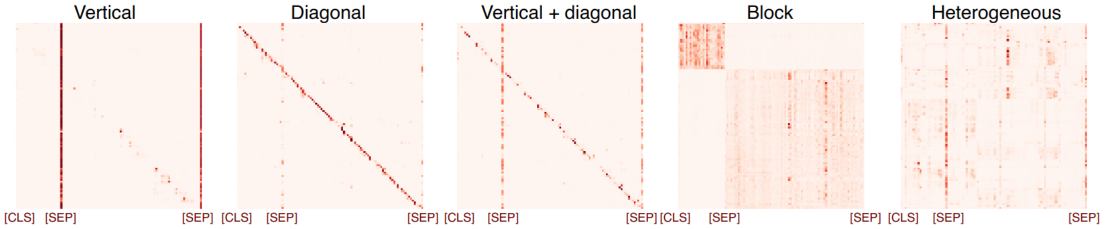


Figure 4.3: *Source: [Kovaleva et al., 2019]. Both axes on every image represent BERT tokens of an input example, and colors denote absolute attention weights (darker colors stand for greater weights). The first three types are most likely associated with language model pre-training, while the last two potentially encode semantic and syntactic information.*

attention heads focus on specific group of tokens and, especially, on the special tokens *[CLS]* and *[SEP]*[1] and on periods and commas. Nevertheless, evaluating the directions of prediction of each attention head, they also showed that certain heads specialize to specific dependency relations: e.g. direct objects attending to their verbs or noun modifiers attending to their nouns. [Kovaleva et al., 2019], instead, identified five frequently occurring patterns within BERT self-attention heads (see Figure 4.3) that are consistently repeated across different heads and tasks. Moreover, they demonstrated that most heads do not directly encode any non-trivial linguistic information when a model is fine-tuned on specific downstream tasks, since only fewer than 50% of heads exhibit an "heterogeneous" pattern. Proposing a novel methodology to observe, for a given syntactic phenomenon, the intrusion effects of distractors on BERT's self-attention mechanism, [Lin et al., 2019] found that BERT encode aspects of syntactic structure that are relevant for subject-verb agreement and reflexive dependencies through attention weights, and that this information is represented more accurately on higher layers. [Vig and Belinkov, 2019] measured the proportion of GPT-2's total attention from a head that focuses on tokens with specific POS tags or connected in a dependency relations. The results obtained showed that: i) attention heads target specific POS tags depending on layer depth; ii) attention aligns with dependency relations most strongly in the middle layers.

Similar to the approaches previously described, the view that attention heads have a clear meaning and, therefore, they could be used as proxies for the interpretation of NLMs' internal mechanisms is currently debated [Jain and Wallace, 2019, Brunner et al., 2020]. For instance, [Serrano and Smith, 2019] found that, modifying attention values of a NLM model by hard-setting the highest attention values to zero, the number of attention values that must be set to zero to alter the model's prediction is often too large, thus concluding that attention is not suitable for interpreting model's decision. Challenging the assumptions underlying [Serrano and Smith, 2019], [Wiegreffe and Pinter, 2019] showed instead that alternative attention distributions found via adversarial training methods perform poorly relative to traditional RNNs attention mechanisms

[1] *[CLS]* is a special symbol added in front of every input example and *[SEP]* is a special separator token.

when used in a diagnostic MLP model. Following these results, the authors suggested that attention mechanisms in RNNs do in fact learn something meaningful about the relationship between tokens and model's prediction.

It is also important to note that the vast majority of work based on the analysis of Transformer-based attention mechanisms focused mainly on attention patterns. However, the Transformer architecture is not only composed of the multi-head attention. For instance, [Kobayashi et al., 2021] extended the analysis to the whole attention block, i.e. multi-head attention, residual connection and layer normalization. Specifically, the authors introduced a decomposition of the operations in the whole attention block exploiting the norm-based analysis previously defined in [Kobayashi et al., 2020]. Testing their approach on 32 variants of the masked language models (BERT with five different sizes, BERT-base trained with 25 different seeds and RoBERTa with two different sizes), they showed that the operation via residual connection and layer normalization contributes more to the internal representations than expected, thus suggesting a relative small impact of the multi-head attention.

4.3 Correlating NLMs and Linguistic Abilities

Alternative approaches have been proposed to analyze the linguistic knowledge implicitly encoded by state-of-the-art NLMs. As mentioned previously, several studies focused on the relationship between internal representations or perplexity scores and grammatical generalization abilities. For instance, [Chrupała and Alishahi, 2019] proposed an approach based on Representation Similarity Analysis (RSA) [Kriegeskorte et al., 2008] to study the correlation between neural representations of strings and structured symbolic representations of these strings. The core idea of RSA is to find connections between data from computational modeling by correlating representations spaces via their pairwise similarities. In order to measure the similarities between these symbolic representations, the authors used *tree kernels*, i.e. a metric to compute the proportion of common substructures between trees. Testing their approach on a sample of data from the English Web Treebank (EWT) [Silveira et al., 2014] and with increasing values of λ[2], they showed that BERT[3] encode a substantial amount of syntactic information compared to random models and simple bag-of-words representations. Moreover, they found that RSA correlation scores peak between layers 15-22, thus indicating that the final layers of the model are more dedicated to encoding properties of sentence other than syntax.

[Saphra and Lopez, 2019], instead, investigated how representations of linguistic knowledge are learned over time in a NLM relying on SVCCA (Singular Vector Canonical Correlation Analysis) [Raghu et al., 2017], a general method to compare the correlation of two vector representations. In particular, they compared the representations of a simple two-layer LSTM language model at each epoch of training with representations of other models trained to predict specific linguistic categories (e.g. POS

[2] The parameter λ is used to scale the relative importance of tree fragments with their size.

[3] They tested the 24-layer version of the BERT model.

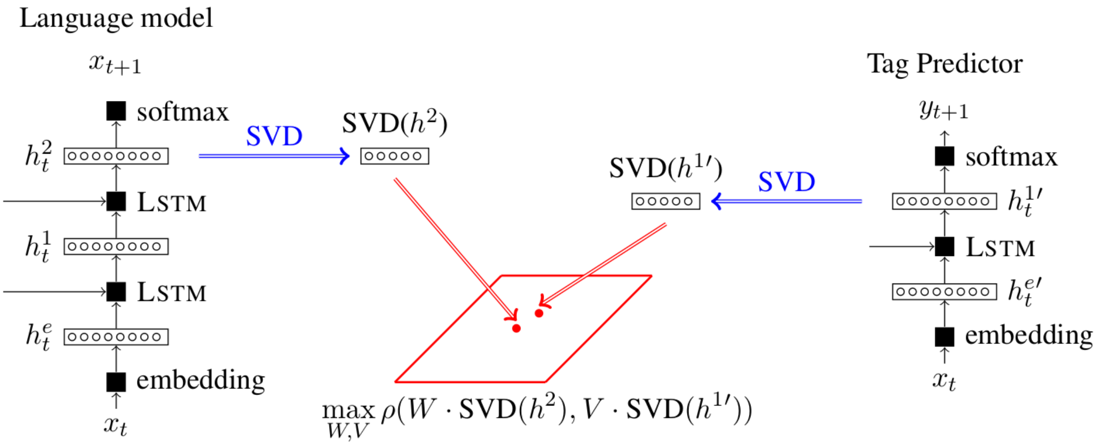


Figure 4.4: *Source: [Saphra and Lopez, 2019]. SVCCA used to compare the layer h^2 of a language model and layer $h^{1\prime}$ of a tagger.*

or semantic tagger, as showed in Figure 4.4). The results showed that different aspects of linguistic structure are learned at different rates within a single recurrent layer: POS tags are acquired earlier, while global topic information are continuously learned during training.

Other studies focused instead on the relationship between NLM perplexity scores and syntactic knowledge. Perplexity is an evaluation metric for language models and it measures how well a probability distribution (or probability model) predicts a sample. With the recent success gained by NLMs across a variety of NLP tasks, the notion of perplexity has started being investigated also to dig into issues related to the interpretability of contextual word representations, with the aim of understanding whether there is a relationship between this metric and the grammatical abilities implicitly encoded by a NLM [Gulordava et al., 2018, Marvin and Linzen, 2018b, Kuncoro et al., 2019]. In this context, [Hu et al., 2020] proposed a systematic evaluation of the syntactic knowledge of NLMs by investigating the relationship between a model's perplexity and its performance on targeted syntactic assessments. Specifically, they assembled a large set of test suites (e.g. syntactic coverage, garden-path effects, gross syntactic expectation, etc.) inspired by the methodology of experimental sentence processing and psycholinguistic research and tested it on four classes of neural models: LSTM, ON-LSTM [Shen et al., 2018], Recurrent Neural Network Grammars (RNNG, [Dyer et al., 2016]) and GPT-2. The results showed that, despite all models achieved high syntactic generalization (SG) scores, there is a dissociation between perplexity and SG scores, thus suggesting that targeted syntactic evaluation can reveal information that may be orthogonal to perplexity. Similarly, [Warstadt et al., 2020], in their work that introduces the Benchmark of Linguistic Minimal Pairs (BLiMP)[4], tested three NLMs (LSTM, GPT-2 and Transformer-XL, [Dai et al., 2019]) by observing whether they assign a higher probability to the acceptable sentence in each minimal pair of the

[4] https://github.com/alexwarstadt/blimp

dataset. The authors concluded that neural LMs are able to acquire robust knowledge of morphological agreement and some syntactic phenomena, e.g. ellipsis and control/raising. Nevertheless, these models show weaker evidence of knowledge about argument structure or the semantic properties of quantifiers. It is also interesting to notice that, in contrast with the results obtained in [Warstadt et al., 2019] with the BERT model, the model tested on the BLiMP benchmark seem to struggle in distinguishing between acceptable and non-acceptable sentences when dealing with NPI licensing.

Part II

Tracking the Evolution of Written Language Competence

CHAPTER 5

Tracking the Evolution of Written Language Competence in L1 Italian Learners

In this chapter we present the work we have carried out in order to monitor the development of written language competence in L1 learners. Specifically, we focus on the experiments we devised in order to assess the evolution of writing skills of Italian L1 learners exploiting a classification approach that makes use of a wide range of linguistic features automatically extracted from morpho-syntactic and syntactic annotated texts.

5.1 Introduction

Moving in the framework of studies based on L1 language development, we introduced an NLP–based stylometric approach to model the evolution of written language competence in Italian L1 learners. According to the core assumptions of computational stylometry [Daelemans, 2013], formal properties of a text characterizing its style can reveal underlying traits about the author, e.g. in terms of gender, age, ethnicity, as well as language proficiency. However, while traditional stylometric techniques are typically based on a close set of ad-hoc linguistic features selected according to specific task in mind (e.g. authorship attribution, authorship verification, gender classification), our approach relies on a wide set of linguistically motivated features extracted from students' essays, which have already shown to be effective in several scenarios, all related to modeling the 'form' of a text, rather than the content: from the prediction of human judgments of perceived linguistic complexity [Brunato et al., 2018] to the automatic identification of the native language of a speaker based on his/her productions in a second language (L2) [Malmasi et al., 2017].

The proposed approach is developed and tested on texts contained in the *CItA* (*Corpus Italiano di Apprendenti L1*) corpus, the first longitudinal corpus of essays written by Italian L1 learners enrolled in the first and second year of lower secondary school [Barbagli et al., 2016]. As stated by their creators, this two–year period is considered as crucial for the development of written language, which undergoes remarkable changes both in the way students write and how they approach to writing, as a consequence of being exposed to a more formal way of writing teaching from the first to the second year of lower secondary school. The longitudinal nature of the corpus, complemented with the emphasis on the importance of the learning period under investigation, makes *CItA* particularly suitable to test the effectiveness of a computational model of writing development in L1 learners. Specifically, we decompose this problem into two main research questions:

- Is it possible to track the individual learning trajectory in writing by automatically predicting the chronological order of two essays written by the same student at different time?
- Which typologies of language phenomena contribute more to the prediction task and how they change according to different temporal spans?

5.1.1 Our Approach

In order to track how written language competence evolves in the two considered school grades, we ask whether the writing development curve of a student can be automatically learned. We model the problem as a binary classification task in which a machine learning classifier has to predict the relative order of two essays using a wide set of linguistically motivated properties automatically extracted from the L1 learners essays contained in the *CItA* corpus.

The classifier uses a Linear Support Vector Machine (LinearSVM) as machine learning algorithm. We rely on LinearSVM rather than more powerful learning algorithms, such as Recurrent Neural Networks (RNNs), in order to obtain meaningful explanations when the classifier outputs its predictions, so as to anchor the observed patterns of language development to explicit linguistic evidence. To prevent overfitting, we train and test our model in a cross-domain manner, using essays of students from different schools during the training and testing phase. Doing so, the algorithm is tested not only on essays written by different students, but also on students coming from different schools.

We further extract and rank the feature weights assigned by the LinearSVM in order to understand which typology of linguistic features contributes more to the classification task. The underlying assumption is that the higher will be the weight associated with a specific feature, the higher will be its importance in solving the classification task and, consequently, in tracking the students written language evolution.

Finally, to provide first insights into the possible influence of background variables on predicting writing development, we ran the same binary classification task distinguishing students enrolled in the center and suburban schools and we assessed the confidence of our classifier in the two scenarios. Since the confidence reflects the uncertainty of the model estimates (i.e. the higher the confidence the easier the prediction was for the classifier), this measure can be viewed as a mean to approximate the degree of changes in the learning curve of each student. That is, we can assume that the classifier is more confident when the two essays for which the relative order has to be predicted show greater differences with respect to the considered features.

In what follows, we first introduce the two main ingredients of our approach, namely the corpus and the set of linguistic features. We then describe the set-up of the experiments and discuss the obtained results in light of the main research questions of the study.

5.1.2 The CItA Corpus

As previously discussed, the availability of authentic texts produced by language learners is of pivotal importance. Such resources can differ according to the modality (i.e. written texts or speech transcriptions), the typologies of learners considered (e.g. preschool children, first or second language students), the goals of analysis (e.g. theoretical studies or development of educational applications). For the purpose of our study, we relied on CItA (*Corpus Italiano di Apprendenti L1*), a longitudinal corpus of essays written by the same students in the first and second year of lower secondary school [Barbagli et al., 2016]. This makes the corpus particularly suitable to track the evolution of L1 written language competence over the time. The corpus was collected during the two school years 2012-2013 and 2013-2014 as part of a broader on-going study carried out in the framework of the IEA[1]-IPS (*Association for the Evaluation of Educational Achievement*) activities [Lucisano, 1984]. As stated by their creators, the collection of essays in *CItA* was motivated by two underlying hypotheses. The former is that students' competence in writing undergoes a variety of relevant changes from the first to the second year of lower secondary school, as a consequence of being exposed to a more formal writing teaching. The latter is that the development of written language competence could be related to background variables of students, such as the city area where the school is located (historical center or suburbs), the language(s) the students speak at home or their parents' employment. To make it possible exploring the effects of these variables, the *CItA* essays were collected from 7 different schools located in Rome, 3 of which in the historical center and 4 in suburbs. In addition, all students whose essays are comprised in the final corpus were asked to answer to a questionnaire of 34 questions to obtain information about their biographical, socio-cultural and sociolinguistic background. For example, they were asked to provide biographical information such as the language(s) the students usually speak at home, when and where they were born, their parents' education, etc.

[1] http://www.iea.nl

	First year				**Second year**		
	School	Students	Essays		School	Students	Essays
Center	1	25	123	Center	1	25	108
	2	27	143		2	28	130
	3	24	138		3	23	117
	School	Students	Essays		School	Students	Essays
Suburbs	4	21	58	Suburbs	4	22	62
	5	19	77		5	19	64
	6	24	66		6	24	146
	7	13	64		7	14	56
Total		153	669	Total		155	683

Table 5.1: *Composition of the corpus.*

Typology	**Prompt example**
Reflexive	*What's your attitude regarding the reading activity?*
Narrative	*Narrative essay in which you describe an episode of bullying*
Descriptive	*Describe a primary school teacher you are particularly close to*
Expository	*Write a news story that the media has been dealing with recently*
Argumentative	*Mobile phones in class: what do you think about it and how do you think it could be solved?*
Common Prompt	*A boy younger than you has decided to enroll at your school. He wrote to you to ask you how to write an essay that can get good grades by your teachers. Send him a friendly letter describing at least five points that you believe are important for your teachers when they evaluate an essay.*

Table 5.2: *Prompts examples according to the different typologies.*

Corpus description

The corpus contains a total of 1,352 essays written by 156 students (see Table 5.1). The essays belong to five textual typologies, which reflect the different writing prompts students were asked to respond: reflexive, narrative, descriptive, expository and argumentative. In addition, a prompt common to all schools was also assigned at the end of each year. Specifically, at the end of second year, students were asked to respond to the Italian version of Task 9 of the IEA-IPS study ([Lucisano, 1984]; [Visalberghi and Costa, 1995]), i.e. a letter of advice to a younger student on how she/he should write in order to get good grades at high school; at the end of the first year, they were presented with a modified version of Task 9 still with the same aim. Table 5.2 shows examples of prompts given to the students according to the different typologies.

As shown in Table 5.3, there are some differences over the two years and the seven schools. First of all, it can be noted that the number of prompts differs among the seven schools: teachers of the schools located in the city center tend to give a higher numbers of prompts than their colleagues in the sub-urban schools. Secondly, if reflexive prompts are the most frequent textual type in the two years, from the first to the second year the distribution of narrative prompts are halved while the expository and argumentative ones are doubled. This different distribution is a consequence of the approach adopted by teachers to teach writing: writing a narrative essay is considered as a simpler task since it requires more rudimentary cognitive and writing skills, than writing an argumentative

Typology	Center	Suburbs	Total
First year			
Reflexive	25	13	38
Narrative	18	4	22
Descriptive	2	1	3
Expository	0	1	1
Argumentative	2	2	4
Sub-total	47	21	68
Second year			
Reflexive	24	5	29
Narrative	3	6	9
Descriptive	0	0	0
Expository	4	5	9
Argumentative	5	4	9
Sub-total	36	20	56

Table 5.3: *Distribution of typologies of prompts.*

or expository essays, for which more complex linguistic and discourse-structuring competences are required [Barbagli, 2016].

Error Annotation

One of the characteristics that mostly distinguishes *CItA* from other corpora of L1 Italian learners, such as those described in [Marconi, 1994], is that the essays were annotated according to different types of linguistic errors occurring in text. Error annotation is a challenging issue since it assumes that a deviation from a linguistic norm is occurring, a norm which is in its turn an arbitrary concept defined only according to social conventions. Moreover, the annotation of errors in L1 corpora is a much less common practice than in L2 corpora, where this level of information is typically used to investigate the properties of *interlanguage* [Brooke and Hirst, 2012] or as a reference resource for automatic error detection and correction tasks [Dahlmeier et al., 2013]. In the absence of a L1 error taxonomy already available for the Italian language, it was defined a new scheme inspired to Berruto's definition of "neo-standard Italian" as linguistic norm [Berruto, 1987] following the literature on the evaluation of written skills of L1 Italian learners ([Visalberghi and Costa, 1995]; [De Mauro, 1983]; [Colombo, 2010]).

As shown in Table 5.4, it is a three-level schema including grammatical, orthographic and lexical errors, which makes it also similar to already existing schemes in other languages (e.g. [Granger, 2003] for French as a second language). Following the the annotation format proposed by [Ng et al., 2013], *CItA* was annotated as follows:

> [...] scapparono al piano di sopra e dal <M t="200" c="buio">buglio</M> <M t="113" c="spuntò">spuntarono</M> un esercito [...]
>
> ([...] they ran away upstairs and from the darkness an army appeared)

Class of error	Type of Modification	**I year**	**II year**
		Freq %	Freq %
Grammar			
Verbs	**Use of tense**	7.78	15.67
	Use of mood	4.25	4.92
	Subject-Verb agreement	2.85	4
Prepositions	**Erroneous use**	6.48	6.75
	Omission/Redundancy	1.03	0.72
Pronouns	Erroneous use	5.09	3.54
	Omission	0.41	0.59
	Redundancy	2.70	1.57
	Erroneous use of relative pronoun	2.13	1.70
Articles	**Erroneous use**	5.81	3.54
Conjunctions	Erroneous use	0.57	0.52
Other		7.31	5.18
Total		46.41	48.7
Orthography			
Double consonants	**Omission**	6.74	5.05
	Redundancy	3.27	3.67
Use of *h*	**Omission**	3.21	1.64
	Redundancy	1.66	1.11
Monosyllables	**Erroneous use of monosyllabic words**	4.87	4.07
	po and *pò* instead of *po'*	1.66	1.64
Apostrophe	**Erroneous use**	4.82	4.52
Other		21.77	23.02
Total		47.63	44.72
Lexicon			
Vocabulary	**Erroneous use**	5.60	6.56

Table 5.4: *Error annotation schema. Errors varying significantly over the two years (i.e. p < 0.05) are bolded.*

where the textual span of error is marked by <*M*> and <*/M*> (*Mistake*), the attribute *t* (*type*) is the macro-class and subclass of error, and *c* (*correction*) reports the corrected form. In the reported example, there is a generic orthographic error (the word *buglio* instead of the correct one *buio*) and a grammatical mistake concerning Subject-Verb agreement (the third person plural of the verb instead of the required third person singular).

The annotation was manually performed by a teacher of lower secondary school and revised by two undergraduate students in digital humanities, who were adequately trained on the task annotation guidelines.

Inspecting the statistical distribution reported in Table 5.4, it can be noted that in both years (Rows *Total*) orthographic and grammatical errors are the most frequent ones (47.63–44.72% and 46.41–48.7% respectively) while lexical errors are far less (about 6%). More specifically, the most frequent errors affect the area of orthography without distinction into specific typologies (i.e. the class *Other*) (22.32%) followed by the erroneous use of verb tenses (11.26%), the grammatical not–classified errors (6.37%) and the erroneous use of prepositions (6.6%). Note that the majority of errors has a statistically significant variation over the two years thus showing that several common trends in the development of writing competence occur during the transition from the first to the second year.

Linguistic Annotation

To allow the extraction of linguistic features used as predictors of writing development in the classification experiments, the *CItA* corpus was firstly automatically annotated using UDPipe [Straka et al., 2016], a NLP pipeline carrying out basic pre-processing steps, i.e. sentence splitting and tokenization, POS tagging, lemmatization and syntactic parsing, according to the Universal Dependencies (UD) annotation framework [Nivre et al., 2016]. Although we used a state–of–the art pipeline, it is well-acknowledged that the accuracy of statistical parsers decreases when tested against texts of a different typology from that used in training [Gildea, 2001]. In this respect, learners' data are particularly challenging for general–purpose text analysis tools since they can exhibit deviation from correct and standard language [Berzak et al., 2016]. For instance, missing or anomalous use of punctuation (especially in 1st grade prompts) could already impact on the coarsest levels of text processing, i.e. sentence splitting, and thus may affect all subsequent levels of annotation. Nonetheless, if we can expect that the predicted value of a given feature might be different from the real one (especially for features extracted from more complex levels of annotation such as syntax), we can also assume that results will be consistent, at least when parsing texts of the same domain. The validity of this claim has been shown in other studies relying on engineered features similar to ours for classification or linear regression analyses. For instance, [Dell'Orletta et al., 2011b] proved that the values of a set morpho-syntactic and syntactic dependency features are comparable when extracted from a gold (i.e. manually annotated) and an automatically annotated corpus of the same domain (i.e. biomedical language). In a study aimed at investigating dependency distance minimization in English using a large diachronic corpus, [Lei and Wen, 2020] checked whether any possible errors from the parser significantly affected the results of their analysis. To this end, they manually revised the annotation of a subset of the automatically parsed corpus under investigation and correlated the values of their examined features (i.e. mean and normalized dependency distance) extracted from the automatically and the manually revised portion, obtaining very high correlation scores. We applied a similar approach to our corpus in order to observe the impact of possible parsing errors on the reliability of the feature extraction process with respect to learner data. Specifically, we randomly extracted a few parsed sentences from both I and II-year CItA essays for a total of $\sim$800 tokens and we manually revised the output of the automatic annotation in every step. We then extracted all monitored features from the manually revised sentences and compared these values to the corresponding ones extracted from the automatically parsed sentences. The resulting Spearman's rank correlation coefficient between the two samples shows that, with the only exception of the distribution of parataxis relations (*dep_dist_parataxis*), linguistic features extracted from automatically annotated and manually revised sentences are extremely highly

Level of Annotation	Linguistic Feature	Label
Raw Text	Sentence Length	tokens_per_sent
	Word Length	char_per_tok
	Document Length	n_sentences
	Type/Token Ratio for words and lemmas	ttr_form, ttr_lemma
POS tagging	Distribution of UD and language–specific POS	upos_*, xpos_*
	Lexical density	lexical_density
	Inflectional morphology of lexical verbs and auxiliaries	verbs_*, aux_*
Dependency Parsing	Depth of the whole syntactic tree	parse_depth
	Length of dependency links and of the longest link	links_len, max_links_len
	Average length of prepositional chains and distribution by depth	prepositional_chain_len, prep_*
	Clause length (n. tokens/verbal heads)	token_per_clause
	Order of subject and object	subj_pre, subj_post, obj_pre, obj_post
	Verb arity and distribution of verbs by arity	verb_edges, verb_edges_*
	Distribution of verbal heads per sentence	verbal_head_sent
	Distribution of verbal roots	verbal_root_perc
	Distribution of dependency relations	dep_*
	Distribution of subordinate and principal clauses	principal_prop, subord_prop
	Length of subordination chains and distribution by depth	subord_chain_len, subord_*
	Relative order of subordinate clauses	subord_post, subord_prep

Table 5.5: *Linguistic features used in the experiments.*

correlated (average $\rho = 0.93$).

5.1.3 Linguistic Features

To extract our set of linguistic features, we relied on Profiling–UD [Brunato et al., 2020], a multilingual tool specifically conceived to carry out linguistic profiling on corpora annotated in UD–style. Universal Dependencies (UD) [Nivre et al., 2016] is an ongoing project aimed at developing corpora with a cross-linguistically consistent annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. The choice of relying on UD–style annotation makes the process of feature extraction language–independent, since similar phenomena are annotated according to a common annotation scheme at morpho–syntactic and syntactic level of analysis.

Profiling–UD allows the computation of a wide set of features encoding a variety of lexical and grammatical properties of a text informed by the literature on linguistic complexity and language development. They range from superficial ones, such as the average length of words and sentences, to morpho–syntactic information concerning the distribution of parts-of-speech (POS) and the inflectional properties of verbs, to more complex aspects of syntactic structure deriving from the whole parse tree and from specific sub–trees (e.g. subordinate clauses.). The set of features is reported in Table 5.5 according to the level of annotation from which they derive.

By looking at the statistical distribution of these features, it can be noted, for example, that the essays written in the second year contain a lower percentage of conjunctions, pronouns (especially clitic and personal ones), and a higher percentage of prepositions and nouns with respect to the essays of the first year (Table 5.6). These statistically significant differences[2] suggest that II–year students possibly exploit more the *pro-drop* potentiality of the Italian language in their writing, thus making less use of overt

[2] The statistical significance for all features reported in the next three Tables was assessed using the Wilcoxon-rank-sum-test.

Feature	I year (%)	II year (%)
Conjunctions	6.88	6.38
Determiners	13.86	14.12
Preposition	10.53	11.21
Pronouns	8.97	8.04
Clitic pronouns	4.58	4.08
Personal pronouns	1.58	1.2
Nouns	16.02	16.38

Table 5.6: *Distribution of major morpho–syntactic features varying significantly between the two school years.*

Features	I year (%)	II year (%)
preverbal subjects	84.19	82.57
postverbal subjects	15.81	17.14
preverbal objects	35.69	30.39
postverbal objects	64.31	69.61
nominal subjects (dep_nsubj)	5.59	5,04
passive subjects (dep_nsubj:pass)	0.19	0.28
adverbial clause modifiers (dep_advcl)	0.53	0.62
copular constructions (dep_cop)	2.13	1.89
coordination (dep_cc)	4.38	4.14
parse depth	4.589	4.716

Table 5.7: *A subset of syntactic features varying significantly between the two school years.*

Feature	I year (%)	II year (%)
Indicative mood	94.83	92.60
Subjunctive mood	2.61	3.31
Imperfect tense	16.48	10.99
Present tense	42.36	49.28
Verbs-1PerSing	15.22	13.55
Verbs-1PerPlu	6.96	5.25

Table 5.8: *Distribution of verbal morphology features (mood, tense and person) varying significantly between the two school years.*

pronouns. At syntactic level (Table 5.7), this speculation seems to be corroborated by the lower distribution in second year's essays of syntactic relations linking a nominal subject (either headed by a noun phrase or realized as a pronoun) to its verbal head (*dep_nsubj*). Moreover, when the subject is overtly expressed, it tends to be placed in the canonical position (i.e. left to the verb since Italian is a SVO language), especially by younger writers.

While the distribution of verbs is almost similar between the two years (i.e. around 13%, without significant variation), the use of verbal morphology changes from the first to the second year (Table 5.8). As could be expected, the indicative mood is predominant in all essays, although in the second year students start using in a slightly higher percentage also more complex moods, such as the subjunctive. Instead, a greater variation affects the use of tenses, especially the imperfect one, which decreases significantly in the second year. On the one hand, this could be expected since imperfect indicative verbs are easier than other past tenses of the Italian verbal morphology. On the other hand, this variation might be related to the different type of essays assigned in the two years. In fact, in the second year the typology of narrative essays, for which is commonly required the use of imperfect tenses, is less predominant across prompts. In this regard, also the more extensive use of first singular and plural person verbs in essays written younger students is indicative of a more subjective writing style.

5.1.4 Tracking the evolution of written language competence

Our first research question was aimed to explore whether it is possible to automatically track the development of students' writing competence across time. We model this problem as a classification task, starting from the assumption described in [Richter et al., 2015]: given a set of chronologically ordered essays written by the same student, a document d_j should show a higher quality level with respect to the ones written previously (d_i). Thus, given two essays d_i and d_j written by the same student, we want to classify whether $t(d_j) > t(d_i)$, where $t(d_i)$ is the time in which the document d_i was written.

For this purpose, we built a classifier operating on morpho-syntactically tagged and dependency parsed essays which assigns to each pair of documents (d_i, d_j) a score expressing its probability of belonging to a given class: 1 if $t(d_j) > t(d_i)$, 0 otherwise. For each pair of essays, we built an *E* vector:

$$E = V_i + V_j + (V_i - V_j) \tag{5.1}$$

where V_i and V_j are, respectively, the feature vectors of the first and second essays, and $V_i - V_j$ is the vector difference between them. Vectors are composed by the values of multi-level linguistic features both automatically extracted, as shown in Sec. 5.1.1, and manually annotated (i.e. features related to the error annotation) in *CItA*. As previously mentioned, the classifier uses linear Support Vector Machines (SVM) as machine learning algorithm.

We split all texts of the *CItA* corpus into four sets, pairing essays written by the same students considering all the possible temporal spans at the same time (*All essays*) and considering only essays written at a distance of one month (*1 month*), one year (*1 year*) and two years (*2 years*). Table 5.9 summarizes the statistics of the four datasets. We evaluated the system with a 7-fold cross validation in which every fold is represented by a different school. It follows that in each experiment the test set is composed by documents which are not included in the corresponding training set.

Each line of the training and test sets follows this structure:

Student code, Label, E vector

where *Student code* is an identifier assigned to the student, *Label* could be 1 or 0 depending on the two essays' order and *E vector* is the feature event associated with the

Temporal span	**Number of samples**
All essays	7,228
1 month distance	1,308
1 year distance	348
2 years distance	208

Table 5.9: *Number of samples/E events within each dataset.*

	Samples	**#1**	**#2**	**#3**	**Baseline**
All essays	7,228	0.53 ± 0.08	0.55 ± 0.09	**0.58 ± 0.09**	0.45 ± 0.06
1 month	1,308	0.49 ± 0.03	0.50 ± 0.05	**0.54 ± 0.04**	0.50 ± 0.03
1 year	348	0.54 ± 0.15	0.63 ± 0.09	**0.65 ± 0.15**	0.61 ± 0.12
2 years	208	0.66 ± 0.16	0.71 ± 0.15	**0.75 ± 0.14**	0.40 ± 0.14

Table 5.10: *Cross-school results (in terms of weighted accuracy ± standard deviation) for the three sets of experiments.*

two essays.

Three different sets of experiments were devised to test the performance of our system, which differ with respect to the number and type of linguistic features extracted for each essays. In the first set (*#1*) we used only the lexical, morpho-syntactic and syntactic features extracted from the parsed corpus. In the second set of experiments (*#2*) we added to them a set of features related to word frequency (*word frequency class*), which was measured as the average class frequency of all lemmas in the document. The class frequency was computed for each lemma and form exploiting the *itWAC* (*Italian Web as Corpus*) corpus[3] as follows:

$$C_l = \lfloor \log_2 \frac{freq(MFL)}{freq(CL)} \rfloor \tag{5.2}$$

$$C_{wf} = \lfloor \log_2 \frac{freq(MFF)}{freq(CF)} \rfloor \tag{5.3}$$

where *MFL* and *MFF* are, respectively, the most frequent lemma and word form of the corpus, and *CL* and *CF* are the considered lemma and word form. In the third experiment (*#3*) we expanded our set of linguistic features with those related to the distribution of the different kinds of errors annotated in *CItA* (Section 5.1.2): grammatical errors; orthographic errors; lexical errors and punctuation errors.

In order to verify the effectiveness of our model, we compared our classification results with the ones obtained with a baseline computed with a LinearSVM that takes as input the average sentence length of the essays for each sample pairs. Classification results are reported in Table 5.10.

As a general remark, we observe that the larger the temporal span between the tested documents, the higher the achieved accuracy. Not only does this suggest that the pairs of essays written by each student at more distant times exhibit a quite divergent

[3] A 1.5 billion words corpus made up of texts collected from the Web [Baroni et al., 2009]

Features	Two years distance
Grammatical errors	**0.74**
Orthographic errors	0.72
Lexical errors	0.70
Punctuation errors	0.68

Table 5.11: *Classification results using different sets of annotated error features.*

linguistic profile – which makes the classification task easier –, but also that linguistic patterns underlying writing development are consistent across students and schools. Remember indeed that in all the experiments the classifier is tested on essays written by different students, but also on students coming from different schools. If we compare the results obtained considering the *1 month* and *2 years* time intervals we can notice an improvement of 20% in terms of accuracy scores. As expected, accuracy scores in the *1 month* temporal span are comparable with those obtained with the simple baseline, proving that the complexity of this task does not allow to obtain reliable results when considering excessively short time intervals.

Focusing on the three different set of experiments, we can see that the results tend to improve as more features are used for classification. In particular, the contribution of vocabulary–related features operationalized in terms of word frequency is particularly effective when considering essays written at a long term distance, such as *1 year* or *2 years*. In addition, differently from what is reported in [Richter et al., 2015], we observe a general improvement when lexical, morpho–syntactic and syntactic features are complemented with features related to the distribution of errors made by students.

With this respect, to have a better understanding of their contribution in the automatic classification, we repeat our experiments using the four sets of error-related features (i.e. grammatical, orthographic, lexical and punctuation) in a separate way. As shown in Table 5.11, the improvement is due to the presence of grammatical errors. Indeed, the accuracy obtained using only this typology of errors, in addition to general linguistic features, is even higher than the one obtained using the four sets of errors together (from 0.73% to 0.74%). These data are in line with the qualitative observations reported in Section 5.1.2: since grammatical errors, as well as orthographic errors, undergo a significant variation over the two school years, they allow the classifier to obtain better results.

Cross-Prompt Testing

As reported in Section 5.1.2, the assigned prompts are differently distributed over the two years. This observation may cast doubts on the effectiveness of our features to serve as real proxies of writing development rather than as prompt–related characteristics. To discard this hypothesis and verify whether the results we obtained generalize across prompts, we replicate the experiments in a cross-prompt scenario. In particular, we used the four datasets previously described (*All essays*, *1 month*, *1 year* and *2 years*) and we performed the experiments with a cross–prompt validation strategy, i.e. testing

	Samples	#1	#2	#3	Baseline
All essays	2,662	0.64 ± 0.04	0.64 ± 0.04	**0.67 ± 0.04**	0.52 ± 0.01
1 month	532	0.47 ± 0.02	0.46 ± 0.05	**0.50 ± 0.01**	0.48 ± 0.04
1 year	128	0.53 ± 0.05	0.53 ± 0.05	**0.68 ± 0.10**	0.65 ± 0.16
2 years	119	0.82 ± 0.04	0.84 ± 0.05	**0.85 ± 0.05**	0.48 ± 0.01

Table 5.12: *Cross-prompt results (in terms of weighted accuracy ± standard deviation) for the three set of experiments along with total test samples size (Samples).*

the resulting model only on pairs of essays that have the same prompt. The new classification results are reported in Table 5.12. As we can notice, our model achieved better results with respect to the length baseline for all the datasets and according to the three sets of experiments, thus allowing us to confirm that the system clearly generalizes across prompts and is actually modeling written language evolution rather than prompt-dependent characteristics.

5.1.5 Studying linguistic phenomena

The results obtained in the previous experiments showed that it is possible to predict the chronological order of two essays written by the same student by using features of different nature. This confirms that relevant transformations occur in L1 writing during the transition from the first to the second year of lower secondary school. However, very little has been said about the contribution of each single feature in the classification tasks. Since we showed that not all the linguistic features vary significantly during the 2-year temporal span, we can reasonably assume that within the set of our features, some of them are also more predictive than others for the classification. To better explore this question, we established a ranking of the most important features according to the different classification scenarios. To do this, we evaluated the importance of each linguistic property by extracting and ranking the feature weights assigned by the LinearSVM model that uses features of all categories (i.e. linguistic features, word class features and error–related ones).

Table 5.13 shows the rankings of the 20 most important features according to three considered temporal spans. As we can see, error–related features acquire relevance as the temporal span increases: in the second classification experiment, where the task was to predict the chronological order of essays written at a distance of one year, three of the ten most significant features derive from error annotation. Similarly, in the third classification scenario, error-related features occur three times in top-ranked positions and one of them, i.e. omission of pronouns, is the first ranked one. The omission of pronouns in required contexts, complemented with their unnecessary use (i.e. *error_pronouns_redundancy*, 12th-ranked), could be indicative of the influence of spoken language phenomena on written texts by middle–school students, which is still pervasive even at longer temporal spans. At syntactic level, this seems to be confirmed by the occurrence of dislocated dependencies (*dep_dislocated*) in the first position of the ranking derived by classifying the order of essays written at a distance of one year.

1 month distance	1 year distance	2 years distance
dep_punct	dep_dislocated	error_pronouns_omission
upos_AUX	xpos_PP	n_tokens
upos_X	xpos_BN	wfc-verbs-lemma
dep_aux	xpos_PD	n_prepositional_chains
upos_PUNCT	xpos_DD	aux_tense_Past
verbs_form_Part	aux_form_Fin	xpos_RI
dep_cop	error_preposition_omission_redundancy	obj_pre
upos_VERB	avg_lexical_errors	obj_post
verbs_form_Fin	error_vocabulary-erroneous-use	n_sentences
xpos_AP	n_sentences	dep_aux
dep_det:poss	dep_vocative	aux_1PerPl
xpos_SP	xpos_RI	error_pronouns_redundancy
dep_conj	wfc-nouns-lemma	verbs_num_pers_2PerSing
wfc-adjectives-lemma	n_prepositional_chains	aux_form_Ger
wfc-nouns-word	n_tokens	xpos_FB
verbs_3PerSing	aux_tense_Imp	dep_conj
dep_root	verb_edges_3	aux_tense_Imp
dep_acl:relcl	error_conjunctions-misuse	wfc-adjectives-word
dep_nsubj	error_full-stop-omission	error_monosyllabes-misuse-*po'*
dep_advcl	avg_punctuation_errors	wfc-nouns-word

Table 5.13: *Ranking of the first 20 features for three different temporal spans.*

According to the UD annotation tagset, this syntactic relation has the specific function of indicating fronted or postposed elements that do not fulfill the usual core grammatical relations of a sentence, which is quite typical in speech. In addition to these features, what helped more the classifier in the same classification scenario is the different use of functional categories by students, specifically pronouns (*xpos_PP*, *xpos_PD*), negative adverb (*xpos_BN*) and determiners (*xpos_DD*, *xpos_RI*).

Beyond error-related features, morpho-syntactic information still has a relevant role in classifying essays when the longest temporal span is considered. However, in this case, features related to verbal inflectional morphology (tense, mood and person) are more highly ranked than those concerning the distribution of core grammatical categories (see, e.g. *aux_tense_Past, aux_form_Ger, aux_tense_Imp*). This is in line with what we observed in the linguistic profiling section (Table 5.6), where differences concerning the use of verbal features in the two years were found to be statistically significant. Interestingly, with the exception of the *words frequency class*, lexical features do not seem to be particularly relevant and this allows us to confirm what already reported in [Barbagli, 2016], namely that vocabulary distribution, lexical density and TTR (*Type Token Ratio*) do not change significantly over the two school years.

Urban area	1 month distance	Two years distance
Center	0.579	0.629
Suburbs	0.513	0.670

Table 5.14: *C_m values according to the two urban areas.*

5.1.6 Investigating relationships between writing competence and background information

The last part of this study presents the first results of a pilot study that we performed in order to explore the hypothesis put forth in [Barbagli, 2016] that there could a relationship between the observed trends in the evolution of writing competence and the school environment of the students. This information was explicitly collected as one of the background variable of each student included in the corpus.

To this end, we inspect again the classification results by computing the confidence of our model (C_m), i.e. the measure that, as mentioned in Sec. 5.1.1, depicts the uncertainty of the classifier estimates. In particular, C_m can be defined as the variation between the two probabilities assigned by our classifier to each labels (1 if $t(d_j) > t(d_i)$, 0 otherwise). On the assumption that the more confident the model in predicting the chronological order of essays written by a given student, the easier is the classification task for that student, we can state that higher C_m values could are indicative of a greater evolution in student's writing competence. On the contrary, if we consider essays for which our classifier is less confident with its predictions, we can infer that the two essays do not present noticeable variations in their linguistic profile, although they were written in two different periods.

Specifically, we performed an experiment by computing the C_m values of our classifier for two different temporal spans (*1 month distance* and *Two years distance*) and then dividing the students according to the two different areas of Rome: historical center and the suburbs. As we can see in Table 5.14 there is no particular difference between the results. However, as the temporal span increases the C_m values for both urban areas show a slight improvement, in particular for the students of the suburban schools. This allows us, partly, to confirm that the evolution of writing competence is more evident for those students attending schools in suburbs, possibly because their entry level is lower, as suggested by the answers obtained from the questionnaires.

CHAPTER 6

Tracking the Evolution of Written Language Competence in L2 Spanish Learners

Starting from the experiments devised in the previous chapter, we present a complementary study in which we applied the same NLP stylometric approach to study the evolution of writing competence of L2 learners of Spanish.

6.1 Introduction

After having investigated the possibility of tracking the evolution of written competence of L1 learners, we decided to extend our methodology in the context of L2 written language development. In particular, we proposed a study aimed at modeling writing development in learners of Spanish as a second and Heritage language. We decomposed the problem into two main research questions: *(i)* verify if it is possible to predict the relative order of two essays written by the same student at different course levels using a wide set of linguistic predictors automatically extracted from Spanish L2 written productions; *(ii)* understand which typologies of language phenomena contribute more to the identification of writing skills' evolution and whether such properties reflect the teaching guidelines of the courses.

Following the approach devised in [Richter et al., 2015] and in the previous experiments we addressed the first research question as a classification task: given a pair of essays written by the same student and ordered according to the course level (d_1, d_2), we classify whether $C(d_2) > C(d_1)$, where $C(d_1)$ and $C(d_2)$ correspond respectively to the course levels during which the student wrote d_1 and d_2. Specifically, we model

Course Level	**Essays**	**Tokens**	**Students**
Beginner (SPA 1-3)	2,058	485,435	1,130
Intermediate (SPA 21-22)	445	120,102	244
Composition (SPA 23-24)	536	151,197	287
Heritage (SPA 31-33)	459	130,684	244
Total	3,498	887,418	1,905

Table 6.1: *Summary of corpus composition.*

Terms Enrolled	**Students**	**Essays**	**Tokens**
2	267	984	290,399
3	111	612	179,306
4	32	242	74,956
5	5	48	13,977

Table 6.2: *Longitudinal data summary.*

the problem as a binary classification task, training a Linear Support Vector Machine (LinearSVM) to predict the relative order of two essays written by the same student using our set of linguistic predictors automatically extracted from the POS tagged and dependency parsed essays.

We further extracted and ranked the feature weights assigned by the linear model in order to understand which typology of linguistic features contributes more to the classification task at different course levels. The assumption is that the higher the weight associated with a specific feature, the greater its importance in solving the classification task and, consequently, in modeling the student's written language evolution.

The contributions of this study are as follows:

- We present, to the best of our knowledge, the first data–driven study which uses linguistic features from student data to model the evolution of written language competence in Spanish as a Second Language (SSL);
- We show that it is possible to automatically predict the relative order of two essays written by the same student at different course levels using a wide spectrum of linguistic features;
- We investigate the importance of linguistic features in predicting language growth at different course levels and whether they reflect the explicit instruction that students receive during each course.

6.1.1 The COWS-L2H Corpus

We analyzed development of student writing from the *Corpus of Written Spanish of L2 and Heritage Speakers*, or COWS-L2H [Davidson et al., 2020]. This corpus consists of 3,498 short essays written by students enrolled in one of ten lower-division Spanish courses at a single American university. Concretely, these courses are organized as follows: Spanish (SPA) 1, 2, and 3 are the introductory courses, which exposes students to the basic morphosyntax of Spanish; SPA 21 and 22 are the intermediate courses, focused on the development of reading and listening skills with a strong emphasis on lexical development; SPA 23 and 24 are two courses that specifically aim at improving writing skills with an emphasis on academic writing in Spanish; SPA 31, 32, and 33 are the Heritage speakers courses. These courses are grouped into four categories based on student proficiency and experience, as shown in Table 6.1.

Student compositions in the corpus are written in response to one of four writing prompts, which are changed periodically. During each period (an academic quarter, which consists of ten weeks of instruction) of data collection, students are asked to submit two compositions, approximately one month apart, in response to targeted writing prompts. These composition themes are designed to be relatively broad, to allow for a wide degree of creative liberty and open-ended interpretation by the writer. Prompts are intended to be accessible to writers at all levels of proficiency. Additionally, the use of broad themes invites the use of a variety of verb tenses and vocabulary. The use of specific writing prompts allows us to control for known topic effects on syntactic complexity among L2 learners [Yang et al., 2015].

The essays in the corpus were submitted by 1,370 unique student participants, with 415 student participants having submitted compositions in two or more academic terms (for a maximum of eight writing samples from each student). Thus, the corpus contains both cross-sectional and longitudinal data on the development of student writing in the context of a university language program. The distribution of the essays across the levels is uneven due to the distribution of student enrollment in Spanish courses. Because more students enroll in beginning Spanish courses than in advanced levels, a larger number of essays submitted to the corpus come from these beginner-level courses. The L2 Spanish learners are primarily L1 speakers of English, but due to the diverse student population of the source university, a large number are L1 speakers of other languages such as Mandarin. However, as English is the university's language of instruction, all students are either L1 or fluent L2 speakers of English. Those students enrolled in the Heritage courses (SPA 31 - 33) are, for the most part, L1 speakers of Spanish, having learned Spanish from a young age in the home, and L2 speakers of English; these Heritage learners have had little-to-no academic instruction in Spanish.

We focused our study on the longitudinal data in the COWS-L2H corpus. We were thus able to model the chronological development of L2 Spanish writing by monitoring how the writing quality of an individual student's compositions increase with time. Student participation is summarized in Table 6.2.

6.1.2 Linguistic Features

The set of linguistic features considered as predictors of L2 written competence evolution is based on those described in [Brunato et al., 2020] already used in the experiments of the *CItA* corpus. Moreover, since it is acknowledged that lexical proficiency plays an

Features	SPA 1	SPA 2	SPA 3	SPA 21	SPA 22	SPA 23	SPA 24	SPA 31	SPA 32	SPA 33
Raw Text Properties										
char_per_tok	4.3	4.4	4.42	4.42	4.43	4.46	4.41	4.42	4.42	4.38
n_sentences	20.0	24.01	23.57	20.8	20.17	19.54	17.92	16.06	16.31	15.46
tokens_per_sent	10.7	13.16	13.74	15.71	16.43	17.11	19.01	19.95	20.07	20.94
Morphosyntactic information										
lexical_density	.51	.5	.5	.49	.48	.48	.47	.48	.47	.47
upos_ADJ	.07	.06	.06	.06	.05	.05	.05	.05	.05	.05
upos_ADP	.09	.1	.11	.11	.11	.12	.12	.13	.12	.13
upos_NOUN	.16	.16	.16	.16	.16	.17	.17	.17	.16	.16
upos_PRON	.07	.07	.07	.07	.07	.07	.07	.07	.08	.08
upos_PUNCT	.14	.13	.12	.12	.11	.11	.11	.09	.09	.09
upos_SCONJ	.01	.02	.03	.03	.04	.04	.04	.04	.05	.05
upos_VERB	.12	.12	.12	.12	.12	.12	.12	.13	.13	.13
Inflectional morphology										
aux_mood_Cnd	.02	.03	.04	.03	.06	.05	.04	.05	.06	.04
aux_mood_Ind	.97	.96	.92	.94	.91	.92	.94	.91	.91	.93
aux_mood_Sub	.01	.01	.03	.02	.03	.02	.03	.03	.03	.03
aux_tense_Imp	.05	.16	.21	.21	.24	.24	.22	.23	.2	.24
aux_tense_Past	.02	.1	.09	.12	.12	.11	.12	.11	.12	.11
aux_tense_Pres	.92	.73	.69	.65	.63	.65	.66	.63	.66	.63
verbs_tense_Imp	.02	.08	.11	.13	.16	.14	.13	.17	.15	.14
verbs_tense_Past	.11	.28	.28	.3	.35	.3	.31	.31	.28	.33
Verbal Predicate Structure										
verb_edges	2.3	2.5	2.52	2.62	2.67	2.63	2.7	2.71	2.68	2.76
verb_edges_4	.09	.13	.13	.16	.16	.15	.16	.16	.16	.16
verbal_head_sent	1.52	1.8	1.92	2.13	2.26	2.3	2.54	2.73	2.86	2.95
Global and Local Parsed Tree Structures										
parse_depth	2.88	3.27	3.37	3.6	3.78	3.94	4.21	4.49	4.59	4.56
max_links_len	.65	.7	.72	.96	.92	.99	1.2	1.24	1.21	1.39
token_per_clause	7.17	7.49	7.28	7.52	7.41	7.55	7.62	7.42	7.16	7.26
Order of elements										
obj_post	.67	.68	.67	.64	.65	.69	.69	.6	.64	.6
obj_pre	.33	.32	.33	.35	.35	.31	.31	.39	.36	.4
subj_pre	.8	.84	.82	.84	.84	.84	.83	.81	.78	.79
Use of Subordination										
subord_chain_len	1.06	1.15	1.18	1.21	1.24	1.24	1.26	1.29	1.33	1.32
subord_2	.08	.11	.13	.15	.17	.17	.18	.19	.2	.2
subord_dist	.24	.33	.38	.4	.44	.47	.5	.56	.58	.57

Table 6.3: *A subset of linguistic features extracted for each course level. For each feature it is reported its average value.*

important role in predicting L2 writing development [Crossley and McNamara, 2012], we also added as training features the *word frequency class* for each word form/lemma was computed exploiting the Spanish Wikipedia (dump of March 2020) using the measures defined in Sec. 5.1.4.

A first overview of how and to what extent all these features vary across the documents of the COWS-L2H Corpus is provided in Table 6.3. Essays written by students in the first course levels are longer in terms of number of sentences but they contain shorter sentences compared with those written in the more advanced courses. As concerns the distribution of POS, essays written in the first years show a lower percentage of e.g. adpositions (*upos_ADP*) and subordinate conjunctions (*upos_SCONJ*) typically contained in longer and well-articulated sentences, while the use of main content words (e.g. *upos_NOUN*, *upos_VERB*) is almost comparable across years. The variation affecting morphosyntactic categories is reflected by the lexical density value, i.e. the ratio between content words over the total number of words, which is slightly higher in beginner essays. If we focus on differences concerning verbal morphology, a linguistic property particularly relevant in the development of Spanish curriculum, we can see how the use of more complex verb forms increases across course levels. Essays of the introductory courses contain a lower percentage of verbs in the past (*verbs_tense_Past*) and imperfect tenses (*verbs_tense_Imp*) (out of the total number of verb tenses) as well as a lower percentage of auxiliary verbs (*aux_**) typically used in more complex verb forms, such as copulative verbs or periphrastic moods and tenses. Interestingly, features related to verb inflectional morphology have the highest standard deviation, suggesting a quite wide variability among learners. A similar trend towards the acquisition of more complex verb structures can also be inferred by considering features extracted from the syntactic level of annotation: essays of the intermediate courses contain for example sentences with a higher average number of dependents of verbs (*verb_edges*) and in particular of verbs with a complex argument structures of 4 dependents (*verb_edges_4*).

As long as Spanish learners start mastering the second language, linguistic properties related to the construction of more complex sentences increase. This is for example the case of the depth of sentence tree (*parse_depth*) and of the length of syntactic relations (*max_links_len*) as well as of features concerning the use of subordination.

6.1.3 Experiments

We train a LinearSVM that takes as input pairs of essays written by the same students according to all the possible pairs of course levels (e.g. SPA 1 - SPA 2, SPA 2 - SPA 3, etc.). Specifically, we extract for each pair the linguistic features corresponding to the first and second essays and the difference between them. We standardize the input features by scaling each component in the range $[0, 1]$. To test the actual efficiency of the model, we perform the experiments with a 5-cross validation using different students during the training and testing phases. In order to provide our system with negative samples, we expand our datasets by adding reversed samples.

Since the students were asked to write essays responding to different prompts, we devise two set of experiments, pairing all the essays written by the same students that have: (i) the same prompt; (ii) both same and different prompts. Also, because of the small number of training samples for certain pairs of course levels we also decide to perform the experiments on a sentence-level, extracting the linguistic features for each sentence in the longitudinal subset of the COWS-L2H corpus and pairing them on the basis of the previously defined criteria. In order to obtain reliable results both on the document and sentence configurations, we consider only datasets at different pairs of course levels that contain at least 50 and 20 samples (including negative pairs) respectively. All the classification experiments are performed using the majority class classifier as baseline and accuracy as the evaluation metric.

Tracking Writing Skills' Evolution

Table 6.4 reports the results obtained at both the document and sentence levels, pairing essays that have the same prompt (*Same* columns) and both the same and different prompts (*All* columns). As a general remark, we observe that best results are those obtained with the document-level experiments. This is quite expected, since sentence-

Course Levels	Documents				Sentences			
	Same		All		Same		All	
	Score	Samples	Score	Samples	Score	Samples	Score	Samples
All Levels	0.68	2,208	0.7	5,536	0.59	1,047,156	0.61	2,570,366
SPA 1 - SPA 2	0.88	280	0.9	624	0.7	143,660	0.71	316,264
SPA 1 - SPA 3	0.97	178	0.95	440	0.75	85,032	0.75	209,048
SPA 1 - SPA 21	#	#	0.91	116	0.61	14,298	0.7	46,738
SPA 2 - SPA 3	0.62	528	0.62	1,192	0.56	323,332	0.56	724,400
SPA 2 - SPA 21	0.61	62	0.61	188	0.57	35,754	0.58	104,442
SPA 2 - SPA 22	#	#	0.59	68	0.55	8,048	0.63	29,670
SPA 2 - SPA 23	#	#	0.77	52	#	#	0.58	27,420
SPA 3 - SPA 21	0.59	158	0.55	364	0.53	82,104	0.54	190,596
SPA 3 - SPA 22	0.61	64	0.58	186	0.54	31,886	0.6	93,486
SPA 3 - SPA 23	#	#	0.89	106	0.59	13,404	0.59	45,804
SPA 3 - SPA 24	#	#	#	#	#	#	0.68	11,276
SPA 21 - SPA 22	0.59	132	0.62	302	0.52	57,326	0.54	132,454
SPA 21 - SPA 23	0.52	58	0.74	154	0.54	27,038	0.57	67,634
SPA 21 - SPA 24	#	#	0.7	92	0.47	9,268	0.56	35,384
SPA 22 - SPA 23	0.71	76	0.69	186	0.55	35,272	0.56	79,168
SPA 22 - SPA 24	0.69	158	0.73	164	0.5	23,446	0.56	66,184
SPA 23 - SPA 24	0.45	168	0.49	386	0.48	61,654	0.49	137,786
SPA 31 - SPA 32	0.8	100	0.63	212	0.55	27,608	0.55	57,790
SPA 31 - SPA 33	0.52	100	0.53	198	0.51	24,830	0.48	48,990
SPA 32 - SPA 33	0.54	96	0.59	256	0.5	24,154	0.55	66,466

Table 6.4: *Classification results in terms of accuracy obtained both at document and sentence levels along with number of samples for each dataset.* ***Same*** *and* ***All*** *columns report the results obtained by pairing essays that have same prompt and both same and different prompts respectively. Since the labels within each dataset has been balanced, baseline accuracy is 0.50.*

level classification is a more complex task that often requires a higher number of features to gain comparable accuracy [Dell'Orletta et al., 2014]. If we focus instead on the distinction between *Same* and *All* results, we notice that higher scores are mainly achieved considering pairs of essays that also have different prompts. Again, this result is not surprising because adding pairs of essays with different prompts within each datasets increases the number of training samples, thus leading to better scores. Despite this, the results obtained according to the *Same* and *All* configurations are quite similar and this allows us to confirm that classification accuracy is not significantly harmed if the two essay's prompts are the same, thus showing that our system is actually focusing on written language competence evolution properties rather than prompt-dependent characteristics.

More interestingly, we notice that considering all the possible course level pairs at the same time our system is able to achieve quite good results, especially at document level classification (0.68 and 0.70 of accuracy for *Same* and *All* configurations respectively), thus showing that it is possible to automatically predict the chronological order of two essays written by the same student by using a wide spectrum of linguistic properties.

In general, our best scores are obtained by considering all the experiments that include essays written by students in the Beginner category (SPA 1, 2 and 3). This is

SPA 1 - SPA 2	SPA 1 - SPA 3	SPA 2 - SPA 3	SPA 3 - SPA 21	SPA 22 - SPA 23	SPA 31 - SPA 32
aux_mood_Ind	lexical_density *	aux_tense_dist_Pres *	lexical_density	upos_PUNCT	upos_ADP *
aux_tense_Pres *	upos_ADP *	aux_mood_Ind	upos_DET	dep_punct	dep_case *
aux_tense_Imp *	upos_VERB *	aux_tense_Imp *	dep_punct	upos_ADV	verbal_head_sent
aux_tense_Past *	upos_NOUN *	aux_tense_Past	upos_VERB	dep_advmod	upos_PUNCT
upos_ADP *	upos_ADJ	dep_punct *	aux_tense_Pres	upos_CCONJ	upos_PRON
verbs_tense_Past *	upos_PRON	upos_PUNCT *	upos_ADJ	dep_cc *	dep_mark
upos_VERB *	dep_det	dep_nsubj *	upos_NOUN	upos_VERB	dep_punct
upos_INTJ *	upos_PUNCT *	dep_iobj	dep_nsubj *	dep_case	aux_tense_Imp
verbal_head_sent *	upos_PROPN	upos_PRON	upos_PRON	aux_form_Part	verbs_tense_Pres
verbs_tense_Imp *	dep_case *	verbal_head_sent *	upos_SCONJ	upos_ADP	subord_dist
upos_ADJ *	upos_SCONJ *	dep_cop	upos_ADV *	dep_mark	dep_cop
ttr_form	upos_AUX	subj_post *	upos_PUNCT	dep_compound	dep_cc
upos_PRON *	dep_punct *	aux_form_Fin	aux_form_Fin	upos_INTJ *	lexical_density
upos_PROPN *	subord_dist *	verbs_tense_Imp *	dep_cc *	dep_nsubj *	upos_AUX
upos_PUNCT *	upos_CCONJ *	upos_AUX	aux_tense_Imp	upos_AUX	upos_ADV

Table 6.5: *Feature rankings obtained with sentence-level (Same) classification results for six different course level pairs. Features that vary in a statistically significant way with Wilcoxon Rank-Sum test are marked with *.*

particularly evident for the experiments that compare essays written during SPA 1 as one of the two considered course levels, most likely because the evolution from knowing nothing at all of a specific L2 to knowing enough to start writing is actually bigger that the difference between knowing a little and then learning a little more. Additionally, students at this beginning stage of L2 acquisition tend to use markedly fewer words per sentence, and the words they user are shorter; these features are particularly salient for the classifier. Observing instead the results obtained pairing student essays belonging to the other three course level categories (Intermediate, Composition and Heritage), we notice a considerable drop in classifier performance. For instance, if we compare essays written by students in the Composition category (SPA 23 - SPA 24) we can see that all the classification results are below the majority class baseline classifier. A possible reason might be that these two courses are specifically aimed at improving learners' writing skills, with an emphasis on academic writing in Spanish, thus involving specific properties, such as discourse-level characteristics, which are possibly not covered by our set of features.

Understanding Linguistic Predictors

Beyond classification results, we were interested in understanding which typologies of linguistic phenomena are more important for solving the classification task and whether such properties correlate to the teaching curriculum. To better explore this second research question, we perform a feature ranking analysis along with the classification experiments, which allows us to establish a ranking of the most important features according to the different classification scenarios. That is, we evaluate the importance of each linguistic property by extracting and ranking the feature weights assigned by the LinearSVM. Table 6.5 reports the feature rankings obtained with sentence-level classification results, including pairs of essays that have the same prompt (*Same* configuration). We considered in particular six different course level pairs which are mostly representative of different stages of writing development. The focus on sentence-level results rather than document-level allows capturing more fine-grained linguistic phenomena.

Because the COWS-L2H corpus was collected from a single university with set curriculum, we are able to compare the features utilized by the LinearSVM with the course curriculum. We find that the feature rankings as obtained from the LinearSVM can in many cases be explained by differences in curriculum at each level. For example, from SPA 1 to SPA 2 the most important features used by the model are all related to verbal morphology, particularly morphology of auxiliary verbs. This can be explained by the fact that SPA 1 and 2 are the courses where students are introduced for the first time to the notions of verb tense and person. SPA 1 is focused on managing the idea of person and number in a tense that is not particularly difficult to understand for a speaker of English: the present tense. SPA 2, however, introduces the difficult difference between the three tenses in the past: imperfect, preterite and plus-perfect. This fact explains why distribution of past tense main verbs (*verbs_tense_Past*) differs between essays written during SPA 1 and SPA 2. Additionally, SPA 2 introduces composed verb tenses that require an auxiliary. Specifically, the auxiliary verbs “haber”, “estar”, and “ser” are introduced in SPA 2 as part of the past tense forms. Thus, it is not surprising that the top four features used by our classifier for distinguishing between essays written in SPA 1 and SPA 2 are related to the use of auxiliary verbs.

Classification of essays written by students while enrolled in SPA 2 and SPA 3 also relies largely on differences in verbal morphology. While the distribution of present tense auxiliary verbs is the most important distinguishing feature, other compound verb tenses play a role at these levels. For example, differences in the distribution of imperfect auxiliary verbs (*aux_tense_Imp*) may be explained by the use of the pluperfect tense.

Between SPA 1 and SPA 3, the most important discriminating feature is lexical density. While there is no specific focus on lexical density in the course curriculum, this feature is a natural extension of increasing sentence complexity. [Davidson et al., 2019] shows that as students progress through the Spanish course sequence, lexical density tends to decrease due to the increased use of function words in more complex sentences. Additionally, one of the final items covered in the SPA 1 curriculum is the use of the prepositions “por” and “para”. Also, at all three beginning levels students are taught to use prepositions in constructing more complex sentence structures. This may explain why preposition usage (*upos_ADP*) is a key discriminating feature between essays written in SPA 1 and SPA2, as well as between SPA 1 and SPA 3. The prominence of this feature indicates that students are learning to more confidently use prepositions as their writing skills develop. The fact that (*upos_ADP*) is not a key discriminating feature between SPA 2 and SPA3 indicates that these changes are occurring primarily at the SPA 2 level, which accords with the course curriculum.

In spite of the still reasonable accuracy in discriminating more advanced levels, making a direct connection between the features used by the SVM and the course curriculum becomes more difficult. At these more advanced levels students have developed an individual writing style which results in a more complex relationship between the curriculum and the syntax used by students. At the SPA 3 - SPA 21 interval, the only three features which vary in a statistically significant way are the distributions of nominal subjects (*dep_nsubj*), adverbs (*upos_ADV*), and coordinating conjunctions (*dep_cc*). While the increased use of adverbs may be seen as a general sign of increased writing complexity, coordinating conjunctions are taught explicitly during SPA 3. Conjunctions are also practiced intensively during both SPA 21 and SPA 22 explaining their importance as a discriminating feature between these levels.

One of the clearest connections between curriculum and the features used by the LinearSVM occurs at the Heritage levels SPA 31 and SPA 32. Heritage learners of Spanish raised in an English-dominant country are known to use "English-like" prepositions in Spanish. For example, [Pascual y Cabo and Soler, 2015] report on preposition stranding (which is grammatical in English by ungrammatical in Spanish) among Heritage speakers of Spanish in the United States. We find that distributional differences in the use of prepositions, represented by the features *upos_ADP* and *dep_case*, is the key distinguishing feature between essays written by the same student during SPA 31 and SPA 32. This difference indicates that students are learning to use prepositions in a more "Spanish-like" manner, which is one of the major areas of feedback which instructors provide to Heritage students.

CHAPTER 7

Discussion and Future Directions

Relying on two learner corpora, our studies examined the potential of a NLP-based stylometric approach to identify relevant transformation occurring in L1 and L2 learners' writing. In particular, the longitudinal nature of the two corpora allowed us to track the evolution of the written competence in Italian L1 and Spanish L2 students, as well as to identify which linguistic features are more predictive of this evolution and how they change according to the considered temporal span.

As regards our first study, the classification results obtained in the three experiments have demonstrated that linguistic features automatically extracted from text not only allow making explicit the relevant transformations occurring in L1 learners' writing competence but can be exploited as effective predictors in the automatic classification of the chronological order of essays written by the same student, especially at more distant temporal spans. Moreover, by testing our approach on a cross-prompt scenario, we show that the considered features capture markers of language evolution which are not related to the textual typology of the essay.

When training our model using also the twenty–six features related to error annotations, we obtained a general improvement in almost all cases. These results demonstrate that analyzing the diverse typologies of errors made by students in their texts is effective to capture aspects of the written language competence evolution. In this regards, we also noticed that the errors which allow the classifier to achieve a better accuracy are the grammatical ones. This could be due both to the larger amount of errors of this category (46.41% and 48.7% of the total in the first and second school year) and by the fact that grammatical errors, as well as orthographic errors, have a significant variation over the two school years, thus probably allow the classifier to obtain better results.

Regarding the second research question, extracting the feature weights assigned by the linear model we were able to establish a ranking of the most important features according to different temporal spans. Changes of the resulting rankings in the different classification scenarios suggest that both linguistic and error-related features contribute in a different way according to time intervals. For instance, it was shown that features related to the error annotation acquire much more relevance as the temporal span increases, and this allows us to confirm that the errors made by the students are an indicative proxy to track the writing competence evolution, especially in the transition from the first to the second year. In a similar fashion, we observed that the classifier is sensitive to changes affecting morpho–syntactic features, especially those related to the use of grammatical categories and to the inflectional properties of verbs: the latter were also found to change in a significant way when comparing the whole subcorpus of essays written in the first and in the second year. This gives additional evidence that mastering verbal morphology in a morphologically-rich language like Italian is an important skill that evolves in writing during the considered school years. This is also in line with the [Weiss and Meurers, 2019] study on German cross-sectional data, which showed that features belonging to morphological complexity play an important role especially in the development of secondary school writing. However, unlike [Weiss and Meurers, 2019] and [Kerz et al., 2020], our analysis showed that features related to lexical sophistication do not seem to be particularly relevant for identifying the evolution of writing competence.

Lastly, we presented a pilot study in which we try to explore the relationships between the developmental patterns in writing and information about students background variables. The obtained results suggested that the student's learning curve varies according to the geographical area where the school is located. In fact, we saw that, when a higher temporal span is considered (e.g. *Two years distance*), the classifier is more confident about its decision for essays written by students who belong to suburban schools. Although preliminary, these results go in the direction of what suggested in [Barbagli, 2016], namely that the evolution of writing skills is strictly related to the socio-cultural context inferred from background variables, and that these aspects affect the linguistic entry level of the students.

In our second study, we demonstrated that our stylometric approach can be also applied in the contest of L1 learners. In fact, we demonstrated that it is possible to automatically predict the relative order of two essays written by the same student at different course levels using our set of linguistic features, especially when considering students enrolled in beginner-level Spanish courses. Moreover, we have shown that the linguistic features most important in predicting essay order often reflect the explicit instruction that students receive during each course.

These works can help instructors and language researchers better understand the specific linguistic factors which contribute to improved writing proficiency. Additionally, the appearance of features in the LinearSVM ranking helps clarify the effect of instruction on writing performance, specifically on effects such as the known delay between students being taught a concept and that concept appearing in the students' writing. We also believe that this work may contribute to the development of better language assessment and placement tools. Moreover, future works works could build upon these findings and investigate more in-depth the influence of student L1 on feature rankings, as L1 (and L2) transfer and interference effects may influence the rate at which students acquire specific linguistic features. Additionally, a possible direction would be to conduct cross-lingual experiments, investigating how the feature rankings of e.g. Spanish writing development relate to those seen in the acquisition of other languages.

To conclude, we would like to draw attention to some other perspectives that the presented studies could enable, which are especially relevant in the field of NLP-based educational applications. Finding theoretically motivated methods to monitor the learning growth of each student can support the assessment process by teachers, which could be a very demanding task in distance learning paradigms. Similarly, we believe that the new educational framework poses new challenges concerning students' engagement in virtual classes. As shown by [Slater et al., 2017], a variety of linguistic features identified by NLP tools can be used as reliable predictors of affective states experienced by students, such as boredom, confusion, frustration, engaged concentration. With this respect, it would be interesting to explore potential correlations between the motivation and level of engagement shown by students and the linguistic properties turned out to be involved in modeling language learning.

Last but not least, the proposed approach can promote comparative studies on the evolution of the written language competence from a cross-linguistic perspective. In fact, one of the main novelties of the proposed approach is that the linguistic features used as predictors of language learning were extracted from corpora annotated according to the Universal Dependencies (UD) framework. Since this annotation is inspired to 'universal' principles aiming at annotating in a consistent way similar constructions across languages, the process of feature extraction can be applicable to other learner corpora for all languages included in the UD project.

Part III

Interpreting Neural Language Models

CHAPTER 8

Testing for Linguistic Competence

As we mentioned in Chapter 1, starting from the hypothesis that NLP methods developed to study the process of written language evolution could be used to interpret the linguistic knowledge encoded by NLMs, we exploit several approaches to investigate the implicit knowledge encoded by these models and how this knowledge is affected (and employed) after a fine-tuning process. In particular, in this chapter we focus on the the studies we have conducted on the basis of the so-called *probing classifier* paradigm, with the aim of understanding which are linguistic properties that are implicitly learned within the internal representations of Transformer-based models.

8.1 Introduction

Approaches based on probing classifiers have become one of the most prominent methodologies for interpreting and analyzing deep neural network models of NLP. As showed in Chapter 4, the approach is quite simple (a classifier is trained to predict some linguistic property from a model's representations) and has been used to investigate a variety of models and language phenomena. In this section we will focus on the experiments we devised to test the linguistic competence of several NLMs relying on a wide range of probing tasks. In particular, we adopted an approach inspired to the 'linguistic profiling' methodology put forth by [van Halteren, 2004], which assumes that wide counts of linguistic features automatically extracted from parsed corpora allow modeling a specific language variety and detecting how it changes with respect to other varieties, e.g. complex vs simple language, female vs male–authored texts, texts written in the same L2 language by authors with different L1 languages. Particularly relevant for our study, is that multi-level linguistic features, as we already showed in the experiments of Chapters 5 and 6, have been shown to have a highly predictive role in tracking the evolution of learners' linguistic competence across time and developmental levels, both in first and second language acquisition scenarios [Lubetich and Sagae, 2014, Miaschi et al., 2020b]. Given the strong informative power of these features to encode a variety of language phenomena across stages of acquisition, we assume that they can be also helpful to dig into the issues of interpretability of NLMs.

The rest of the section is organized as follows. Sec. 8.2 discusses a study aimed at investigating the linguistic knowledge implicitly encoded by BERT internal representations before and after a fine-tuning process and how this knowledge affects its ability on a downstream task. We then focus on an in-depth study aimed at understanding the linguistic competence encoded in a contextual (BERT) and a contextual-independent (word2vec) model in Sec. 8.3. Sec. 8.4 investigates the relationship between linguistic knowledge encoded by BERT and the number of individual units involved in the encoding of such knowledge. Sec. 8.5 presents a comparison between the probing performances of 7 Italian NLMs over multiple linguistic feature categories and according to different architectures of probing models and textual genres. Finally, Sec. 8.6 introduces a methodology to test the reliability of probing tasks by building control tasks at increasing level of complexity for an Italian Transformer model.

8.2 Linguistic Profiling of a Neural Language Model

In this study, we extended prior work by studying the linguistic properties encoded by one of the most prominent NLM, BERT [Devlin et al., 2019], and how these properties affect its predictions when solving a specific downstream task. We defined three research questions aimed at understanding: (i) what kind of linguistic properties are already encoded in a pre-trained version of BERT and where across its 12 layers; (ii) how the knowledge of these properties is modified after a fine-tuning process; (iii) whether this implicit knowledge affects the ability of the model to solve a specific downstream task, i.e. Native Language Identification (NLI). To tackle the first two questions, we adopted the 'linguistic profiling' methodology defined above. In particular, we investigated whether the features successfully exploited to model the evolution of language competence can be similarly helpful in profiling how the implicit linguistic knowledge of a NLM changes across layers and before and after tuning on a specific downstream task. We chose the NLI task, i.e. the task of automatically classifying the L1 of a writer based on his/her language production in a learned language [Malmasi et al., 2017]. As shown by [Cimino et al., 2018], linguistic features play a very important role when NLI is tackled as a sentence–classification task rather than as a traditional document–classification task. This is the reason why we considered the sentence-level NLI classification as a task particularly suitable for probing the NLM linguistic knowledge. Finally, we investigated whether and which linguistic information encoded by BERT is involved in discriminating the sentences correctly or incorrectly classified by the fine-tuned models. To this end, we tried to understand if the linguistic knowledge that the model has of a sentence affects the ability to solve a specific downstream task involving that sentence.

The contributions of this work are as follows:

- we carried out an in-depth linguistic profiling of BERT's internal representations;
- we showed that contextualized representations tend to lose their precision in encoding a wide range of linguistic properties after a fine-tuning process;
- we showed that the linguistic knowledge stored in the contextualized representations of BERT positively affects its ability to solve NLI downstream tasks: the more BERT stores information about these features, the higher will be its capacity of predicting the correct label.

8.2.1 Our Approach

To probe the linguistic knowledge encoded by BERT and understand how it affects its predictions in several classification problems, we relied on a suite of 68 probing tasks, each of which corresponds to a distinct feature capturing lexical, morpho–syntactic and syntactic properties of a sentence. Specifically, we defined three sets of experiments. The first consisted in probing the linguistic information learned by a pre-trained version of BERT (BERT-base, cased) using gold sentences annotated according to the Universal Dependencies (UD) framework [Nivre et al., 2016]. In particular, we defined a probing model that uses BERT contextual representations for each sentence of the dataset and predicts the actual value of a given linguistic feature across the internal layers. The second set of experiments consisted in investigating variations in the encoded linguistic information between the pre-trained model and 10 different fine-tuned ones obtained training BERT on as many Native Language Identification (NLI) binary tasks. To do so, we performed again all probing tasks using the 10 fine-tuned models. For the last set of experiments, we investigated how the linguistic competence contained in the models affects the ability of BERT to solve the NLI downstream tasks.

Data We used two datasets: (i) the UD English treebank (version 2.4) for probing the linguistic information learned before and after a fine-tuning process; (ii) a dataset used for the NLI task, which is exploited both for fine-tuning BERT on the downstream task and for reproducing the probing tasks in the third set of experiments. The UD dataset includes three UD English treebanks: UD_English-ParTUT, a conversion of a multilingual parallel treebank consisting of a variety of text genres, including talks, legal texts and Wikipedia articles [Sanguinetti and Bosco, 2015a]; the Universal Dependencies version annotation from the GUM corpus [Zeldes, 2017]; the English Web Treebank (EWT), a gold standard universal dependencies corpus for English [Silveira et al., 2014]. Overall, the final dataset consists of 23,943 sentences.

As regards the second dataset, we used the 2017 NLI shared task dataset, i.e. the TOEFL11 corpus [Blanchard et al., 2013]. It contains test responses from 13,200 test takers (one essay and one spoken response transcription per test taker) and includes 11

Level of Annotation	Linguistic Feature	Label
Raw Text	**Raw Text Properties (*RawText*)**	
	Sentence Length	sent_length
	Word Length	char_per_tok
Vocabulary	**Vocabulary Richness (*Vocabulary*)**	
	Type/Token Ratio for words and lemmas	ttr_form, ttr_lemma
POS tagging	**Morphosyntactic information (*POS*)**	
	Distibution of UD and language–specific POS	upos_dist_*, xpos_dist_*
	Lexical density	lexical_density
	Inflectional morphology (*VerbInflection*)	
	Inflectional morphology of lexical verbs and auxiliaries	xpos_VB-VBD-VBP-VBZ, aux_*
Dependency Parsing	**Verbal Predicate Structure (*VerbPredicate*)**	
	Distribution of verbal heads and verbal roots	verbal_head_dist, verbal_root_perc
	Verb arity and distribution of verbs by arity	avg_verb_edges, verbal_arity_*
	Global and Local Parsed Tree Structures (*TreeStructure*)	
	Depth of the whole syntactic tree	parse_depth
	Average length of dependency links and of the longest link	avg_links_len, max_links_len
	Average length of prepositional chains and distribution by depth	avg_prep_chain_len, prep_dist_*
	Clause length	avg_token_per_clause
	Order of elements (*Order*)	
	Relative order of subject and object	subj_pre, obj_post
	Syntactic Relations (*SyntacticDep*)	
	Distribution of dependency relations	dep_dist_*
	Use of Subordination (*Subord*)	
	Distribution of subordinate and principal clauses	principal_prop_dist, subordinate_prop_dist
	Average length of subordination chains and distribution by depth	avg_subord_chain_len, subordinate_dist_1
	Relative order of subordinate clauses	subordinate_post

Table 8.1: *Linguistic Features used in the probing tasks.*

native languages (L1s) with 1,200 test takers per L1. We selected only written essays and we created pairwise subsets of essays written by Italian L1 native speakers and essays for all the other languages. At the end of this process, we obtained 10 datasets of 2,400 documents (33,756 sentences in average): 1,200 for the Italian L1 speakers and 1,200 for each of the other L1s included in the TOEFL11 corpus.

Probing Tasks and Linguistic Features Our experiments are based on the probing tasks approach defined in [Conneau et al., 2018], which aims to capture linguistic information from the representations learned by a NLM. In our study, each probing task consists in predicting the value of a specific linguistic feature automatically extracted from the parsed sentences in the NLI and UD datasets. The set of features is based on the ones described in [Brunato et al., 2020] and already tested for the experiments described in Chapters 5 and 6. As shown in Table 8.1, the considered features are intended to probe whether the NLMs encode in their representations 9 main aspects of the structure of a sentence. They range from quite simple aspects related to the knowledge of raw text properties (i.e. sentence and word length), to the vocabulary richness (in terms of type/token ratio), to morpho-syntactic and inflectional properties specific in particular of verbal predicates. More challenging probing features concerns the NLMs ability to encode complex aspects of sentence structure, including both global structure, such as the depth of the whole syntactic tree, and local features. We paid a specific attention to testing the models knowledge of the sub-trees of the nuclear elements of a sentence. In this respect we included a group of features modelling the verbal predicate structure, e.g. in terms of number of dependents of verbal heads, and a group referring to the order of subjects and objects with respect to their verbal head. In line with the focus

Level	BERT	Baseline
Raw text	0.68	0.52
Vocabulary	0.78	0.23
POS	0.68	0.27
Verb inflection	0.72	0.35
Verb predicate	0.60	0.48
Tree structure	0.78	0.70
Order	0.72	0.51
Syntactic dep	0.69	0.34
Subordination	0.71	0.48
All features	0.69	0.38

Table 8.2: *BERT ρ scores (average between layers) for all the linguistic features (AllFeatures) and for the 9 groups corresponding to different linguistic phenomena. Baseline scores are also reported.*

on specific sub-trees, we also considered a group of features capturing the use of subordination in terms of the distribution of subordinate clauses, of the internal structure of the subordinate clause sub–trees and of their relative order with respect to the main clause.

Models We relied on the pre–trained English version of BERT (BERT-base cased, 12 layers, 768 hidden units) for both the extraction of contextual embeddings and the fine-tuning process for the NLI downstream task. To obtain the embeddings representations for our sentence-level tasks we used for each of its 12 layers the activation of the first input token (*[CLS]*), which somehow summarizes the information from the actual tokens, as suggested in [Jawahar et al., 2019].

As mentioned above, each of our probing tasks consists in predicting the actual value of a given linguistic feature given the inner sentence representations learned by a NLM for each of its layers. Therefore, we used a linear Support Vector Regression (LinearSVR) as probing model.

8.2.2 Profiling BERT

Our first experiments investigated what kind of linguistic phenomena are encoded in a pre-trained version of BERT. To this end, for each of the 12 layers of the model (from input layer *-12* to output layer *-1*), we firstly represented each sentence in the UD dataset using the corresponding sentence embeddings according to the criterion defined in Sec. 8.2.1. We then performed for each sentence representation our set of 68 probing tasks using the LinearSVR model. Since most of our probing features are strongly correlated with sentence length, we compared the probing model results with the ones obtained with a baseline computed by measuring the Spearman's rank correlation coefficient (ρ) between the length of the UD dataset sentences and the corresponding probing values. The evaluation is performed with a 5-fold cross validation and using Spearman correlation (ρ) between predicted and gold labels as evaluation metric. Since the majority of probing experiments are based on classification tasks with a limited number of classes,

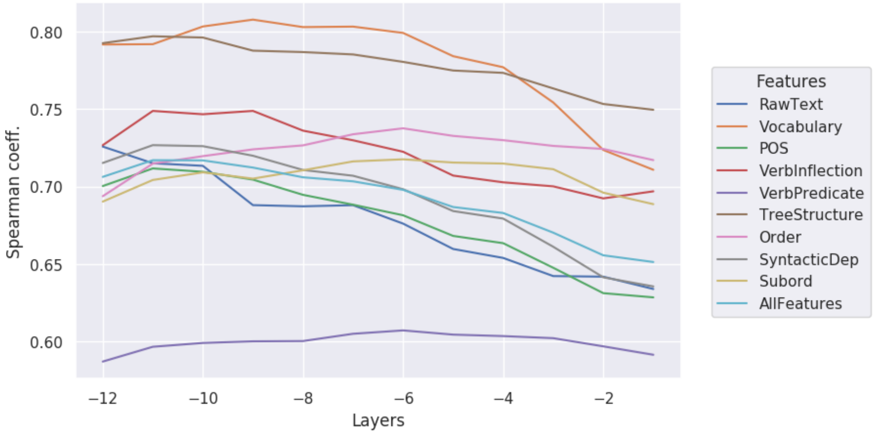


Figure 8.1: *BERT average layerwise ρ scores.*

we decided to rely on such metrics in our regression tasks since our interest was to check whether the model is able to capture the main differences and variations between the values assumed by our set of linguistic features.

As a first analysis, we probed BERT's linguistic competence with respect to the 9 groups of probing features. Table 8.2 reports BERT (average between layers) and baseline scores for all the linguistic features and for the 9 groups corresponding to different linguistic phenomena. As a general remark, we can notice that the scores obtained by BERT's internal representations always outperform the ones obtained with the correlation baseline. For both BERT and the baseline, the best results are obtained for groups including features highly sensitive to sentence length. For instance, this is the case of syntactic features capturing global aspects of sentence structure (*Tree structure*). However, differently from the baseline, the abstract representations of BERT are also very good at predicting features related to other linguistic information such as morpho-syntactic (*POS*, *Verb inflection*) and syntactic one, e.g. the structure of verbal predicate and the order of nuclear sentence elements (*Order*).

We then focused on how BERT's linguistic competence changes across layers. These results are reported in Figure 8.1, where we see that the average layerwise ρ scores are lower in the last layers both for all distinct groups and for all features together. As suggested by [Liu et al., 2019a], this could be due to the fact that the representations that are better-suited for language modeling (output layer) are also those that exhibit worse probing task performance, indicating that Transformer layers trade off between encoding general and probed features. However, there are differences between the considered groups: competences about raw texts features (*RawText*) and the distribution of POS are lost in the very first layers (by layer -10), while the knowledge about the order of subject/object with respect to the verb, the use of subordination, as well as features related to verbal predicate structure is acquired in the middle layers.

Interestingly, if we consider how the knowledge of each feature changes across layers (Figure 8.2), we observe that not all features belonging to the same group have an

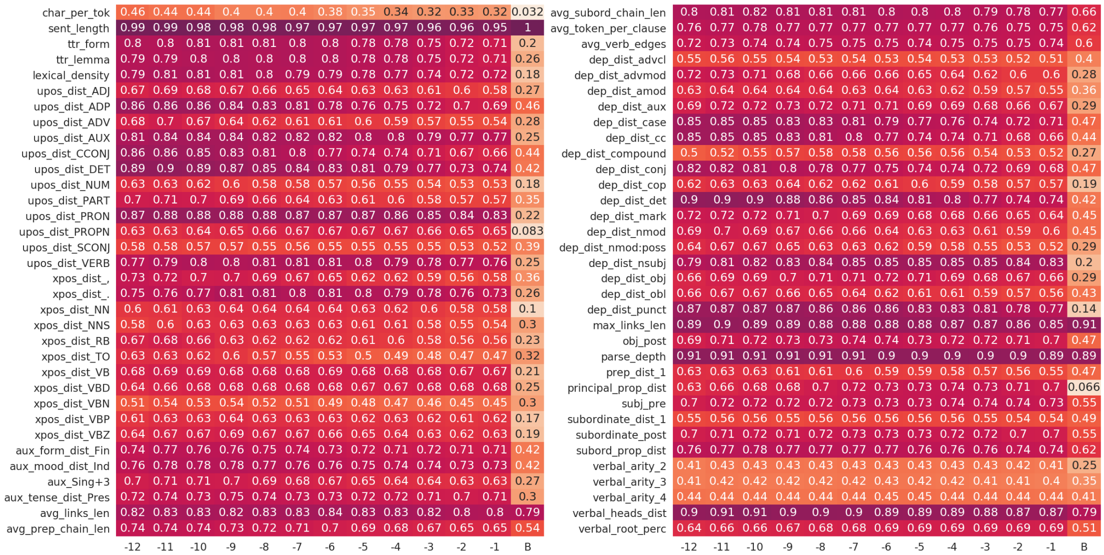


Figure 8.2: *Layerwise ρ scores for the 68 linguistic features. Absolute baseline scores are reported in column B.*

homogeneous behaviour. This is for example the case of the two features included in the *RawText* group: word length (*char_per_tok*) achieves quite lower scores across all layers with respect to the *sent_length* feature. Similarly, the knowledge about POS differs when we consider more granular distinctions. For instance, within the broad categories of verbs and nouns, worse predictions are obtained by sub–specific classes of verbs based on tense, person and mood features (see especially past participle, *xpos_dist_VBN*), and by inflected nouns both singular and plural (*_NN*, *_NNS*). Within the broad set of features extracted from syntactic annotation, we also see that different scores are reported for features referring e.g. to types of dependency relations: those linking a functional POS to its head (e.g. *dep_dist_case, dep_dist_cc, dep_dist_conj, dep_dist_det*) are better predicted than others relations, such as *dep_dist_amod, advcl*. Besides, within the *VerbPredicate* group, lower ρ scores are obtained by features encoding subcategorization information about verbal predicates, such as the distribution of verbs by arity (*verbal_arity_2,3,4*), which also remains almost stable across layers.

Since we observed these not homogeneous scores within the groups we defined a priori, we investigated how BERT hierarchically encodes across layers all the features. To this end, we clustered the 68 linguistic characteristics according to layerwise probing results: specifically, we performed hierarchical clustering using Euclidean distance as distance metric and Ward variance minimization as clustering method. Interestingly enough, Figure 8.3 shows that the traditional division of features with respect to the linguistic annotation levels has not been maintained. On the contrary, BERT puts together features from all linguistic groups into clusters of different size. In addition, these clusters gather features that are differently ranked according to the baseline scores (ranking positions are bolded in the figure). For example, the first cluster includes

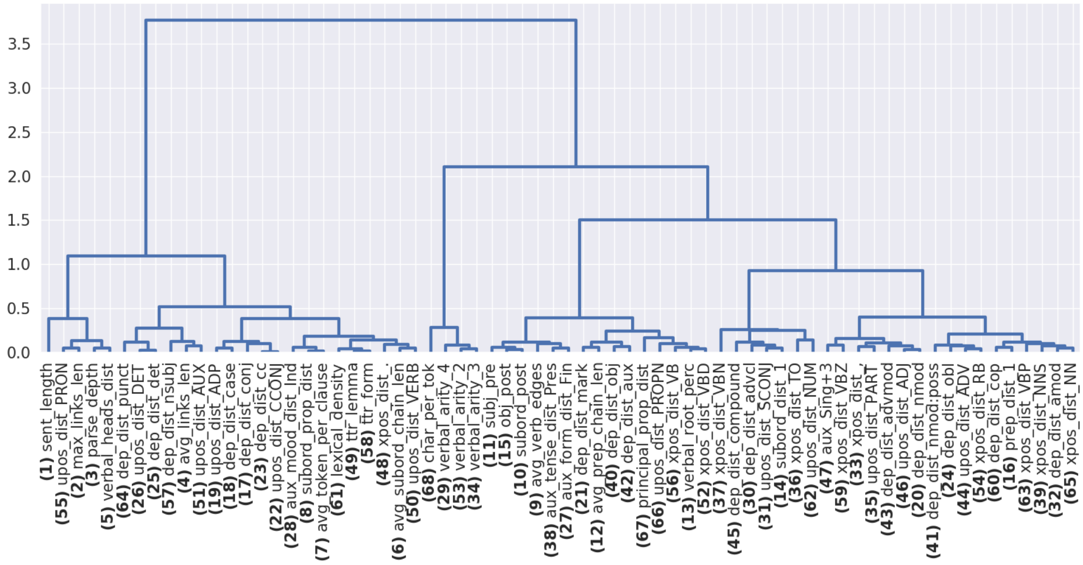


Figure 8.3: *Hierarchical clustering of the 68 probing tasks based on layerwise ρ values. Bold numbers correspond to the ranking of each probing feature based on the correlation with sentence length.*

	KOR	TEL	HIN	JPN	CHI	TUR	ARA	GER	FRE	SPA
Baseline	59.05	51.32	54.09	56.27	55.68	55.66	52.92	59.29	56.03	52.61
BERT	85.74	85.18	84.75	84.19	82.78	79.29	76.38	72.78	72.50	70.03

Table 8.3: *NLI classification results in terms of accuracy. We used the Zero Rule algorithm as baseline. Note that, for each task, sentences of the 10 languages are paired with the Italian ones (e.g. KOR = KOR-ITA).*

features with similar ρ scores, and both highly and lower ranked by the baseline. All these features model aspects of global sentence structure, e.g. *sent_length*, functional POSs (e.g. *upos_dist_DET*, *_ADP*, *_CCONJ*), parsed tree structures (e.g. *parse_depth, verbal_heads_dist, avg_links_len*), nuclear elements of the sentence such as subjects (*dep_dist_nsubj*), verbs (*_VERBS*), pronouns (*_PRON*). This behaviour seem to partially contradict the results observed by [Tenney et al., 2019a], where it was shown that BERT represents the steps of the traditional NLP pipeline in an interpretable and localizable way. In contrast, our results show that using a more granular set of linguistic features actually makes it much more difficult to identify well-defined patterns and regularities among the implicit competences learned by the model.

8.2.3 The Impact of Fine–Tuning on Linguistic Knowledge

Once we have probed the linguistic knowledge encoded by BERT across its layers, we investigated how it changes after a fine-tuning process. To do so, we started with the same pre-trained version of the model used in the previous experiment and performed a fine-tuning process for each of the 10 subsets built from the original NLI corpus (Sec. 8.2.1). We decided to use 50% of each NLI subset for training (40% and 10% for

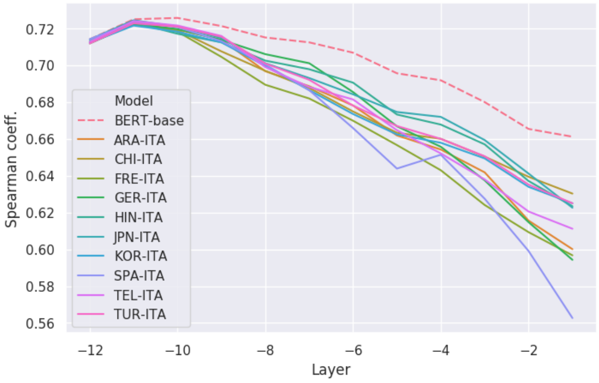


Figure 8.4: *Layerwise mean ρ scores for the pre-trained and fine-tuned models.*

training and development set) and the remaining 50% for testing the accuracy of the newly generated models. Table 8.3 reports the results for the 10 binary NLI tasks. As we can notice, BERT achieves good results for all downstream tasks, meaning that is able to discriminate the L1 of a native speaker on a sentence-level regardless of the L1 pairs taken into account. The best performance is achieved by the model that was fine-tuned on the Korean and Italian pairwise subset, while the lowest scores are obtained with the model trained on the subset consisting of essays written by Spanish and Italian L1 speakers (SPA-ITA). Interestingly, these results seem to reflect typological distances among L1 pairs, with higher scores for languages that are more distant from Italian (Korean, Telugu or Hindi) and lower scores for L1s belonging to the same language family (FRE-ITA or SPA-ITA).

After fine-tuning the model on NLI, we performed again the suite of probing tasks on the UD dataset using the 10 newly generated models and following the same approach discussed in Section 8.3.2. Figure 8.4 reports layerwise mean ρ correlation values for all probing tasks obtained with BERT-base and the other fine-tuned models. It can be noticed that the representations learned by the NLM tend to lose their precision in encoding our set of linguistic features after the fine-tuning process. This is particularly noticeable at higher layers and it possibly suggests that the model is storing task–specific information at the expense of its ability to encode general knowledge about the language. Again, this is particularly evident for the models fine–tuned on the classification of language pairs belonging to the same family, SPA–ITA above all. To study which phenomena are mainly involved in this loss, we computed the differences between the probing tasks results obtained before and after the fine-tuning process. We focused in particular on the scores obtained on the output layer representations (layer *-1*), since it is the most task-specific [Kovaleva et al., 2019]. For each subset, Figure 8.5 reports the difference between the score of each linguistic feature obtained with the pre–trained model and the fine–tuned one. Not surprisingly, the loss of linguistic knowledge reflects the typological trend observed for overall classification performance. In fact, when the task is to distinguish Italian vs German, French and Spanish L1, BERT loses much

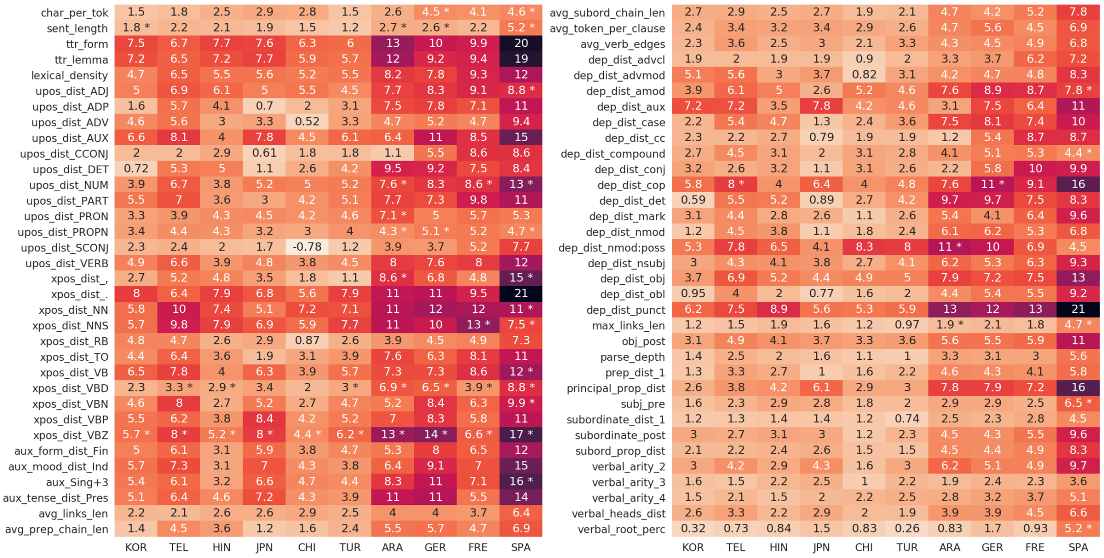


Figure 8.5: *Differences between BERT–base and fine–tuned models ρ scores (multiplied by 100) computed using the output layer representations (-1). Statistically significant variations (Wilcoxon Rank-sum test) are marked (*).*

of its encoded knowledge for almost all the considered features. This is particularly evident for the morpho-syntactic features (i.e. distribution of *upos_dist* and *xpos_dist*) and for features related to lexical variety (i.e. *ttr_form*, *ttr_lemma*). It seems that for typologically similar languages BERT needs more task-specific knowledge mostly encoded at the level of morpho-syntactic information rather than the structural level. On the contrary, the drop is less pronounced and in most cases not significant for models fine–tuned on the classification of more distant languages (e.g. models fine–tuned on KOR-ITA or TUR-ITA). In this case, the quite stable performance on the probing tasks may suggest that those features were still useful to perform the downstream task. Interestingly, the class of features that decreases significantly in all models are those encoding the knowledge about the tense of verbs. This is particularly the case of the third-person singular verbs in the present tense (*xpos_dist_VBZ*) and of verbs in the past tense (*xpos_dist_VBD*). A possible explanation could be related to the prompts of essays, which are the same across the NLI dataset. Thus, the textual genre could have favored a quite homogeneous use of verbal morphology features by students of all L1s. This makes this class of features less useful for the identification of native languages.

8.2.4 Are Linguistic Features useful for BERT's predictions?

As a last research question we investigated whether the implicit linguistic knowledge affects BERT's predictions when solving the NLI downstream task. To answer this question we have split each NLI subset into two groups, i.e. sentences correctly classified according to the L1 and those incorrectly classified. For the two groups of each NLI subset, we performed the probing tasks using the pre–trained BERT-base and the specific

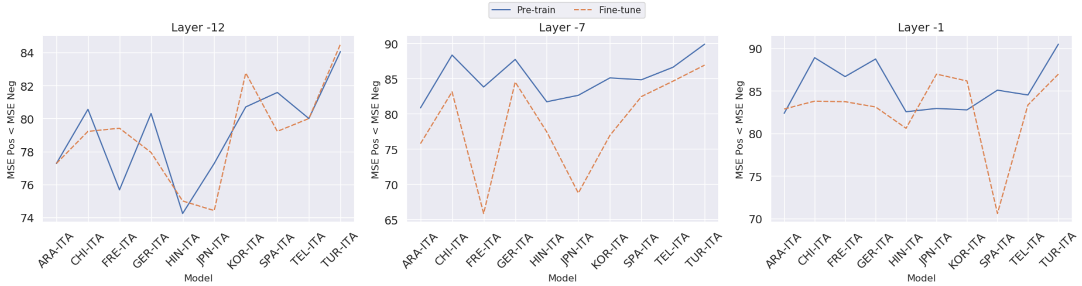


Figure 8.6: *% of probing features for which the MSE of the sentences correctly classified by BERT-base (Pre-train) and the fine-tuned models (Fine-tune) is lower than that of the incorrectly ones. Results are reported for layers -12, -7 and -1.*

NLI fine-tuned model. For each sentence of the two groups, we calculated the variation between the actual and predicted feature value obtaining two lists of absolute errors. We used the Wilcoxon Rank-sum test to verify whether the two lists were selected from samples with the same distribution. As a general remark, we observed that much more than half of features vary in a statistically significant way between correctly and incorrectly classified sentences. This suggests that BERT's linguistic competence on the two groups of sentences is very different. To deepen the analysis of this difference, we calculated the accuracy achieved by BERT in terms of Mean Square Error (MSE) only for the set of features varying in a significant way. Figure 8.6 reports the percentage of features for which the MSE of the sentences correctly classified (*MSE Pos*) is lower than that of the incorrectly ones (*MSE Neg*). This percentage is significantly higher, thus showing that BERT's capacity to encode different kind of linguistic information could have an influence on its predictions: the more BERT stores readable linguistic information into the representations it creates, the higher will be its capacity of predicting the correct L1. Moreover, we noticed that this is true also (and especially) using the pre-trained model. In other words, this result suggests that the evaluation of the linguistic knowledge encoded in a pre–trained version of BERT on a specific input sequence could be an insightful indicator of its ability in analyzing that sentence with respect to a downstream task.

Since this behavior might be simply due to the complexity of the sentences rather than the model itself, to investigate more in depth this phenomenon we analyze the average length of corrected and incorrected classified sentences. Interestingly, we notice the correct ones are much more longer than the others for all tasks (from 3 tokens more for SPA-ITA to 9 for TEL-ITA). This is quite expected for the NLI task, since a higher number of linguistic events possibly occurring in longer sentences are needed to classify the L1 of a sentence [Dell'Orletta et al., 2014]. At the same time, longer sentences make more complex the probing tasks because the output space is larger for almost all them. This is an additional evidence that BERT's linguistic knowledge is not strictly related to sentence complexity, but rather to the model's ability to solve a specific downstream task. To confirm this hypothesis and verify whether such tendency does not only depend on sentence length, we trained another LinearSVR that takes as input the sentence length and predict our probing tasks according to correctly or incorrectly classified NLI sentences. Table 8.4 reports the average Spearman's correlation coefficients between gold and predict probing features for the two classes of sentences. Results showed that, for all the considered language pairs, the LinearSVR achieved higher accuracy for the probing tasks computed with respect to the incorrectly NLI classified sentences. This is an additional evidence that deeper linguistic knowledge is needed for BERT to correctly classify the L1 of a sentences.

Model	**ARA**	**CHI**	**TUR**	**SPA**	**GER**	**FRE**	**JPN**	**KOR**	**TEL**	**HIN**
Correct	0.226	0.225	0.236	0.223	0.215	0.224	0.276	0.239	0.234	0.229
Incorrect	0.248	0.251	0.249	0.235	0.244	0.239	0.290	0.255	0.258	0.257

Table 8.4: *Average ρ scores for sentences correctly and incorrectly classified using only sentence length as input feature.*

8.3 Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation

After investigating the linguistic competence implicitly encoded by BERT within its internal representations and how it changes following a fine-tuning process, we decided to apply our methodology to study differences and similarities with the knowledge encoded by a contextual-independent LM. In fact, despite several works provided evidences that recent NLMs are able to encode a wide range of linguistic phenomena, less study focused on the analysis and the comparison of contextual and non-contextual NLMs according to their ability to encode implicit linguistic properties in their representations.

In this study we performed a large number of probing experiments to analyze and compare the implicit knowledge stored by a contextual and a non-contextual model within their inner representations. In particular, we define two research questions, aimed at understanding: (i) which is the best method for representing BERT or word2vec [Mikolov et al., 2013] word representations into sentence embeddings and how each model and sentence representation approach differently encodes properties related to the linguistic structure of a sentence; (ii) whether such sentence-level knowledge is preserved within BERT single-word representations. To answer our questions, we relied on a the same methodology and set of probing features used in Section 8.2.

The contributions are as follows:

- we perform an in-depth study aimed at understanding the linguistic knowledge encoded in a contextual (BERT) and a contextual-independent (word2vec) Neural Language Model;

- we evaluate the best method for obtaining sentence-level representations from BERT and word2vec according to a wide spectrum of probing tasks;

- we compare the results obtained by BERT and word2vec according to the different combining methods;
- we study whether BERT is able to encode sentence-level properties within its single word representations.

8.3.1 Our Approach

We studied how layer-wise internal representations of BERT encode a wide spectrum of linguistic properties and how such implicit knowledge differs from that learned by a context-independent model such as word2vec. Following the probing task approach and the suite of 68 probing features defined in Sec. 8.2, we defined two sets of experiments. The first consists in evaluating which is the best method for generating sentence-level embeddings using BERT and word2vec single-word representations. In particular, we performed the probing experiments using as input layer-wise BERT and word2vec combined representations for each sentence of the English UD dataset and we compared the results to understand which model performs better according to different levels of linguistic sophistication.

In the second set of experiments, we measured how many sentence-level properties are encoded in single-word representations. To do so, we performed our set of probing tasks using the embeddings extracted from both BERT and word2vec individual tokens. In particular, we considered the word representations corresponding to the first, last and two internal tokens for each sentence of the UD dataset.

Experimental Setting We relied on a pre-trained English version of BERT (BERT-base uncased, 12 layers) for the extraction of the contextual word embeddings. To obtain the representations for our sentence-level tasks we experimented four different combining methods: *Max-pooling*, *Min-pooling*, *Mean* and *Sum*. Each of this four combining methods returns a single $\vec{s}$ vector, such that each s_n is obtained by combining the n^{th} components $w_{1n}, w_{2n}, ..., w_{mn}$ of the embedding of each word in the input sentence. In order to conduct a comparison of context-based and word-based representations when solving our set of probing tasks, we performed all the probing experiments using also the embeddings extracted from a pre-trained version of word2vec. In particular, we trained the model on the English Wikipedia dataset (dump of March 2020), resulting in 300-dimensional vectors. In the same manner as BERT's contextual representations, we experimented four combining methods: *Max-pooling*, *Min-pooling*, *Mean* and *Sum*. We used a linear Support Vector Regression model (LinearSVR) as probing model.

In order to perform the probing experiments on gold annotated sentences, we relied on the UD English dataset.

8.3.2 Evaluating Sentence Representations

The first set of experiments consists in evaluating which is the best method for combining word-level embeddings into sentence representations in order to understand what kind

Categories	BERT	word2vec	Baseline
Raw text	**0.65**	0.51	0.37
Morphosyntax	0.49	**0.57**	0.28
Syntax	0.55	**0.56**	0.44
All features	0.53	**0.56**	0.38

Table 8.5: *BERT (average between layers) and word2vec ρ scores computed by averaging Max-, Min-, Mean and Sum scores according to the three linguistic levels of annotations and considering all the probing features (All features). Baseline scores are also reported.*

Categories	Sum	Min	Max	Mean
Raw text	**0.56**	0.51	0.51	0.46
Morphosyntax	0.59	0.52	0.54	**0.61**
Syntax	**0.61**	0.55	0.55	0.54
All features	**0.60**	0.54	0.55	0.57

Table 8.6: *word2vec probing scores obtained with the four sentence combining methods.*

of implicit linguistic properties are encoded within both contextual and non-contextual representations using different combining methods. To do so, we firstly extracted from each sentence in the UD dataset the corresponding word embeddings using the output of the internal representations of word2vec and BERT layers (from input layer *-12* to output layer *-1*). Secondly, we computed the sentence-representations according to the different combining strategies defined in 8.5.1. We then performed our set of 68 probing tasks using the LinearSVR model for each sentence representation. Since the majority of our probing features is correlated to sentence length, we compared probing results with the ones obtained with a baseline computed by measuring the ρ coefficient between the length of the UD sentences and each of the 68 probing features. Evaluation was performed with a 5-cross fold validation and using Spearman correlation score (ρ) between predicted and gold labels as evaluation metric.

Table 8.11 report average ρ scores aggregating all probing results (*All features*) and according to raw text (*Raw text*), morphosyntactic (*Morphosyntax*) and syntactic (*Syntax*) levels of annotations. Scores are computed by averaging *Max-*, *Min-pooling*, *Mean* and *Sum results*. As a general remark, we notice that the scores obtained by word2vec and BERT's internal representations outperforms the ones obtained with the correlation baseline, thus showing that both models are capable of implicitly encoding a wide spectrum of linguistic phenomena. Interestingly, we can notice that word2vec sentence representations outperform BERT ones when considering all the probing features in average.

We report in Table 8.6 and Figure 10.9 the probing scores obtained by the two models. For what concerns word2vec representations, we notice that the *Sum* method prove to be the best one for encoding raw text and syntactic features, while morophosyntactic properties are better represented averaging all the word embeddings (*Mean*). In general, best results are obtained with probing tasks related to morphosyntactic and syntactic features, like the distribution of POS (e.g. *upos_dist_PRON*, *upos_dist_VERB*) or the

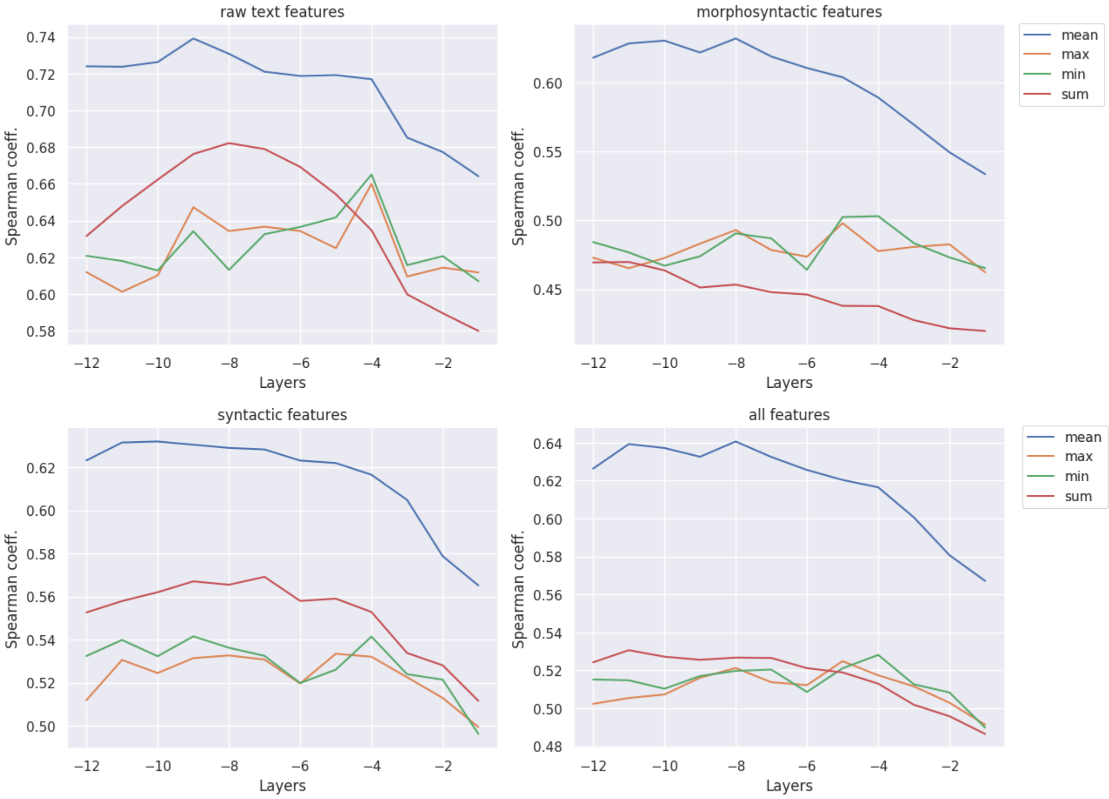


Figure 8.7: *Layerwise ρ scores for the three categories of raw-text, morphosyntactic and syntactic features. Layerwise average results are also reported. Each line in the four plots corresponds to a different aggregating strategy.*

maximum depth of the syntactic tree (*parse_depth*). If we look instead at the average ρ scores obtained with BERT layerwise representations (Figure 10.9), we observe that, differently from word2vec, best results are the ones related to raw-text features, such as sentence length or Type/Token Ratio. The *Mean* method prove to be the best one for almost all the probing tasks, achieving highest scores in the first five layers. The only exceptions mainly concern some of the linguistic features related to syntactic properties, e.g. the average length of dependency links (*avg_links_len*) or the maximum depth of the syntactic tree (*parse_depth*), for which best scores across layers are obtained with the *Sum* strategy. The *Max*- and *Min-pooling* methods, instead, show a similar trend for almost all the probing features.

In order to investigate more in depth how the linguistic knowledge encoded by BERT across its layers differs from that learned by word2vec, we report in Table 8.7 average ρ differences between the two models according to the four combining strategies. As a general remark, we can notice that, regardless of the aggregation strategy taken into account, BERT and word2vec sentence representations achieve quite similar results on average. Hence, although BERT is capable of understanding the full context of each word in an input sequence, the amount of linguistic knowledge implicitly encoded in its aggregated sentence representations is still comparable to that which can be achieved

Layers	Mean	Max-pooling	Min-pooling	Sum
-12	.052	-.058	-.038	-.091
-11	.065	-.055	-.038	-.084
-10	.063	-.053	-.043	-.088
-9	.058	-.044	-.036	-.089
-8	.066	-.039	-.034	-.088
-7	.058	-.046	-.033	-.088
-6	.051	-.048	-.045	-.094
-5	.046	-.035	-.032	-.096
-4	.042	-.043	-.025	-.102
-3	.026	-.049	-.041	-.113
-2	.006	-.057	-.045	-.119
-1	-.007	-.069	-.063	-.128

Table 8.7: *Average ρ differences between BERT and word2vec probing results according to the four embedding-aggregation strategies.*

with a non-contextual language model.

In Figure 8.8 we report instead the differences between BERT and word2vec scores for all the 68 probing features (ordered by correlation with sentence length). For the comparison, we used the representations obtained with the *Mean* combining method. As a first remark, we notice that there is a clear distinction in terms of ρ scores between features better predicted by BERT and word2vec. In fact, features most related to syntactic properties (left heatmap) are those for which BERT results are generally higher with respect to those obtained with word2vec. This result demonstrates that BERT, unlike a non-contextual language model as word2vec, is able to encode information within its representations that involves the entire input sequence, thus making more simple to solve probing tasks that refer to syntatic characteristics.

Focusing instead on the right heatmap, we observe that word2vec non-contextual representations are still capable of encoding a wide spectrum of linguistic properties with higher ρ values compared to BERT ones, especially if we consider scores closer to BERT's output layers (from *-4* to *-1*). This is particularly evident for morphosyntactic features related to the distribution of POS categories (*xpos_dist_**, *upos_dist_**), most likely because non-contextual representations tend to encode properties related to single tokens rather than syntactic relations between them.

8.3.3 Evaluating Word Representations

Once we have probed the linguistic knowledge encoded by BERT and word2vec using different strategies for computing sentence embeddings, we investigated how much information about the structure of a sentence is encoded within single-word contextual representations. For doing so, we performed our sentence-level probing tasks using a single BERT word embedding for each sentence in the UD dataset. We tested four different words, corresponding to the first, the last and two internal tokens for each sentence in the UD dataset. In particular, we extracted the embeddings from the output

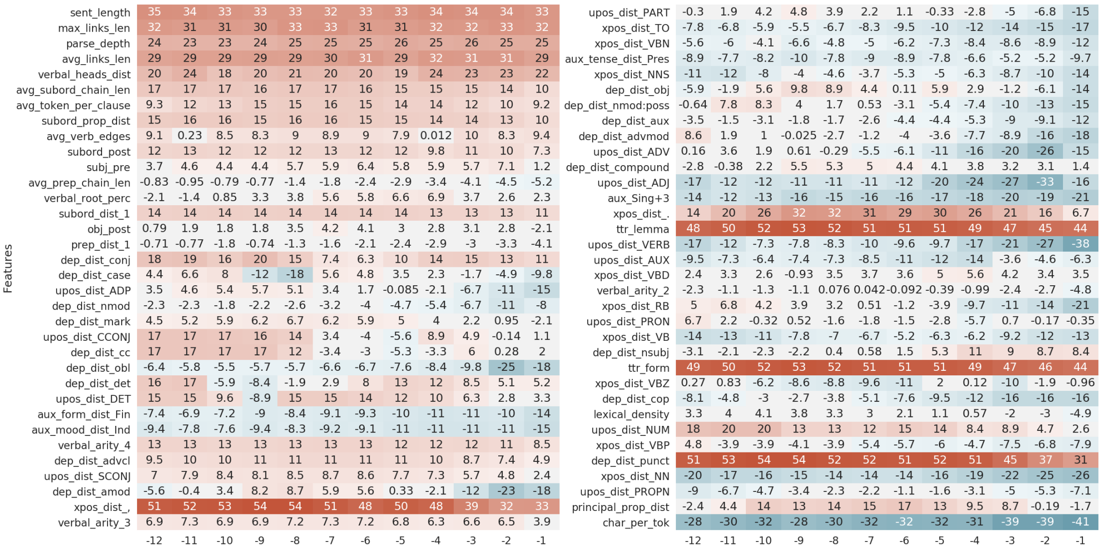


Figure 8.8: *Differences between BERT and word2vec scores (multiplied by 100) for all the 68 probing features (ranked by correlation with sentence length), obtained with the Mean aggregation strategy. BERT scores are reported for all the 12 layers. Positive (red) and negative (blue) cells correspond to scores for which BERT outperforms word2vec and vice versa.*

layer (*-1*) and from the layer that achieved best results in the previous experiments (*-8*). We used probing scores obtained with word2vec embeddings for the same tokens as baseline. In Table 8.8 we report average ρ scores obtained by BERT (*BERT-**) and word2vec (*word2vec-**) according to word-level representations extracted from the four tokens mentioned above. Results were computed aggregating all probing results (*All*) and according to raw text (*Raw*), morphosyntactic (*Morphosyntax*) and syntatic (*Syntax*) levels of annotation. For comparison, we also report average scores obtained with BERT *[CLS]* token.

As a first remark, we can clearly notice that even with a single-word embedding BERT is able to encode a wide spectrum of sentence-level linguistic properties. This result allows us to highlight the main potential of contextual representations, i.e. the capability of capturing linguistic phenomena that refer to the entire input sequence within single-word representations. An interesting observation is that, except for the raw text features, for which the best scores are achieved using *[CLS]*, higher performance are obtained with the embeddings corresponding to *BERT-4*, i.e. the last token of each sentence. This result seems to indicate that *[CLS]*, although being used for classification predictions, does not necessarily correspond to the most linguistically informative token within each input sequence.

Comparing the results with those achieved using word2vec word embeddings, we notice that BERT scores greatly outperform word2vec for all the probing tasks. This is a straightforward result and can be easily explained by the fact that the lack of contextual knowledge does not allow single-word representations to encode information that are related to the structure of the whole sentence.

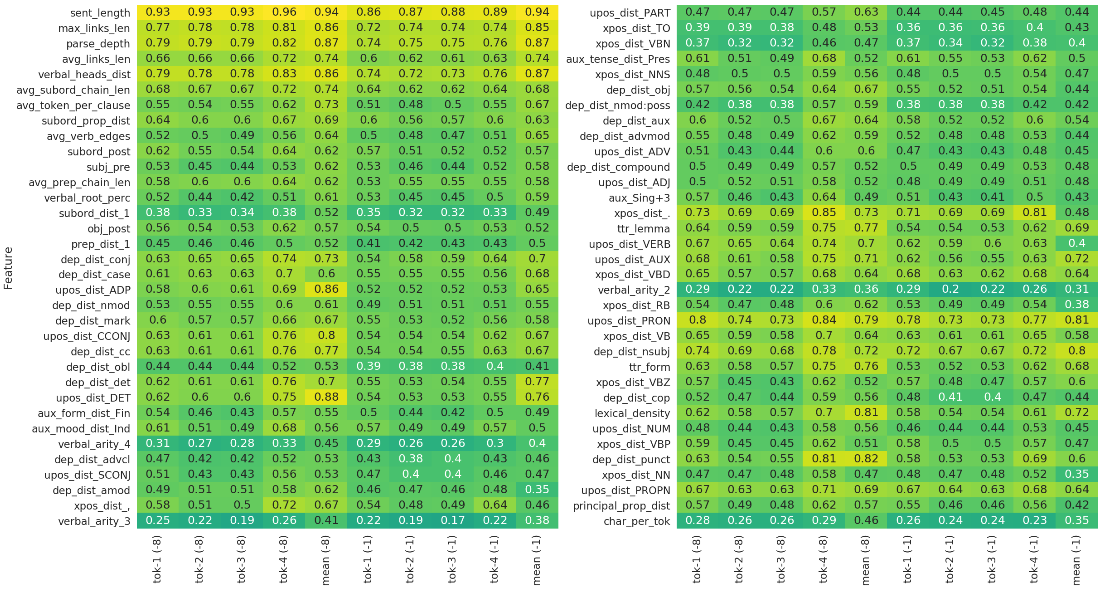
Embeddings	**Raw**	**Morphoyntax**	**Syntax**	**All**
BERT-1 (-8)	0.62	0.57	0.55	0.57
BERT-2 (-8)	0.59	0.53	0.53	0.53
BERT-3 (-8)	0.59	0.52	0.52	0.53
BERT-4 (-8)	0.65	**0.66**	**0.62**	**0.64**
BERT-1 (-1)	0.55	0.55	0.51	0.53
BERT-2 (-1)	0.54	0.51	0.49	0.50
BERT-3 (-1)	0.54	0.51	0.49	0.50
BERT-4 (-1)	0.59	0.57	0.53	0.55
[CLS] (-8)	**0.66**	0.47	0.52	0.51
[CLS] (-1)	0.61	0.45	0.49	0.48
word2vec-1	0.26	0.26	0.22	0.24
word2vec-2	0.17	0.21	0.18	0.19
word2vec-3	0.17	0.19	0.17	0.18
word2vec-4	0.13	0.15	0.12	0.13

Table 8.8: *Average ρ scores obtained by BERT and word2vec according to word representations corresponding to the first, the last and two internal tokens of each input sentence. Results are computed according to the three linguistic levels of annotation and considering all the probing features (All). Average scores obtained with the [CLS] token are also reported.*

Since the latter results demonstrated that BERT is capable of encoding many sentence-level properties within its single word representations, as a last analysis, we decided to compare these results with the ones obtained using sentence embeddings. In particular, Figure 8.9 reports probing scores obtained by BERT single word (*tok_**) and *Mean* sentence representations (*sent*) extracted from the output layer (*-1*) and from the layer that achieved best results in average (*-8*).

As already mentioned, for many of these probing tasks, word embeddings performance is comparable to that obtained with the aggregated sentence representations. Nevertheless, there are several cases in which the difference between performance is particularly significant. Interestingly, we can notice that aggregated sentence representations are generally better for predicting properties belonging to the left heatmap, i.e. to the group of features more related to syntactic properties. This is particularly noticeable for the average number of tokens per clause (*avg_token_per_clause*) or the distribution of subordinate chains by length (*subord_dist*), for which we observe an improvement from word-level to sentence-level representations of more than .10 ρ points. On the contrary, probing features belonging to the right heatmap, therefore more close to raw text and morphosyntactic properties, are generally better predicted using single word embeddings, especially when considering the inner representations corresponding to the last token in each sentence (*tok_4*). The property most affected by the difference in scores between word- and sentence-level embeddings is the the distribution of periods (*xpos_dist_.*).

Focusing instead on differences in performance between the two considered layers, we can notice that regardless of the method used to predict each feature, the representations learned by BERT tend to lose their precision in encoding our set of linguistic



Feature	tok-1 (-8)	tok-2 (-8)	tok-3 (-8)	tok-4 (-8)	mean (-8)	tok-1 (-1)	tok-2 (-1)	tok-3 (-1)	tok-4 (-1)	mean (-1)
sent_length	0.93	0.93	0.93	0.96	0.94	0.86	0.87	0.88	0.89	0.94
max_links_len	0.77	0.78	0.78	0.81	0.86	0.72	0.74	0.74	0.74	0.85
parse_depth	0.79	0.79	0.79	0.82	0.87	0.74	0.75	0.75	0.76	0.87
avg_links_len	0.66	0.66	0.66	0.72	0.74	0.6	0.62	0.61	0.63	0.74
verbal_heads_dist	0.79	0.78	0.78	0.83	0.86	0.74	0.72	0.73	0.76	0.87
avg_subord_chain_len	0.68	0.67	0.67	0.72	0.74	0.64	0.62	0.62	0.64	0.68
avg_token_per_clause	0.55	0.54	0.55	0.62	0.73	0.51	0.48	0.5	0.55	0.67
subord_prop_dist	0.64	0.6	0.6	0.67	0.69	0.6	0.56	0.57	0.6	0.63
avg_verb_edges	0.52	0.5	0.49	0.56	0.64	0.5	0.48	0.47	0.51	0.65
subord_post	0.62	0.55	0.54	0.64	0.62	0.57	0.51	0.52	0.52	0.57
subj_pre	0.53	0.45	0.44	0.53	0.62	0.53	0.46	0.44	0.52	0.58
avg_prep_chain_len	0.58	0.6	0.6	0.64	0.62	0.53	0.55	0.55	0.55	0.58
verbal_root_perc	0.52	0.44	0.42	0.51	0.61	0.53	0.45	0.45	0.5	0.59
subord_dist_1	0.38	0.33	0.34	0.38	0.52	0.35	0.32	0.32	0.33	0.49
obj_post	0.56	0.54	0.53	0.62	0.57	0.54	0.5	0.5	0.53	0.52
prep_dist_1	0.45	0.46	0.46	0.5	0.52	0.41	0.42	0.43	0.43	0.5
dep_dist_conj	0.63	0.65	0.65	0.74	0.73	0.54	0.58	0.59	0.64	0.7
dep_dist_case	0.61	0.63	0.63	0.7	0.6	0.55	0.55	0.55	0.56	0.68
upos_dist_ADP	0.58	0.6	0.61	0.69	0.86	0.52	0.52	0.52	0.53	0.65
dep_dist_nmod	0.53	0.55	0.55	0.6	0.61	0.49	0.51	0.51	0.51	0.55
dep_dist_mark	0.6	0.57	0.57	0.66	0.67	0.55	0.53	0.52	0.56	0.58
upos_dist_CCONJ	0.63	0.61	0.61	0.76	0.8	0.54	0.54	0.54	0.62	0.67
dep_dist_cc	0.63	0.61	0.61	0.76	0.77	0.54	0.54	0.55	0.63	0.67
dep_dist_obl	0.44	0.44	0.44	0.52	0.53	0.39	0.38	0.38	0.4	0.41
dep_dist_det	0.62	0.61	0.61	0.76	0.7	0.55	0.53	0.54	0.55	0.77
upos_dist_DET	0.62	0.6	0.6	0.75	0.88	0.54	0.53	0.53	0.55	0.76
aux_form_dist_Fin	0.54	0.46	0.43	0.57	0.55	0.5	0.44	0.42	0.5	0.49
aux_mood_dist_Ind	0.61	0.51	0.49	0.68	0.56	0.57	0.49	0.49	0.57	0.5
verbal_arity_4	0.31	0.27	0.28	0.33	0.45	0.29	0.26	0.26	0.3	0.4
dep_dist_advcl	0.47	0.42	0.42	0.52	0.53	0.43	0.38	0.4	0.43	0.46
upos_dist_SCONJ	0.51	0.43	0.43	0.56	0.53	0.47	0.4	0.4	0.46	0.47
dep_dist_amod	0.49	0.51	0.51	0.58	0.62	0.46	0.47	0.46	0.48	0.35
xpos_dist_,	0.58	0.51	0.5	0.72	0.67	0.54	0.48	0.49	0.64	0.46
verbal_arity_3	0.25	0.22	0.19	0.26	0.41	0.22	0.19	0.17	0.22	0.38

Feature	tok-1 (-8)	tok-2 (-8)	tok-3 (-8)	tok-4 (-8)	mean (-8)	tok-1 (-1)	tok-2 (-1)	tok-3 (-1)	tok-4 (-1)	mean (-1)
upos_dist_PART	0.47	0.47	0.47	0.57	0.63	0.44	0.44	0.45	0.48	0.44
xpos_dist_TO	0.39	0.39	0.38	0.48	0.53	0.36	0.36	0.36	0.4	0.43
xpos_dist_VBN	0.37	0.32	0.32	0.46	0.47	0.37	0.34	0.32	0.38	0.4
aux_tense_dist_Pres	0.61	0.51	0.49	0.68	0.52	0.61	0.55	0.53	0.62	0.5
xpos_dist_NNS	0.48	0.5	0.5	0.59	0.56	0.48	0.5	0.5	0.54	0.47
dep_dist_obj	0.57	0.56	0.54	0.64	0.67	0.55	0.52	0.51	0.54	0.44
dep_dist_nmod:poss	0.42	0.38	0.38	0.57	0.59	0.38	0.38	0.38	0.42	0.42
dep_dist_aux	0.6	0.52	0.5	0.67	0.64	0.58	0.52	0.52	0.6	0.54
dep_dist_advmod	0.55	0.48	0.49	0.62	0.59	0.52	0.48	0.48	0.53	0.44
upos_dist_ADV	0.51	0.43	0.44	0.6	0.6	0.47	0.43	0.43	0.48	0.45
dep_dist_compound	0.5	0.49	0.49	0.57	0.52	0.5	0.49	0.49	0.53	0.48
upos_dist_ADJ	0.5	0.52	0.51	0.58	0.52	0.48	0.49	0.49	0.51	0.48
aux_Sing+3	0.57	0.46	0.43	0.64	0.49	0.51	0.43	0.41	0.5	0.43
xpos_dist_.	0.73	0.69	0.69	0.85	0.73	0.71	0.69	0.69	0.81	0.48
ttr_lemma	0.64	0.59	0.59	0.75	0.77	0.54	0.54	0.53	0.62	0.69
upos_dist_VERB	0.67	0.65	0.64	0.74	0.7	0.62	0.59	0.6	0.63	0.4
upos_dist_AUX	0.68	0.61	0.58	0.75	0.71	0.62	0.56	0.55	0.63	0.72
xpos_dist_VBD	0.65	0.57	0.57	0.68	0.64	0.68	0.63	0.62	0.68	0.64
verbal_arity_2	0.29	0.22	0.22	0.33	0.36	0.29	0.2	0.22	0.26	0.31
xpos_dist_RB	0.54	0.47	0.48	0.6	0.62	0.53	0.49	0.49	0.54	0.38
upos_dist_PRON	0.8	0.74	0.73	0.84	0.79	0.78	0.73	0.73	0.77	0.81
xpos_dist_VB	0.65	0.59	0.58	0.7	0.64	0.63	0.61	0.61	0.65	0.58
dep_dist_nsubj	0.74	0.69	0.68	0.78	0.72	0.72	0.67	0.67	0.72	0.8
ttr_form	0.63	0.58	0.57	0.75	0.76	0.53	0.52	0.53	0.62	0.68
xpos_dist_VBZ	0.57	0.45	0.43	0.62	0.52	0.57	0.48	0.47	0.57	0.6
dep_dist_cop	0.52	0.47	0.44	0.59	0.56	0.48	0.41	0.4	0.47	0.44
lexical_density	0.62	0.58	0.57	0.7	0.81	0.58	0.54	0.54	0.61	0.72
upos_dist_NUM	0.48	0.44	0.43	0.58	0.56	0.46	0.44	0.44	0.53	0.45
xpos_dist_VBP	0.59	0.45	0.45	0.62	0.51	0.58	0.5	0.5	0.57	0.47
dep_dist_punct	0.63	0.54	0.55	0.81	0.82	0.58	0.53	0.53	0.69	0.6
xpos_dist_NN	0.47	0.47	0.48	0.58	0.47	0.48	0.47	0.48	0.52	0.35
upos_dist_PROPN	0.67	0.63	0.63	0.71	0.69	0.67	0.64	0.63	0.68	0.64
principal_prop_dist	0.57	0.49	0.48	0.62	0.57	0.55	0.46	0.46	0.56	0.42
char_per_tok	0.28	0.26	0.26	0.29	0.46	0.26	0.24	0.24	0.23	0.35

Figure 8.9: *Probing scores obtained by BERT word (tok_*) and sentence (mean) representations extracted from layers -1 and -8. Sentence embeddings are computed using the Mean method.*

properties, most likely because the model is storing task-specific information (Masked Language Modeling task) at the expense of its ability to encode general knowledge about the language.

8.4 How Do BERT Embeddings Organize Linguistic Knowledge?

Once probed the linguistic competence implicitly encoded by a NLM, another issue concern the way in which this information is localized within its internal representations. Relying on the same set of linguistic features exploited in 8.2 and 8.3, we proposed an in-depth investigation aimed at understating how sentence-level linguistic knowledge encoded by BERT is arranged within its representations. In particular, we defined two research questions, aimed at: (i) investigating the relationship between the sentence-level linguistic knowledge encoded in a pre-trained version of BERT and the number of individual units involved in the encoding of such knowledge; (ii) understanding how these sentence-level properties are organized within the internal representations of BERT, identifying groups of units more relevant for specific linguistic tasks.

8.4.1 Approach

To study how the information used by BERT to implicitly encode linguistic properties is arranged within its internal representations, we relied on a variable selection approach based on Lasso regression [Tibshirani, 1996], which aims at keeping as few non-zero coefficients as possible when solving specific regression tasks. Our aim was to identify which weights within sentence-level BERT internal representations can be set to zero, in order to understand the relationship between hidden units and linguistic competence and whether the information needed to perform similar linguistic tasks is encoded in similar positions.

Lasso regression consists in adding an L_1 penalization to the usual ordinary least square loss. To do so, one of the most relevant parameters is λ, which tunes how relevant the L_1 penalization is for the loss function. We performed a grid search with cross validation for each feature-layer pair, in order to identify the best suited value for λ according to each task. Specifically, our goal was to find the most suited value for seeking the best performance when having as few non-zero coefficients as possible.

Model and Data We used the same pre-trained BERT model used in the previous experiments and we experimented with both the activation of the first input token (*[CLS]*) and the mean of all the word embeddings in a sentence (*Mean-pooling*).

As regards the probing features, we relied once again on our set of 68 probing features extracted from each sentence of the English UD treebank.

8.4.2 Linguistic competence and BERT units

As a first analysis, we investigated the relationship between the implicit linguistic properties encoded in the internal representations of BERT and the number of individual units involved in the encoding of these properties. Figure 8.10 and 8.11 report layerwise R^2 results for all the probing tasks along with the number of non-zero coefficients obtained with the sentence representations computed with the *[CLS]* token and the *Mean-pooling* strategy respectively. As a first remark, we can notice that the *Mean-pooling* method proved to be the best one for almost all the probing features across the 12 layers. Moreover, in line with the results obtained with the previous experiments performed with the LinearSVR probing model, we noticed that there is high variability among different tasks, whereas less variation occurs among the model layers. Focusing instead on the relationship between R^2 scores and number of non-zero coefficients, we can notice that although best scores are achieved at lower layers (between layers 12 and 8 for both configurations), the highest number of non-zero coefficients occurs instead at layers closer to the output. This is particularly evident for the results achieved using the *[CLS]* token, for which we observe a continuous increase across the 12 layers in the number of units used by the the probing models.

For both configurations, features more related to the structure of the whole syntactic tree are those for which less units were set to zero during regression (e.g. *max_links_len*, *parse_depth*, *n_prepositional_chains*), while properties belonging to word–based properties (i.e. features related to POS and dependency labels) were predicted relying on less units. Moreover, we can clearly notice that features related to specific POS and dependency relationships are also those that gained less units through the 12 layers (e. g. *xpos_dist_.*, *xpos_dist_AUX*). On the contrary, features belonging to the structure of the syntactic tree tend to acquire more non-zero units as the output layer is approached. This is particularly evident for the linguistic features predicted using sentence representations computed using the *[CLS]* token (e.g. *subj_pre*, *parse_depth*, *n_prepositional_chains*).

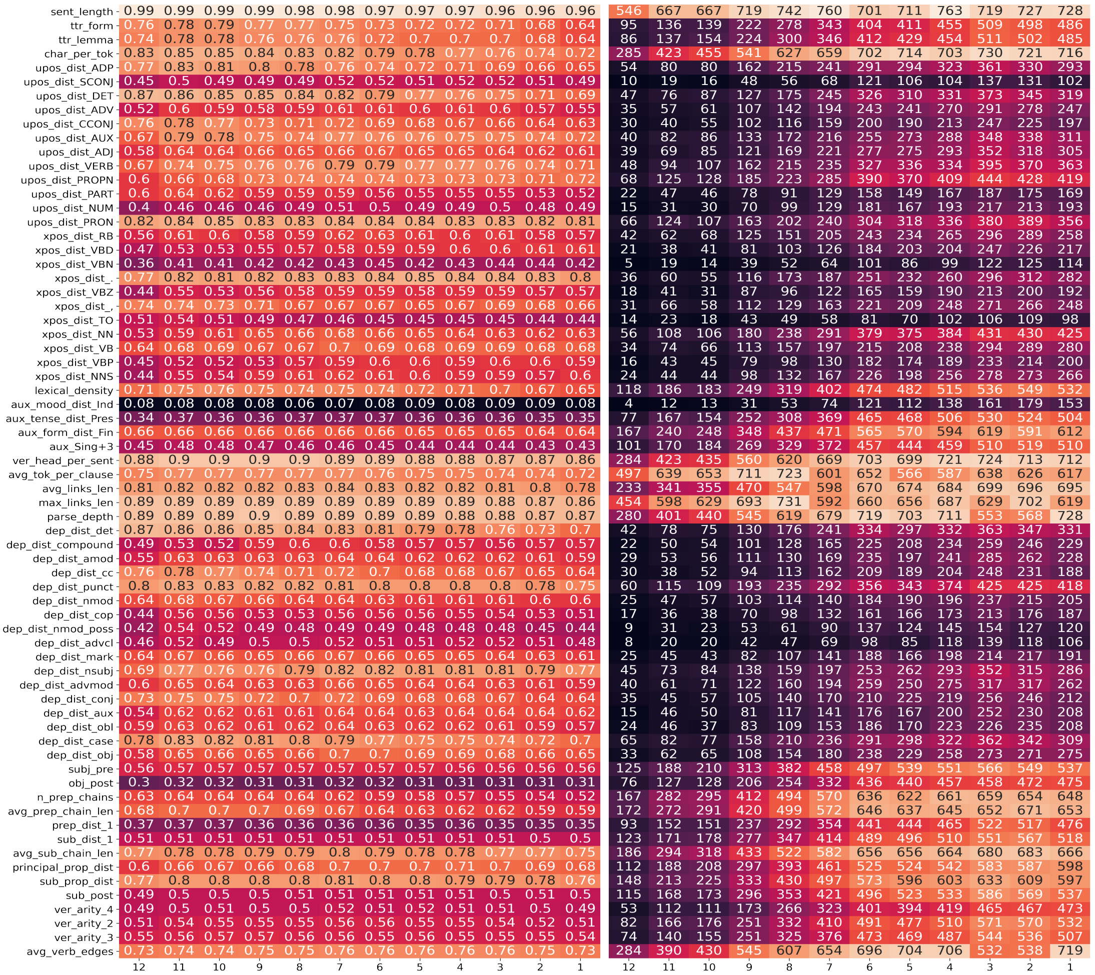


Figure 8.10: *Layerwise R^2 results for all the probing tasks (left heatmap) along with the number of non-zero coefficients (right heatmap) obtained with the sentence representations computed using the [CLS] token.*

We believe this is due to the fact that the interdependence between different units in each representation tend to increase across layers, thus making the information less localized especially for those features that belong to the whole structure of the syntactic tree. This is coherent with the fact that using the *Mean-pooling* strategy a higher number of non-zero coefficients was preserved also in the very first input layers, suggesting that this strategy increases the interdependence between each unit and makes the extraction of localized information more complex.

In order to focus more closely on the relationship between R^2 scores and non-zero units, we reported in Figure 8.12 average R^2 scores versus average number of non-zero coefficients, along with the line of best fit, for each layer and according to the *[CLS]* token and to the *Mean-pooling* strategy respectively. Interestingly, for both *[CLS]* and *Mean-pooling* representations, R^2 scores tend to improve as the number of non-zero coefficients increases. Moreover, when considering sentence representations computed with the *[CLS]* token, this behaviour becomes more pronounced as the output layer is

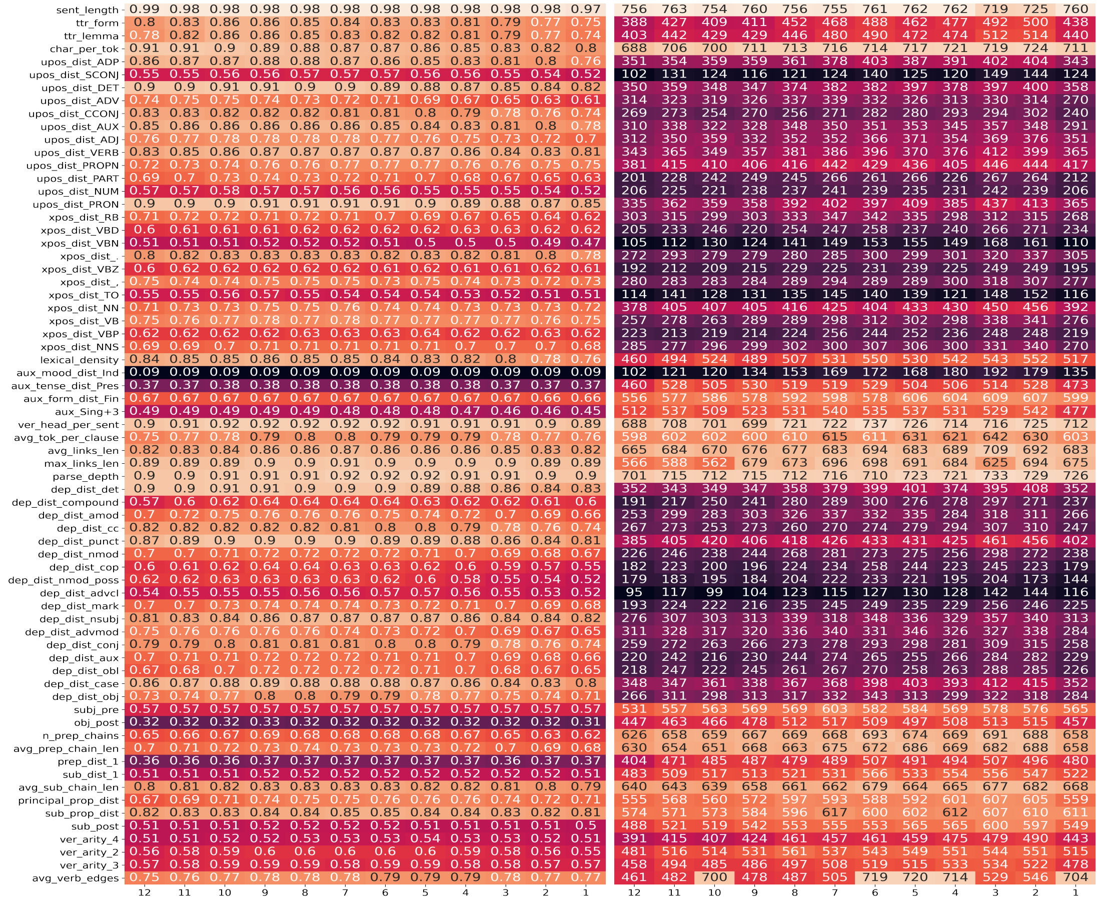


Figure 8.11: *Layerwise R^2 results for all the probing tasks (left heatmap) along with the number of non-zero coefficients (right heatmap) obtained with the sentence representations computed with the Mean-pooling strategy.*

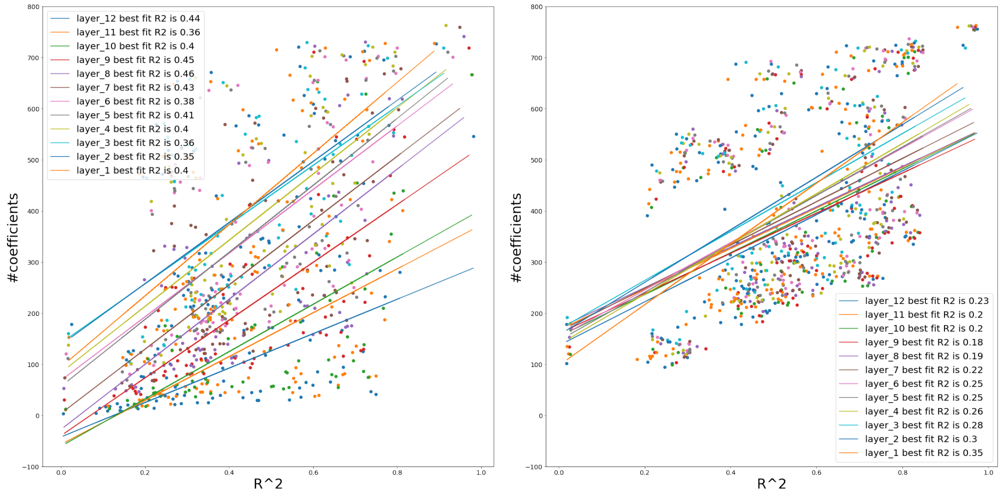


Figure 8.12: *Average R^2 scores versus average number of non-zero coefficients, along with the line of best fit, for each layer and according to [CLS] and Mean-pooling strategy respectively.*

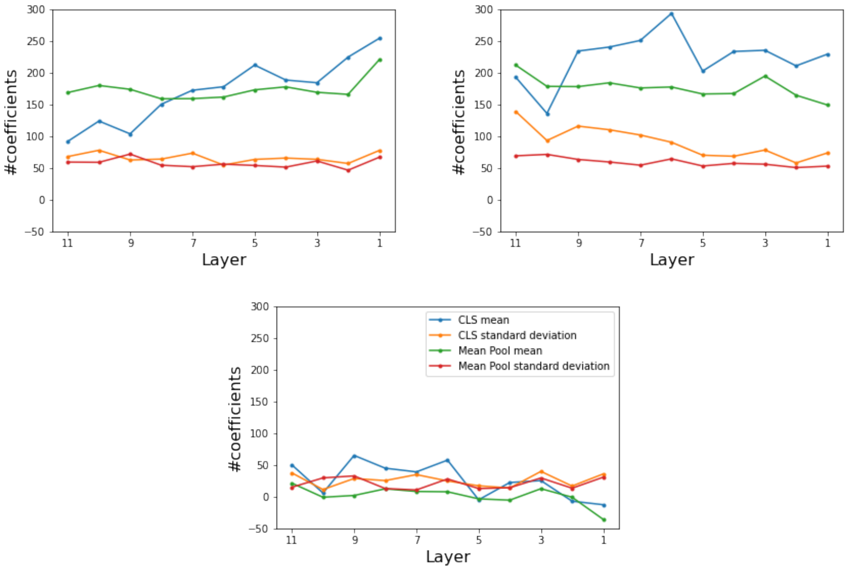


Figure 8.13: *Average number of non-zero coefficients in a layer that are set to zero in the following one (average number of dropped coefficients), average number of zero coefficients in a layer that are set to non-zero in the following one (average number of gained coefficients) and the value of the difference between the number of non-zero coefficients at pairs of consecutive layers (average number of changed coefficients).*

reached. This is in line with what we already noticed, namely that the interdependence between different units tend to increase across layers, especially when taking into account representations extracted without using a mean-pooling strategy.

In order to investigate more in depth the behaviour of BERT hidden units when solving the probing tasks, we focused more closely at how the different units in the internal representations are kept and lost across subsequent layers. Figure 8.13 reports, respectively, the average number of non-zero coefficients in a layer that are set to zero in the following one, the average number of zero coefficients in a layer that are set to non-zero in the following one and the average value of the difference between the number of non-zero coefficients at pairs of consecutive layers. As it can be observed, there is high coherence between each layer and its subsequent one, meaning that the variation in the number of selected coefficient is stable. However, the first two plots also show that there is a higher variation when considering non-zero coefficients in the same positions between pairs of layers. This underlines the fact that the information is not localized within BERT's internal representations, since the algorithm shows a degree of freedom in which units can be zeroed and which cannot.

In Figure 8.14 we report instead how many times each individual unit in the *[CLS]* and *Mean-pooling* internal representations has been kept non-zero when solving the 68 probing tasks for all the 12 BERT layers (816 regression task). In general, we can observe that the regression tasks performed using sentence-level representations obtained with the *Mean-pooling* strategy tend to use more hidden units with respect to the *[CLS]*

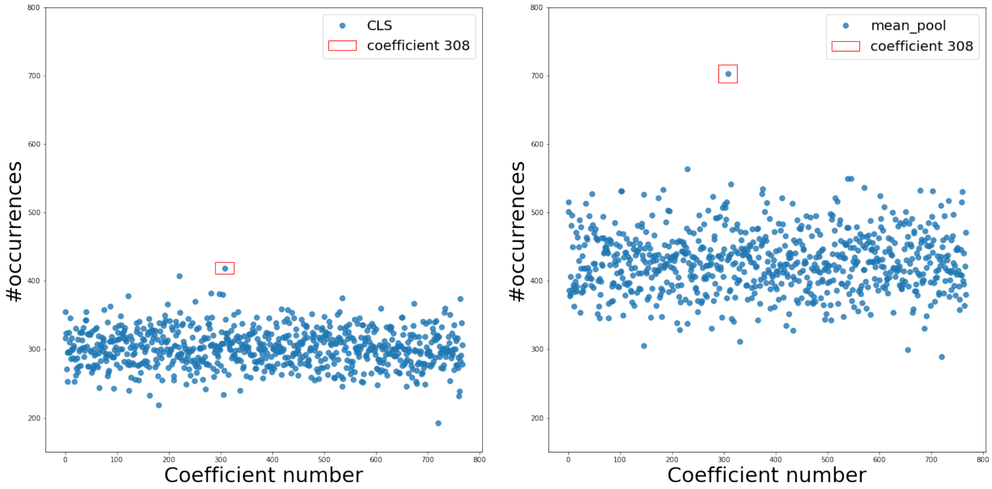


Figure 8.14: *Number of times in which each BERT individual unit (computed with [CLS] and with Mean-pooing aggregation strategy respectively) has been kept as non-zero when solving all the probing tasks for all the 12 layers.*

ones. It is also interesting to notice that there is a highly irregular unit (number 308) that has been kept different from zero in a number of tasks and layers much higher than the average. This could suggest that this unit is particularly relevant for encoding almost all the linguistic properties devised in our probing tasks.

8.4.3 Is information linguistically arranged within BERT representations?

Once we have investigated the relationship between the linguistic knowledge implicitly encoded by BERT and the number of individual units involved in it, we verified whether we can identify groups of units particularly relevant for specific probing tasks. To this end, we clustered the 68 probing features according to the weights assigned by the regression models to each BERT hidden unit. Specifically, we perform hierarchical clustering using correlation distance as distance metric. Figure 8.15 and 8.16 report the hierarchical clustering obtained with the *[CLS]* and *Mean-pooling* internal representations at layers 12, 8 and 1. We chose layers 12 and 1 in order to study differences of the clustering of linguistic features taking into account the representations that were more distant and more closer to the language modeling task respectively, while layer 8 was chosen since it was the layer after which BERT's representations tend to lose their precision in encoding our set of linguistic properties.

As a general remark, we can notice that, despite some variations, the linguistic features are organized in a similar manner across the tree layers and for both the configuration. This is to say that, despite the number of non-zero coefficients varies significantly between layers and according to the strategy for extracting the internal representations, the way in which linguistic properties are arranged within BERT embeddings is quite consistent. This suggests that there is a coherent organization of linguistic features according to non-zero coefficients that is independent from the layer and the aggregation

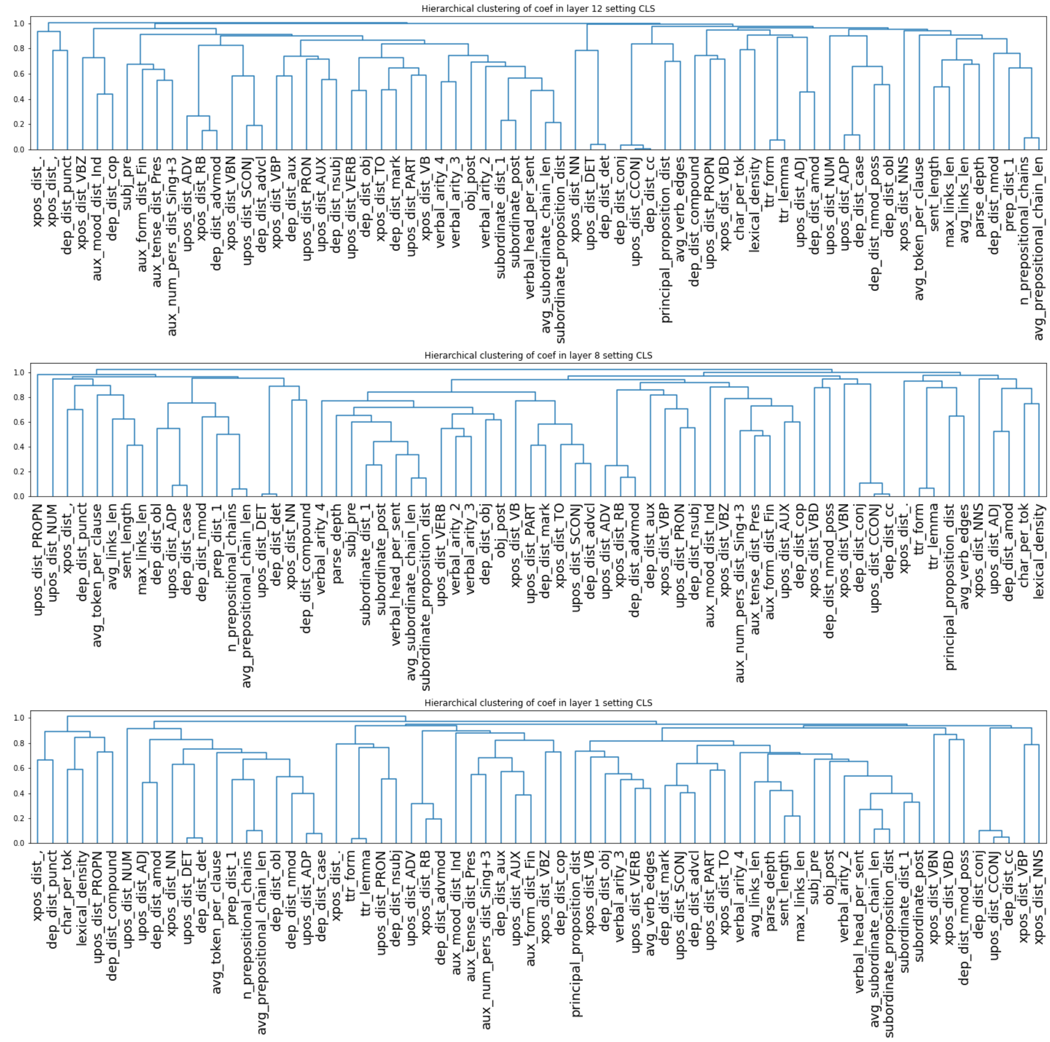


Figure 8.15: *From top to bottom, the hierarchical clustering for the* [CLS] *setting of all the tasks respectively at layers 12, 8 and 1.*

techniques taken into account.

Focusing on specific groups of features, we observe that, even if the traditional division with respect to the linguistic annotation levels (see Table 8.1) has not been completely maintained, it is possible to identify different clusters of features referable to the same linguistic phenomena for all the 3 layers taken into account and for both configurations. In particular, we can clearly observe groups of features related to the length of dependency links and prepositional chains (e.g. *max_links_len*, *avg_links_len*, *n_prepositional_chains*), to vocabulary richness (*ttr_form*, *ttr_lemma*), to properties related to verbal predicate structure and inflectional morphology of auxiliaries (e.g. *xpos_dist_VBD*, *xpos_dist_VBN aux_form_dist_Fin*, *aux_tense_dist_pres*) and to the use of punctuation (*xpos_dist_.*, *xpos_dist_,*, *dep_dist_punct*) and subordination (e.g. *subordinate_dist_1*, *subordinate_post*). Interestingly enough, BERT representations also tend to put together features related to each other but not necessarily belonging

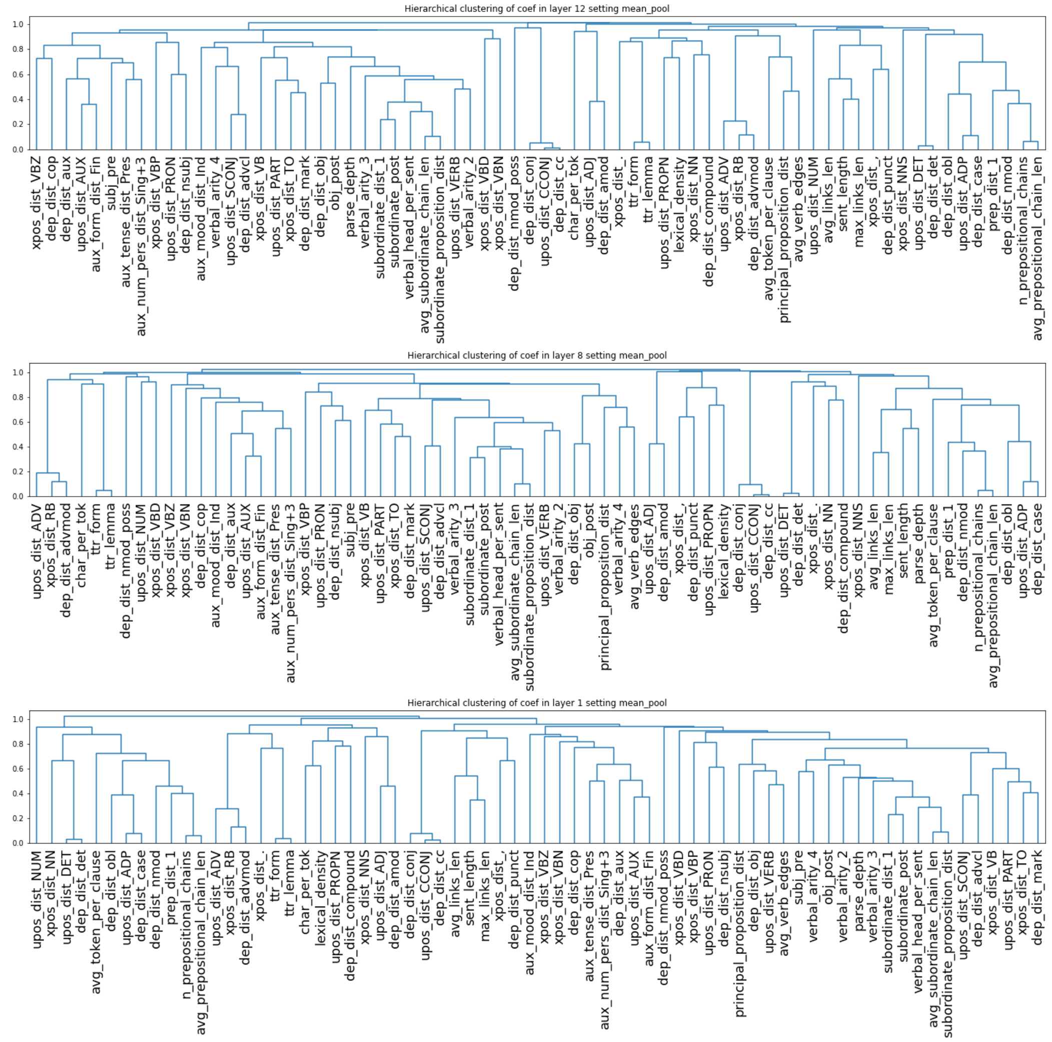


Figure 8.16: *From top to bottom, the hierarchical clustering for the* Mean-pooling *setting of all the tasks respectively at layers 12, 8 and 1.*

to the same linguistic macro-category. This is the case, for instance, of characteristics corresponding to functional properties (e.g. *upos_dist_ADP*, *dep_dist_det*).

8.5 Italian Transformers Under the Lingusitic Lens

While the vast majority of the works aimed at understanding the linguistic competence of NLMs have focused on models trained on the English language, as we already discussed in Chapter 4, relatively little work has been done to understand the inner working of non-English models. Starting from this premise, we decided to apply our methodology in order to carry out an in-depth investigation of the linguistic knowledge implicitly encoded by 6 Italian monolingual models and multilingual BERT. Besides the focus on Italian, which represents a scarcely considered language in the scenario of the NLM interpretation studies, a further novelty of our approach the comparative analysis of how and to which extent the different architectures on which the probing model rely on influence the probing accuracy. To address this point, for each Transformer, we perform the same suite of probing tasks using both a LinearSVR and a multilayer perceptron (MLP), and compare whether and how each probing task's resolution is affected by the two architectures. Since all experiments were carried out on different sections of Italian Universal Dependency Treebank [Nivre et al., 2016], we were also able to investigate how linguistic knowledge of NLMs varies according to different textual genres and language varieties.

To the best of our knowledge, this is the first study aimed at comparing the linguistic knowledge encoded in the representations of multiple non-English pre-trained transformer models. In particular:

- we compared the probing performances of 6 Italian NLMs and a multilingual one spanning three models over multiple linguistic feature categories;
- we investigated whether and how using different architectures of probing models affects the performance of transformers in encoding specific features;
- we showed how the implicit knowledge learned by these models differs across textual genres and language varieties.

8.5.1 Approach

To inspect the inner knowledge encoded by Italian Transformers, we relied on a suite of 82 probing tasks and we tested our approach testing two different probing architectures: a LinearSVR and a three-layer feedforward network with ReLU activations (Multi-layer perceptron, MLP). If the linear architecture is the most commonly used approach to infer information inside NLMs, the MLP was selected to investigate the presence of nonlinear relations in representations, which could hamper the probing performance of the LinearSVR probe. Regardless of the architecture, the two probing models take as input layer-wise sentence-level representations extracted from the Italian models. These representations are produced for each sentence of different sections of the Italian Universal Dependency Treebank (IUDT), version 2.5 [Zeman et al., 2019], and used to predict the actual value of each probing feature. Starting from the results obtained we performed three complementary investigations. In the first one we compared the results obtained by the two probing architectures according to different groups of probing tasks. Then, we move to compare the linguistic competence of the 7 Italian Transformers. Finally, the impact of the considered linguistic varieties on the NLMs linguistic generalization abilities is discussed.

Models and Data

We relied on 7 pre-trained Italian models based on three different Transformers architectures: BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019b] and GPT-2 [Radford

Table 8.9: *NLMs used in the experiments.*

Name	Training data
BERT Architecture	
Multilingual-BERT	Wikipedia
BERT-base-italian	Wikipedia + OPUS (13GB)
AlBERTo	TWITA (191GB)
RoBERTa Architecture	
GilBERTo	OSCAR (71GB)
UmBERTo-Commoncrawl	OSCAR (69GB)
UmBERTo-Wikipedia	Wikipedia (7GB)
GPT-2 Architecture	
GePpeTto	Wikipedia + ItWAC (14GB)

et al., 2019]. In particular, we investigated the linguistic competence of: three BERT-based models, Multilingual-BERT, BERT-base-italian[1] and AlBERTo [Polignano et al., 2019], trained respectively on Wikipedia (102 languages), Italian Wikipedia + texts from the OPUS corpus [Tiedemann and Nygaard, 2004] and TWITA [Basile et al., 2018]; three RoBERTa-based models, GilBERTo[2] and two versions of UmBERTo[3], trained respectively on OSCAR [Suárez et al., 2019] (GilBERTo and UmBERTo-Commoncrawl) and Italian Wikipedia (UmBERTo-Wikipedia); a GPT-2 based model, GePpeTto [De Mattei et al., 2020], trained on Italian Wikipedia + ItWAC [Baroni et al., 2009]. Models statistics are reported in Table 8.9. Sentence level representations were computed performing a *Mean-pooling* operation over the word embeddings provided by the models.

NLM's linguistic competences are probed against 5 sections of the Italian treebank representative of different language varieties and textual genres, as shown in Table 8.10. The considered sections can be categorised in two main groups: a first one that includes sentences acquired from documents of diverse nature, ranging from Wikipedia pages, to newspaper articles, novels, speech transcriptions, etc., and a second group collecting examples of the social media language, in particular of Twitter. In the first group we included the Italian version of the multilingual Turin University Parallel Treebank (ParTUT) [Sanguinetti and Bosco, 2015b], the Venice Italian Treebank (VIT) [Delmonte et al., 2007] and Italian Stanford Dependency Treebank (ISDT) [Bosco et al., 2013], which we considered representative of the standard Italian language. The group of treebanks composed of PoSTWITA [Sanguinetti et al., 2018] and TWITTIRÒ [Cignarella et al., 2019] was originally built to enhance the performances of systems in processing social media texts, and in particular, for irony detection purposes. Being representative of a non-standard variety of the Italian language, for our specific scopes, they are intended to be a quite challenging testbed for probing the linguistic knowledge of NLMs also when they are trained on standard language variety.

Note that the linguistic abilities of the 7 NLMs were also tested against a number of

[1] https://github.com/dbmdz/berts
[2] https://github.com/idb-ita/GilBERTo
[3] https://github.com/musixmatchresearch/umberto

Table 8.10: *Sections of the Italian Universal Dependency Treebank (IUDT).*

Short Name	Types of texts	# sent
ParTUT	Multi-genre	2,090
VIT	Multi-genre	10,087
ISDT	Multi-genre	14,167
ISDT_tanl	Newswire	4,043
ISDT_tut	Legal/Newswire/Wiki	3,802
ISDT_quest	Interrogative sentences	2,162
ISDT_2parole	Simplified Italian news	1,421
ISDT_europarl	EU Parliament debates	497
PoSTWITA	Tweets	6,713
TWITTIRÒ	Ironic Tweets	1,424
Total		35,481

sub-portions of the largest Italian UD treebank, i.e. ISDT. They have been chosen since they are representative of language sub-varieties possibly infrequently seen during the NLMs training phase. Accordingly, they can be conceived as a favorite point of view to investigate whether general-purpose NLMs encode less standard linguistic phenomena. For this purpose, in addition to sub-sections including newspapers (ISDT_tanl) and miscellaneous documents (ISDT_tut), we considered sub-portions including sentences in interrogative form (ISDT_quest), newspaper articles specifically written to be linguistically simple (ISDT_2parole) and transcriptions of the European parlament oral debates (ISDT_europarl).

The linguistic features used for testing the linguistic knowledge of Italian Trasformers are, once again, based on the ones described in [Brunato et al., 2020] and extracted from the sentences of the IUDT treebank. Specifically, for the experiments we relied on a subset of 83 features, that can be grouped in the 9 macro-groups earlier discussed.

8.5.2 Experiments and Results

In this section we report the results of the three different investigations we carried out starting from the probing strategies devised.

Comparison of Probing Model Architectures Our first analysis concerns the comparison of the two considered architectures for probing the linguistic knowledge encoded by the Italian Transformers. Since many of our probing features are strongly related to sentence length, we compared these results with the ones obtained by a baseline corresponding to a LinearSVR model trained using only sentence length as input feature. Table 8.11 reports average (layer-wise) R^2 results[4] for all the 7 NLMs obtained with the LinearSVR and the MLP probing architectures, along with baseline scores.

As a first remark, we notice that both probing architectures outperform the baseline. This suggests that all NLMs encode a spectrum of phenomena that, although related

[4] The Coefficient of determination (R^2) is a statistical measure of how close the data are to the fitted regression line and corresponds to the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

Table 8.11: *Average R^2 scores for all the NLMs obtained with the LinearSVR and the MLP probing models. Baseline scores are also reported.*

Groups	LinearSVR	MLP	Baseline
RawText	**0.84**	0.80	0.50
Vocabulary	**0.70**	0.34	0.19
POS	**0.69**	0.68	0.03
VerbInflection	0.50	**0.61**	0.03
VerbPredicate	0.32	**0.43**	0.08
TreeStructure	0.61	**0.64**	0.40
Order	0.46	**0.55**	0.06
SyntacticDep	0.65	**0.74**	0.04
Subord	0.49	**0.60**	0.16
AllFeatures	0.60	**0.64**	0.10

to sentence length, require a more sophisticated linguistic knowledge to be accurately predicted. However, if we compare the results achieved by the two architectures on all groups of linguistic phenomena (*AllFeatures*), we can see that MLP architecture achieves higher R^2 scores. This is specifically the case of the group of features which refer to distribution of syntactic dependency relations (*SyntacticDep*) and to complex aspects of sentence structure (*TreeStructure*). On the contrary, if we compare the ranking of the linguistic phenomena ordered by decreasing scores, we can see that for both architectures raw text properties and the distribution of morpho-syntactic categories (*POS*) appear in the first positions, while the order of subject and object (*Order*) and the structure of verbal predicates (*VerbPredicate*) are ranked in the lower part of the ranking. Interestingly enough, the LinearSVR architecture outperforms the MLP by more than .30 R^2 points when predicting features related to vocabulary richness (*Vocabulary*).

In order to ensure that our probes are actually showing the linguistic generalization abilities of the NLMs rather than learning the linguistic tasks, we also tested the probing models using the *control task* approach devised in [Hewitt and Liang, 2019]. We produced a control version of the IUDT corpus by randomly shuffling the linguistic features assigned to each sentence and performed the same probing tasks with the two probing classifiers for all NLMs representations. The correlation and R^2 scores between regressors' predictions and shuffled scores were low (< 0.05) and comparable for both the SVR and the MLP. These results support the claim that NLMs representations encode information closely related to linguistic competence and that our probing models are not relying on spurious signals unrelated to our linguistic properties to solve the regression task.

Comparison of Italian Transformers To investigate to which extent each transformer encodes the considered set of linguistic phenomena, we compared the performances achieved by the 7 NLMs, using the two probing architectures. Results are reported in Figure 8.17 where we can notice that the 7 transformers achieve quite similar results

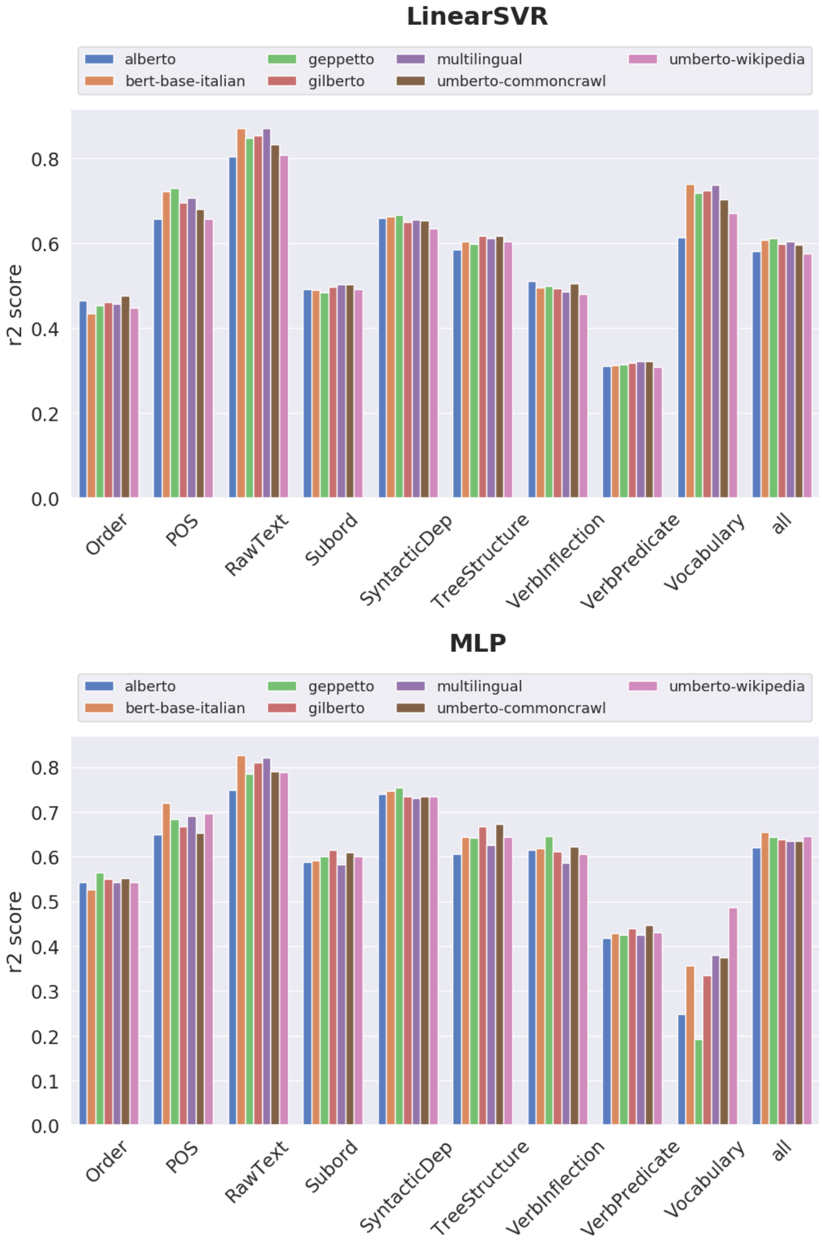


Figure 8.17: *Average (layer-wise) R^2 scores obtained by each NLM with the two probing models.*

when considering all features as a whole. Nevertheless, a more in depth analysis highlights a number of small differences. Namely, we can see that BERT-base-italian and GePpeTto are among the first three best models, while AlBERTo is among the two models resulting less able to encode the sentence linguistic properties. If this can be observed for both the probing architectures, a main difference concerns the performances achieved by the UmBERTo model trained on the Italian Wikipedia: it is the second best model using the MLP architecture while it represents the last one with the SVR probing architecture.

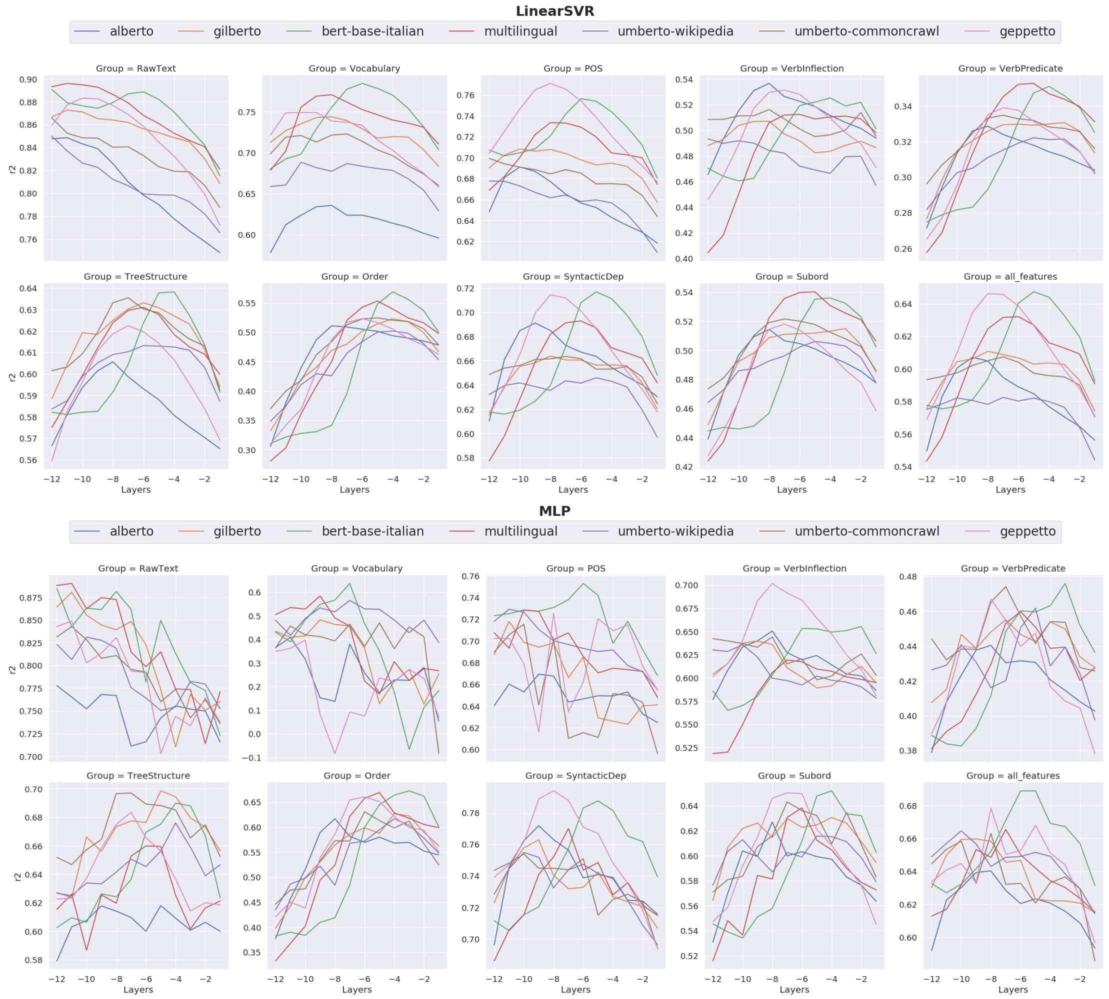


Figure 8.18: *Average layerwise R^2 scores obtained with the LinearSVR (top) and the MLP (bottom) using the internal representations of the 7 NLMs.*

However, this trend does not hold when we analyse the NLMs performances with respect to the encoding of the different groups of linguistic phenomena. For instance, we can notice that, for both the probing architectures, tree structure properties (*TreeStructure*) are predicted more accurately by RoBERTa-style models, i.e. by GilBERTo and UmBERTo-Commoncrawl, than by models based on BERT or GPT-2. This can be similarly observed for the prediction of two other linguistic properties referring to sub-trees of the whole syntactic structure of a sentence. Namely, it can be seen that GilBERTo and UmBERTo-Commoncrawl are the two best models able to encode the use of subordination (*Subord*) and the verb predicate structures (*VerbPredicate*). However, this holds only if we consider the MLP probing architecture. Further differences in terms of probing architectures can be inspected considering NLMs abilities to encode competencies related to vocabulary richness (*Vocabulary*): while UmBERTo-Wikipedia extensively outperforms all the other transformers using the MLP model, the best transformer is BERT-base-italian when these competences are probed with the LinearSVR model.

Additional observations can be made if we move to the analysis of how NLMs prediction abilities change and evolve across layers. As it can be seen in Figure 8.18, regardless of the architecture, for all transformers linguistic competences tend to decrease across the 12 layers. This is in line with previous findings [Liu et al., 2019a, Miaschi et al., 2020a] and it could be due to the fact that transformer layers trade off between task-oriented (e.g. Masked Language Modeling) information and general linguistic competence. Such decreasing trend can be specifically observed for example for the ability to predict raw text features, or the distribution of the UD morpho-syntactic categories (*POS*) and syntactic dependencies (*SyntacticDep*): they represent sentence properties mainly encoded in the first layers by all NLMs. On the contrary, we can observe that there is a number of more complex linguistic features whose knowledge increases consistently across layers, even if it decreases in the output layer. This is the case of features referring to structural sentence knowledge, such as the order of subject/object with respect to the verbal head (*Order*) and the use of subordination (*Subord*). In addition, contrarily to what was observed by [de Vries et al., 2020], Mulilingual-BERT's linguistic knowledge is not encoded systematically earlier than in monolingual transformers.

This perspective of analysis also reveals other differences among the considered transformers which were unseen. By inspecting the trend of the R^2 scores across layers, we can for example see that even though GePpeTto has a lower average competence on verb inflection (see Figure 8.17), it achieves the highest scores in the middle layers. Or, even if we previously noted that RoBERTa-style transformers are more able to predict features related to the structure of a sentence (*TreeStructure*), the highest accuracy is achieved by a BERT-style model, i.e. BERT-base-italian, in the -4 layer. A similar observation also concerns the use of subordination and the verb predicate structure: the two group of features are in general predicted more accurately by GilBERTo and UmBERTo-Commoncrawl but the highest R^2 scores are achieved by Mulilingual-BERT and BERT-base-italian in the -5 and -4 layers.

Focusing instead on differences between layerwise scores obtained by the two probing architectures, we can clearly notice that the encoding of linguistic knowledge shows a quite rough trend for what concerns the results obtained with the MLP. This is particularly the case of features belonging to the vocabulary, POS and tree structure groups.

If we deepen our investigation and we focus on the linguistic generalization ability of the NLMs with respect to each individual feature (see Figure 8.19), we can clearly observe that the rankings according to R^2 scores are quite similar regardless the probing architecture and the transformer model. It is also interesting to note that, despite some deviations, the distinction into macro-groups of linguistic phenomena seems to be preserved across the rankings. In fact, raw-text features, as well as the distributions of POS-tags (*upos_dist_**, *xpos_dist_**) and dependency relations (*dep_dist_**), are those that were better predicted by the two probing models, while features more related to

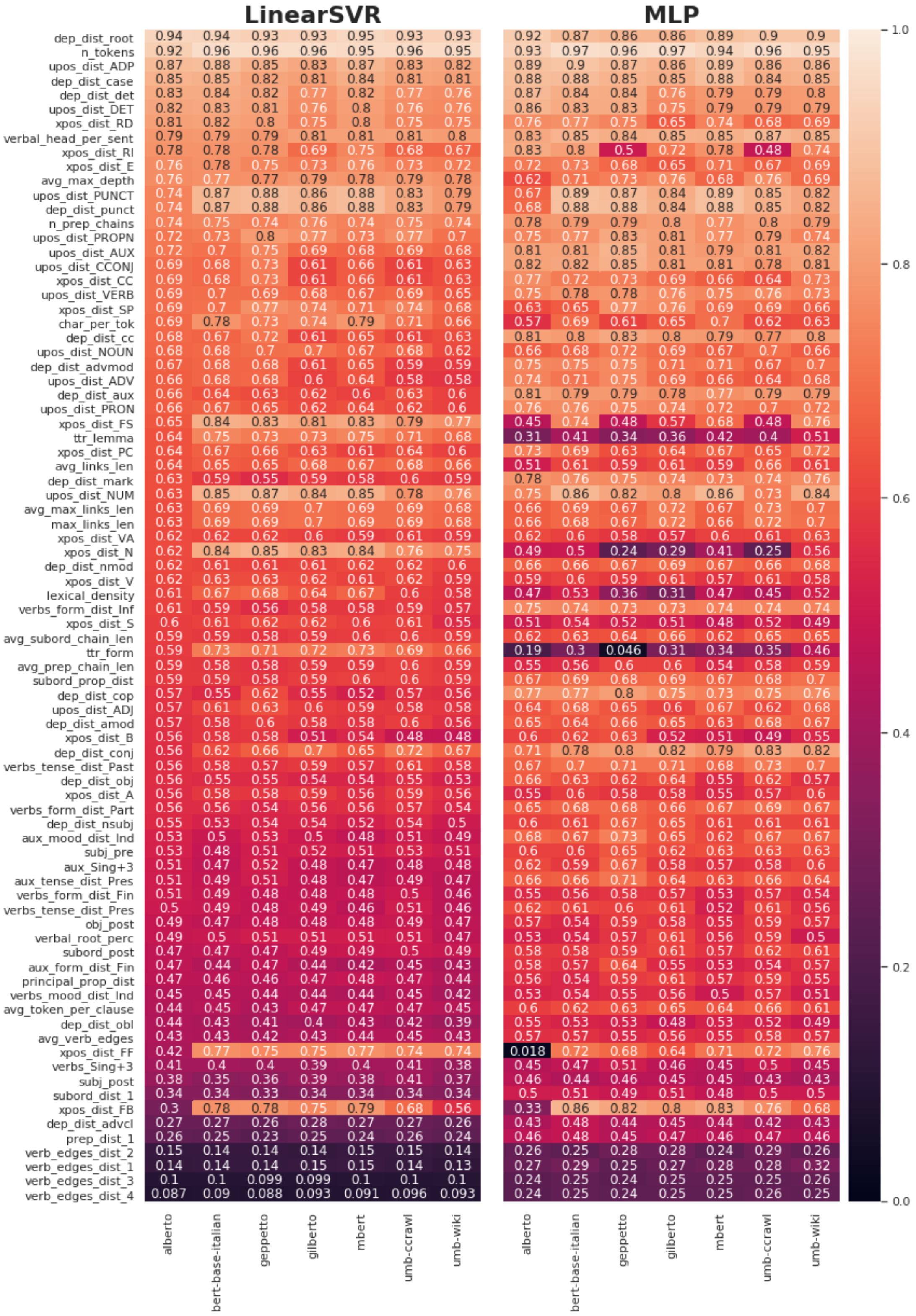


Figure 8.19: *Average R^2 scores obtained for each probing features using the two probing architectures tested with the internal representations of the 7 NLMs. Both heatmaps are ordered on the basis of the feature ranking as predicted by the AlBERTo model using the LinearSVR architecture.*

the structural information of a sentence, such as the order of elements (e.g. *subj_pre*, *subj_post* and *obj_post*) or the structure of parsed tree (e.g. *avg_token_per_clause*, *avg_prep_chain_len*) achieved lower probing scores. Worse predictions are also related to morphological features of both lexical and auxiliary verbs, namely for example their mood (*verb_mood_**) or tense (*verb_tense_**). Taking a closer look at the differences between the 7 NLMs, we can see that in few cases the linguistic competence of the AlBERTo model is significantly different (lower) from that of the other models. The most remarkable case concerns the distribution of punctuation marks in general, both at the level of morpho-syntactic category (*upos_dist_PUNCT*) and of dependency relation (*dep_dist_punct*), as well as the distribution of commas (*xpos_dist_FF*) and balanced punctuation (*xpos_dist_FB*). This appears particularly evident using MLP as probing architecture and it is possibly related to the typology of texts the AlBERTo model was trained on, i.e. Twitter. It is well known that social media represents a not standard language variety, characterised by specific linguistic properties mostly different from ordinary language [Farzindar and Inkpen, 2015], such as short sentences where punctuation marks, especially weak ones, are rarely used. Accordingly, the low frequency of punctuation in the training corpus possibly yields AlBERTo's reduced generalization abilities with respect to this specific set of features.

Comparison of Italian Language Varieties Our last analysis concerns the impact of the considered Italian language varieties on NLMs linguistic abilities. For this purpose, we inspected whether the overall linguistic competence encoded in the contextual representations of each model changes according to the different IUDT sections. The results reported in Figure 8.20 show that all transformers, regardless of the probing architecture, achieve lower performance when they have to predict the value of features extracted from treebanks representative of social media language (PoSTWITA and TWITTIRÒ) and from the sub-set of ISDT sentences in interrogative form (ISDT_quest). In both cases, this seems supporting our starting intuition that NLMs trained on standard language varieties, represented for example by Wikipedia pages, websites or web-crawled documents, may be less robust to non-standard varieties that were possibly unseen, or rarely seen, during the pre-training process. Quite surprisingly, even if AlBERTo has been trained on Twitter data, it obtains the lowest R^2 scores also when its internal representations are used to predict the feature values of the two social media Italian treebanks. A possible explanation is that, although PoSTWITA and TWITTIRÒ contain sentences representative of Twitter language, these sentences are still quite close to the Italian standard language, in order to be compliant with the UD morpho-syntactic and syntactic annotation schema. On the contrary, AlBERTo's training set is derived from Twitter's official streaming API that included all possible typologies of sentences.

It also worth noting that BERT-base italian and GePpeTto are the two models slightly less affected by the non-standard linguistic peculiarities of the social media variety. As noted in Section 8.5.2, they represent the two best performing models in terms of overall linguistic competence. This may explain why they are more robust in the

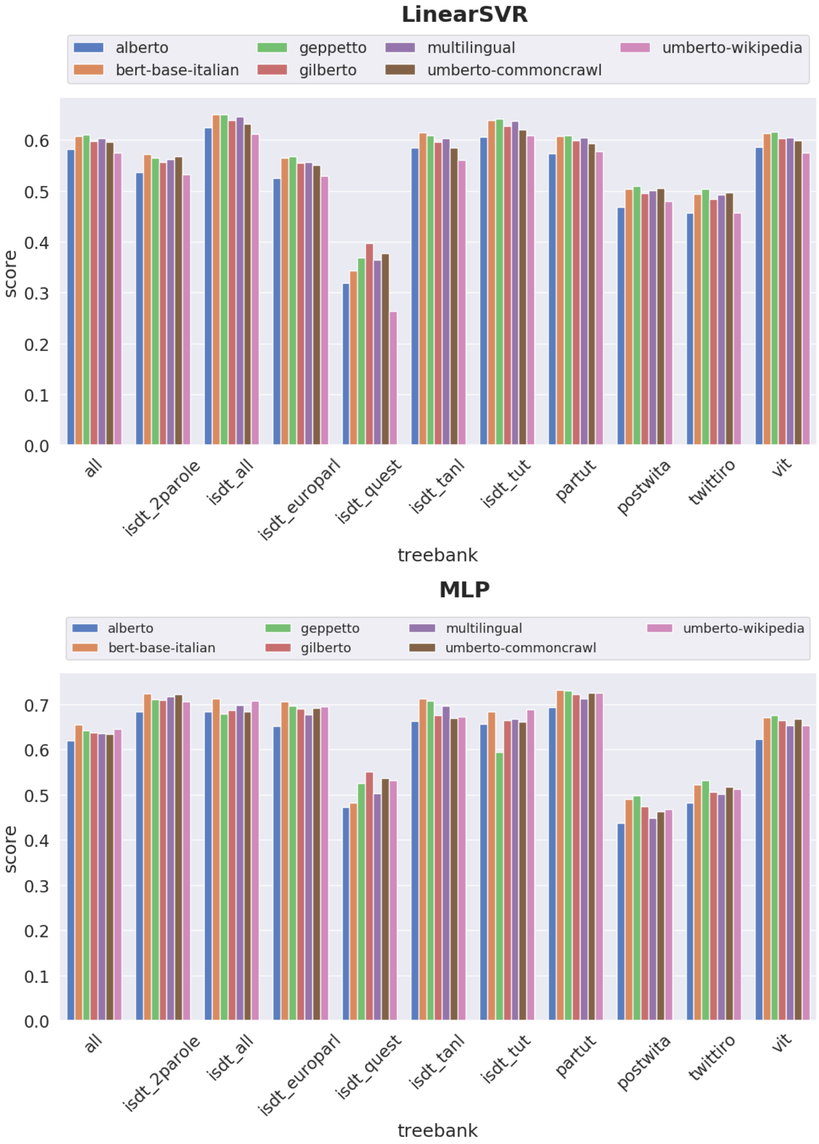


Figure 8.20: *Average LinearSVM R^2 score considering all the UD Italian sentences (all) and according to the 10 treebanks previously described.*

accurate prediction of the features values of all the considered IUDT sub-sections. This holds both with the LinearSVR and MLP probing architecture, even if in the latter case the two versions of UmBERTo achieve comparable or slightly better scores. A main exception is represented by the ISDT sub-section including sentences in interrogative form (ISDT_quest), which, as we noted above, are hardly mastered by all models. This is possible due to the fact that interrogative sentences are more likely to display a less canonical distribution of morpho-syntactic and syntactic phenomena, hence being more difficult to encode effectively. In this case, the transformer based on GPT-2, i.e. GePpeTto, results to be the NLM with the highest linguistic knowledge of this type of sentences.

Table 8.12: *Spearman correlations between rankings of features as predicted by the 7 NLMs on four sections of the IUDT treebank: IUDT_2parole (2par), IUDT_tanl (tanl), IUDT_quest (quest) and IUDT_postwita (ptw). Highest correlations are bolded, while lowest ones are marked in italics.*

Model	Section	LinearSVR				MLP			
		2par	tanl	quest	ptw	2parole	tanl	quest	ptw
alberto	2par	1				1			
	tanl	.72	1			**.85**	1		
	quest	*.38*	*.38*	1		.62	*.56*	1	
	ptw	.76	**.82**	.45	1	.75	.80	.58	1
bert-base-italian	2par	1				1			
	tanl	.68	1			.82	1		
	quest	*.34*	.41	1		.62	*.47*	1	
	ptw	.72	**.91**	.47	1	.75	**.88**	*.47*	1
geppetto	2par	1				1			
	tanl	.65	1			.80	1		
	quest	*.30*	.38	1		.64	50	1	
	ptw	.70	**.92**	.48	1	.72	**.88**	*.47*	1
gilberto	2par	1				1			
	tanl	.61	1			.77	1		
	quest	*.30*	.40	1		.58	54	1	
	ptw	.66	**.88**	.46	1	.69	**.82**	*.49*	1
mbert	2par	1				1			
	tanl	.65	1			.76	1		
	quest	*.30*	.37	1		.55	.47	1	
	ptw	.71	**.90**	.45	1	.71	**.83**	*.46*	1
umberto-commoncrawl	2par	1				1			
	tanl	.58	1			.71	1		
	quest	*.28*	.33	1		.55	.47	1	
	ptw	.69	**.8**	.39	1	.65	**.75**	*.35*	1
umberto-wikipedia	2par	1				1			
	tanl	.57	1			.70	1		
	quest	-	-	1		.50	.44	1	
	ptw	.66	**.72**	*.36*	1	.69	**.72**	*.36*	1

A further analysis of the impact of language varieties on the ability of NLMs to encode the considered group of linguistic phenomena can be appreciated in Table 8.12. It shows, for each probing architecture, the Spearman correlations between the rankings of features predicted by all NLMs considering three ISDT sub-sections, i.e. ISDT_tanl, ISDT_2parole and ISDT_quest, and PoSTWITA, and ordered by decreasing R^2 scores. For each NLM, higher correlations correspond to similar linguistic generalization abilities across the paired treebanks, while lower correlations suggest that the inner representations of the NLM allow predicting effectively diverse linguistic features. As we can see, regardless of the probing architecture, for all NLMs, the highest correlated rankings are those obtained comparing ISDT_tanl and PoSTWITA predicted features. Even if it is quite surprising, this result can be explained assuming that the morpho-syntactic and syntactic features of the Twitter sentences contained in PoSTWITA are not so dramatically different from those characterising ISDT_tanl newspaper articles. In fact, among all the three ISDT sections considered here they resulted to be the two most similar treebanks with respect to the distribution of the set of linguistic features reported in Table 8.1. In particular, the main differences concern the distribution of some morpho-syntactic categories (e.g. punctuation, nouns) and main features related to the inflectional morphology of verbs, e.g. the distribution of present tenses, higher in PoSTWITA (51.11% out of the total verb tenses) than in ISDT_tanl (34.95%), or of the past tenses that in the Twitter sentences are less than half than in the newspaper ones. Interestingly, these characteristics belong to the group of features that the NLMs are able to master quite accurately, regardless of the language variety. Even if these differences had a negative impact on the overall probing abilities of the PoSTWITA sentence characteristics, as shown in Figure 8.20, the higher knowledge of these specific features did not possibly have a great consequence on the ranking of the predicted features, thus yielding quite high correlations.

On the contrary, the lowest correlations can be observed when we compare the rankings obtained for the pairs of treebanks containing the set of sentences in the interrogative form, i.e. ISDT_quest. Even if the correlation values are slightly higher using MLP, this trend holds for the two probing architectures and for all NLMs. Note that the correlations between the ranking obtained with UmBERTo-Wikipedia for the pairs ISDT_quest/ISDT_2parole and ISDT_quest/ISDT_tanl are even not statistically significant. Let us remind that this is the NLM that achieved the lowest prediction accuracy using the LinearSVR probing architecture (see Figure 8.17). Our intuition is that this may have made it less robust in the prediction of non-standard linguistic forms, such as interrogative sentences. Similarly to what aforementioned, these results can be explained if we analyse the feature values in the considered treebanks. ISDT_quest resulted to be quite different from all the other treebanks particularly with respect to complex aspects of sentence structure. For example, the canonical order of the nuclear elements of a sentence (i.e. subject and object) is largely subverted in sentences in the interrogative form. Thus, they contain a very high percentage of post-verbal explicit subjects (68.69% of the total), half an order of magnitude higher than ISDT_tanl (15.21%) and PoSTWITA (12.63%) and an order of magnitude higher than ISDT_2parole (7.55%). Sentences in the interrogative form also have a lower percentage of post-verbal objects (17.31%), which instead represent the majority of cases in other treebanks, and they are characterised by a very low distribution of subordinate clauses in general and in particular of subordinates following the principal clause, i.e. 4% vs. 43% in ISDT_tanl, 35.78% ISDT_2parole and 44.36%. These and other similar features all concern structural aspects of a sentence that may have undermined the overall NLM linguistic competence thus yielding not only lower probing scores on ISDT_quest but also different feature rankings with respect to the other treebanks.

8.6 Probing Tasks Under Pressure

In Chapter 4 we showed that, despite the emerging amount of work, there are several open questions concerning the design of probing tasks and that those questions fostered complementary lines of research. Among these, a number of studies have started to investigate the effectiveness of the probing paradigm, as in [Ravichander et al., 2021]. Starting from the *control tasks* approach defined in [Hewitt and Liang, 2019], we introduced a new approach to put increasingly under pressure the effectiveness of a suite of probing tasks to test the linguistic knowledge implicitly encoded by BERT [Devlin et al., 2019]. To achieve this goal, we set up a number of experiments aimed at comparing the probing results obtained by Italian BERT when predicting i) our set of linguistic features extracted from the the Italian Universal Dependency Treebank [Zeman et al., 2019] ii) a set of linguistic features built from a suite of *control datasets* we specifically built for the purpose of this study. We define a control dataset as a set of linguistic features whose values were automatically altered in order to be increasingly different from the values in the treebank, referred to as *gold* values. Our underlying hypothesis is that if the predictions of the altered values diverge from the predictions of the gold values, this possibly suggests the effectiveness of probing tasks to test the linguistic knowledge embedded in BERT representation.

To the best of our knowledge this is the first paper that:

- introduces a methodology to test the reliability of probing tasks by building control tasks at increasing level of complexity
- puts under pressure the probing approach considering the Italian language.

8.6.1 Methodology

Our methodology seeks to investigate the effectiveness of probing tasks for evaluating the linguistic competences encoded in NLM representations. To this aim, we trained a LinearSVR model using BERT sentence representations and then tested its performances when predicting the values of a set of linguistic features in multiple scenarios. In one scenario, the model shall predict gold values, thus corresponding to the real values of the features in the corpus. In the other scenarios, we automatically altered the feature values at different control levels each corresponding to increasing degrees of pressure for the probing model.

Our methodology will allow us to test whether the probing model really encodes linguistic competences or simply learns regularities in the task and data distributions by checking the results obtained in the different scenarios. If the predictions of the probing model will be more similar to the gold values than to the automatically altered ones, then we might assume that the information captured by the probed feature is encoded in the representations.

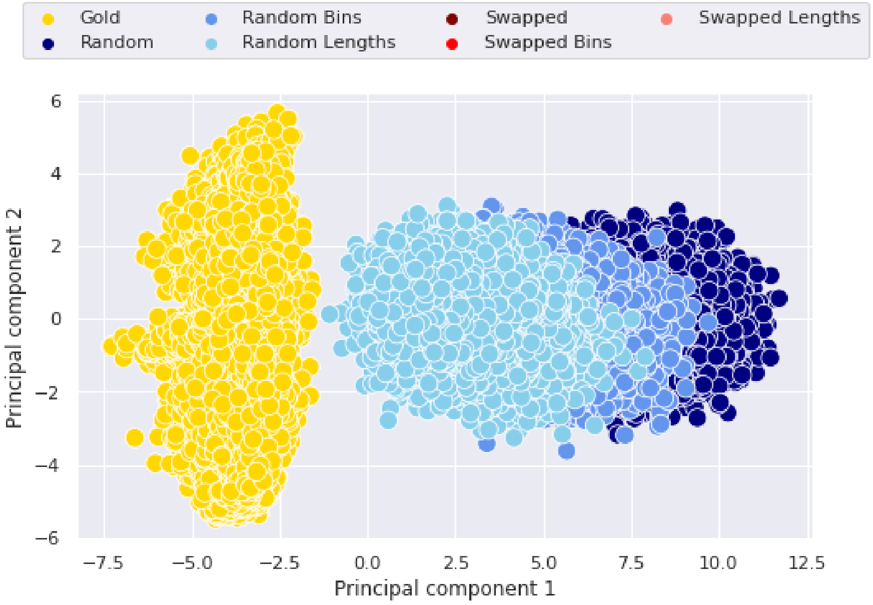


Figure 8.21: *2-dimensional PCA projection of the feature values in the gold and control datasets. All Swapped datasets overlap with the Gold one.*

Models and Data We tested our approach of the base cased Italian BERT developed by the MDZ Digital Library Team[5]. For the sentence-level representations, we leveraged the activation of the first input token *[CLS]*.

Our experiments are carried out on the Italian Universal Dependencies Treebank (IUDT), containing a total of 35,480 sentences. Due to the IUDT high variability in terms of sentence length[6], we focused on a sub-set of sentences with a ± 10 tokens variation with respect to the median sentence length (i.e. 20 tokens). As a result, we selected 21,991 sentences whose length ranges between 10 and 30 tokens. We chose to consider only this sub-set since all groups of sentences of the same length included in this interval are composed by an amount of elements, i.e. 1,000, which makes our results reliable and comparable across groups of different lengths.

Starting from the set of linguistic features devised in 8.5, we relied on a subset of 77 features, modeling 7 main aspects of the structure of a sentence: morphosyntactic information, inflectional morphology, verbal predicate structure, global and local parsed tree structures, relative order of elements, syntactic relations and use of subordination.

Control datasets We created two main types of control datasets, obtained by automatically altering gold feature values. The first main type (hereafter referred to as *Swapped*) is built by shuffling the original values of each feature across sentences; while the second type (*Random*) contains values randomly generated within the maximum and the minimum value that each feature shows in the whole gold dataset. Since the values

[5] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

[6] IUDT contains sentences ranging from 1 to 308 token long.

of the considered features are strongly related to the length of the sentence, for each type of control dataset we built two sub-types of datasets. In a first sub-type (*Bins*), we grouped sentences falling into the same predefined range of sentence lengths (i.e., 10-15, 15-20, 20-25 and 25-30 tokens). In a second sub-type (*Lengths*), we included groups of sentences having exactly the same length.

Note that the different data altering strategies are conceived to represent increasingly challenging testbeds to assess the effectiveness of our probing tasks. The *Swapped* control datasets are the most challenging ones as the swapped feature values might be quite similar to the gold ones, thus possibly predicted with an high accuracy by the probing model. Such intuition is confirmed by the results of the 2-dimensional Principal Component Analysis (PCA) reported in Figure 8.21[7]. As we can see, all the data points representing the feature values contained in the *Swapped* datasets fully overlap with the gold ones, thus confirming their similarity. On the contrary, randomly generated values are progressively more distant being less plausible, even if the constraints of sentence length yield values that are closer to the gold ones.

From now on, the values of each feature acquired from IUDT represent the *gold dataset* and they have been automatically altered in order to generate additional *control datasets*.

8.6.2 Results

For both gold and control datasets, probing scores are computed as a Spearman correlation between the feature values predicted by the probing model and the values contained in each dataset. Such correlation values are computed by averaging the NLM's layer–wise scores as, for all datasets, we observed small differences between the scores obtained across the 12 layers. We experimentally verified that these differences were not significant by computing the slope of a linear regression line between BERT layers and the scores of the gold dataset, obtaining -0.0017 as mean value considering all features. Our intuition is that the small range of lengths of the sentences here considered may have yielded such insignificant variation across layers, which turned out to be on the contrary significant on the whole set of IUDT sentences (see 8.5). Namely, being highly related to the length of the sentence, the feature values have little variations.

Figure 8.22 shows the scores obtained on the gold and the 6 control datasets, both for the 7 macro-groups of linguistic features and on average (*AVG*). Additionally, in order to properly appreciate the differences between the results obtained on the *gold* and control datasets, in Figure 8.23 we report the error reduction rate for each control dataset computed as the difference between the scores obtained when predicting gold and altered features.

General Results. We can observe that on average the highest probing scores are obtained on the gold dataset and that, accordingly, there is a great difference (i.e. almost 1.0, see

[7] PCA is a classical data analysis method that reduces the dimensionality of the data while retaining most of the variation in the data set by identifying *n* principal components, along which the variation of the data is maximal [Jolliffe and Cadima, 2016].

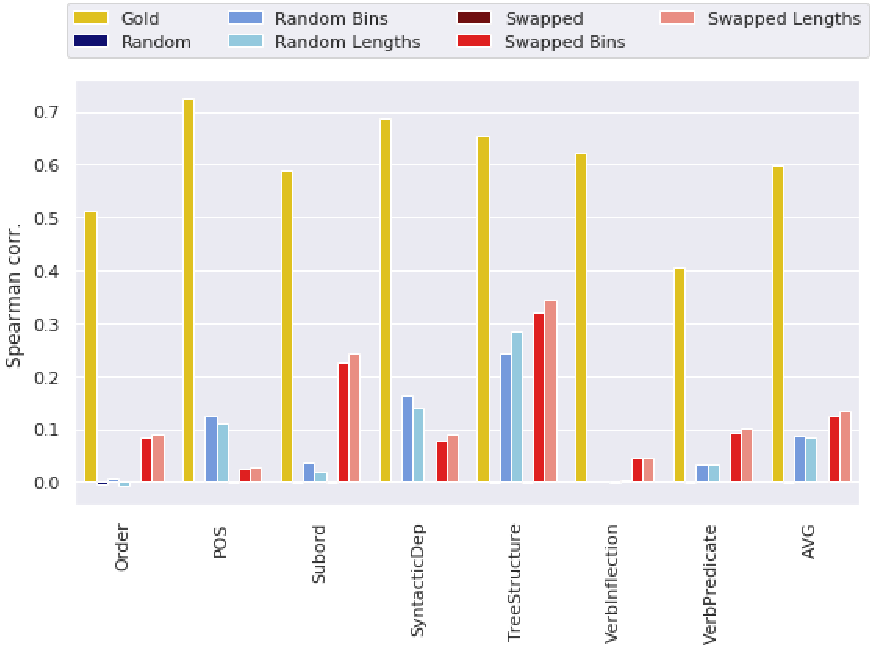


Figure 8.22: *Average probing scores (as Spearman correlation) obtained by the LinearSVR model when predicting gold and control linguistic features. Results are reported for each feature group and on average ('AVG' column).*

Fig. 8.23) between the accuracy of the probing model when predicting the authentic and altered feature values. This seems suggesting that the model is able to recognize that the feature values contained in the control datasets have been altered, even when they are not fully random but plausible, i.e. in the *Swapped* datasets. As a consequence, we can hypothesize that the model is relying on some implicit linguistic knowledge when it predicts the authentic feature values, rather than learning some regularities possibly found in the dataset.

However, if we take a closer look at the scores obtained for the *Random* and *Swapped* datasets when we constrain the length of the sentences, we can observe that the accuracy in predicting the feature values contained in the *Swapped* datasets is sightly higher than in the *Random* ones (see 'AVG' column in Figure 8.22). This is in line with our starting hypothesis and shows that feature values artificially created simply by shuffling gold ones across sentences of the same lengths (or of the same range of lengths) are more similar to the gold values and thus are predicted with higher accuracy than randomly altered values. Nevertheless, their error rate, namely the difference from the accuracy of gold predictions, is still quite high, i.e. about 0.80 (see the 'AVG' column, Figure 8.23).

Linguistic Features Analysis. Also when we focus on the results obtained with respect to the 7 macro-groups of linguistic features, we can observe that the probing model is more accurate in the prediction of the gold values. Again, the scores on the control datasets are slightly higher when we constrain the values with respect to sentence length, since

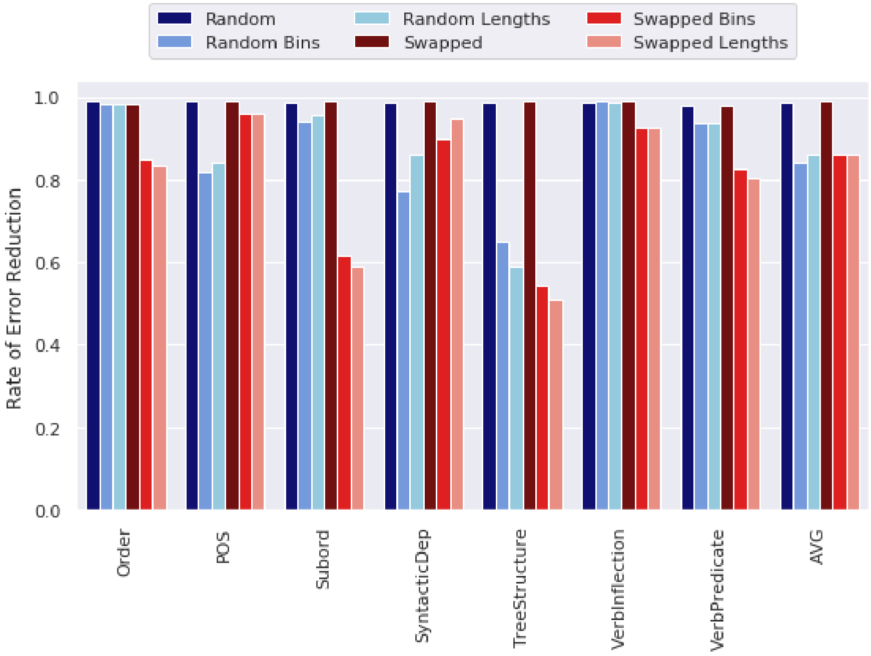


Figure 8.23: *Error reduction rates reporting the difference between the probing scores obtained on the Gold dataset and each control dataset. Result are reported for each feature group and on average ('AVG' column).*

Dataset	Spearman corr.
Random	0.08
Random Bins	0.46 *
Random Lengths	0.33 *
Swapped	-0.15
Swapped Bins	0.05
Swapped Lengths	0.06

Table 8.13: *Spearman correlations between the rankings of features obtained with the Gold dataset and the 6 control datasets. Statistically significant correlations are marked with * (p-value $<$ 0.05).*

we narrow the range of possible values. In particular, we see that the feature values related to the sentence tree structure are those predicted most closely to the gold ones (see column 'TreeStructure', Figure 8.23). Note that these sentence properties are the most sensitive to the sentence length, that BERT encodes with a very high accuracy. This may suggest that in the resolution of these tasks the probing model is possibly relying on some regularities related to sentence length. Similar observations hold for the results achieved in the resolution of the probing tasks related to the use of subordination, which heavily depends on sentence length. Interestingly, we can note that the values of all the other groups of features contained in the control datasets are predicted by the probing model with a very low accuracy, possibly making the results not significant.

Gold	Random Bins	Swapped Lengths
dep_dist_root	dep_dist_root	dep_dist_root
dep_dist_punct	avg_max_links_len	avg_max_links_len
upos_dist_PUNCT	max_links_len	max_links_len
xpos_dist_FS	xpos_dist_FB	avg_max_depth
upos_dist_ADP	avg_token_per_clause	verbal_head_per_sent
dep_dist_det	xpos_dist_FS	xpos_dist_FS
upos_dist_PROPN	n_prep_chains	avg_links_len
upos_dist_DET	avg_max_depth	subord_prop_dist
xpos_dist_RD	verbal_head_per_sent	avg_subord_chain_len
dep_dist_case	xpos_dist_RI	n_prep_chains
verbal_head_per_sent	dep_dist_cop	subord_post
xpos_dist_FF	xpos_dist_PC	subord_dist_1
xpos_dist_SP	dep_dist_conj	avg_prep_chain_len
xpos_dist_E	xpos_dist_B	obj_post
upos_dist_NOUN	xpos_dist_VA	avg_verb_edges

Table 8.14: *15 top-ranked Gold and control features (Random Bins and Swapped Lengths) predicted by BERT sentence-level representations.*

Features Correlations. Once we showed that the probing tasks accuracy is very different if the feature values are authentic or altered, in this section we compare the ranking of linguistic features ordered by decreasing prediction accuracy in the gold and control scenarios. As we can see in Table 8.13, which reports the Spearman correlations between the rankings, the *control rankings* are almost not related to the gold one and the existing correlations are not statistically significant. The only exceptions are represented by the rankings of values that were randomly generated with sentence length constraints, which have a weak and moderate correlation. Note that however, as shown before, the ranked scores are very low.

A more qualitative feature ranking analysis can be carried out by inspecting Table 8.14 where we report the first 15 top-ranked features predicted in the gold and in the two most highly correlated *Swapped* and *Random* datasets. As we can see, the *gold ranking* diverges from the rankings of the altered values with respect to the majority of top-ranked features. The most visible exception is represented by the distribution of syntactic root that the probing model always predicts with the highest accuracy. The result is quite expected since this feature can be seen as a proxy of the length of the sentence, a linguistic property properly encoded by BERT. Similarly, other two features influenced by sentence length appear, as expected, on the top positions of all rankings, namely the distribution of the sentence boundary punctuation (*xpos_dist_FS*) and of verbal heads (*verbal_head_per_highly*).

8.7 Discussion

As discussed in Chapter 4, the probing tasks approach is a natural way to estimate the mutual information shared by a neural network's parameters and some latent property that the model could have implicitly learned during training. Although the effectiveness and the reliability the approach is still under debate, previous works provided several insights about the linguistic properties implicitly learned by state-of-the-art NLMs. Using a suite of probing tasks inspired to the ’linguistic profiling’ methodology [van Halteren, 2004], in our experiments we showed that pre-trained Transformer-based models are capable of encoding a wide spectrum of linguistic proprieties of sentence structure.

Focusing our study on an English pre-trained version of BERT, we showed that this model is able to encode linguistic phenomena across its 12 layers, but, differently from previous studies (e.g. [Lin et al., 2019, Tenney et al., 2019a]), we found that the order in which probing features are stored in the internal representations does not necessarily reflect the traditional division with respect to the linguistic annotation levels. In line with previous work (e.g. [Liu et al., 2019a]), we noticed that in the last layers the linguistic competence encoded by the model tends to decrease, probably because BERT is getting more specified for the MLM task.

In a follow-up study dedicated to the comparison between a contextual and a non-contextual NLM, we noticed that word2vec acquire sentence-level linguistic competence in a similar way to BERT. More specifically, we showed that BERT is able in storing features that are mainly related to raw text and syntactic properties, while word2vec is good at predicting morphosyntactic characteristics. Moving instead from sentence-level to word-level representations, we learned that BERT encodes sentence-level linguistic phenomena even within single-word embeddings, exhibiting comparable or even superior performance than those obtained with aggregated sentence representations. Relying instead on a variable selection approach applied to our set of probing tasks, we showed the existence of a relationship between the implicit linguistic knowledge encoded by BERT and the number of individual units involved in the encoding of this knowledge. Specifically, according to the strategy for obtaining sentence-level representations, the amount of hidden units devised to encode linguistic properties varies differently across BERT layers: while the number of non-zero units used in the *Mean-pooling* strategy remains more or less constant across layers, the *[CLS]* representations show a continuous increase in the number of used coefficients. Moreover, we noticed that this behaviour is particularly significant for linguistic properties related to the whole structure of the syntactic tree, while features belonging to POS and dependency tags tend to acquire less non-zero units across layers.

Another issue in the investigation of the linguistic competence implicitly learned by NLMs is how this knowledge is modified after a fine-tuning process and how it affects the decisions they make when solving specific downstream tasks. The results obtained in our study suggested that BERT tends to lose its precision in encoding the set of probing features after a fine-tuning process (i.e. Native Language Identification), probably because it is storing more task–related information for solving NLI. Nevertheless, we found that the linguistic knowledge implicitly encoded by the model positively affects its ability to solve the tested downstream tasks: the more it stores readable linguistic information of a sentence, the higher will be its capacity of predicting the expected label assigned to that sentence.

With findings reported in Sec. 8.5 we tried to achieve a deeper understanding of the linguistic competence learned by Transformer models in a language other than English. Specifically, presenting an in-depth comparative investigation of the linguistic knowledge encoded in by 7 Italian transformers relying on two different probing architectures, we first showed experimentally how non-linear architectures such as the multi-layer perceptron (MLP) capture a broader range of information encoded in learned representations with respect to their linear counterparts, and as such they can be considered more suitable for studying highly nonlinear models such as NLM. In this sense, our results support the information-theoretic operationalization of probing proposed by [Pimentel et al., 2020a]. However, the rankings of this and of the LinearSVR model in terms of their probing ability are quite similar. Namely, both are particularly able to probe raw text properties, as well as the distribution of Parts-Of-Speech and dependency relations; while they obtained lower scores for features referring to the order of subject and object with respect their verbal head and to the verbal predicate structure. The following comparison of the linguistic generalization abilities of the 7 models showed that if we analyse the results considering all the probing features as a whole few differences can be observed. More interesting outcomes result when we focus on the embedded knowledge of each group of linguistic characteristics. We noticed for example that global and local tree structure properties are predicted more accurately by RoBERTa-style models, i.e. by GilBERTo and UmBERTo-Commoncrawl, than by models based on BERT or GPT-2. We obtained additional information when we narrowed our analysis on how NLMs prediction abilities evolve across models' layers, showing for example that the highest competence about the tree structure is achieved by a BERT-style model, i.e. BERT-base-italian, in the -4 layer. A more in-depth comparison with respect to the ranking of each individual feature by R^2 scores also revealed that, even if the 7 Transfomers are quite similar, a main exception is represented by the AlBERTo model. In particular, it showed to have reduced generalization abilities concerning the use of punctuation. Our intuition is that it is possibly related to the typology of texts the AlBERTo model was trained on, i.e. Twitter, where punctuation marks are rarely used. Finally, we showed that the level of NLMs linguistic competence changes according to the diverse linguistic varieties of IUDT. All Transformers resulted to be less robust in the prediction of the linguistic properties characterising sentences representative of social media language and of sentences in the interrogative form. This is possible due to the fact that the two types of sentences are characterised by non-canonical distribution of morpho-syntactic and syntactic phenomena, possibly rarely or never seen during the training phase. Surprisingly, also the AlBERTo model, even if it was trained on Twitter data, achieved very low performances, while on the contrary, BERT-base italian and GePpeTto are the two models slightly less affected by the non-standard linguistic varieties. Despite both social media and questions seem representing two quite challenging testbeds, our in-depth investigation of how each probing feature is ranked by the NLMs allowed highlighting noteworthy differences. We observed that the most diverse rankings concern the test on the sentences in the interrogative form, which result to be characterised by distributions of structural aspects of sentence very different from other IUDT sections.

Finally, in Sec. 8.6 we described a methodology to test the effectiveness of a suite of probing tasks for evaluating the linguistic competence encoded by NLMs. To this aim, we analysed the performances of a probing model trained with Italian BERT representations to predict the authentic and automatically altered values of a set of linguistic features derived from IUDT. We observed general higher performances in the prediction of authentic values, thus suggesting that the probing model relies on linguistic competences to predict linguistic properties. However, when we constrained automatically altered values with respect to sentence length, the model tends to learn surface patterns in the data.

CHAPTER 9

Assessing NLMs Linguistic Abilities

In this chapter we discuss the experiments that focused on the relationship between perplexity scores and grammatical generalization abilities and on the performance of NLMs on diagnostic tests built to probe their sensitivity to specific language phenomena.

9.1 Introduction

In Chapter 4 we showed that alternative methodologies to the probing tasks approach have been proposed to analyze the linguistic competence of NLMs. They range from the analysis of the relationship between perplexity scores and grammatical generalization abilities [Hu et al., 2020] to the investigation of NLMs performance of diagnostic datasets built to probe their ability on targeted syntactic phenomena [Warstadt et al., 2020]. As regards the first approach, in this section we present two complementary studies. In the first one (Sec. 9.2) we study how the linguistic structure of a sentence affects the perplexity of a NLM and whether it is possible to predict NLMs perplexity scores using our set of linguistic features already expolited during the probing tasks experiments. In the second study (Sec. 9.3), we first investigate the relationship between NLM perplexity scores and the readability scores assigned to the same sentences by a supervised readability assessment tool and then we verify whether the two metrics are equally affected by the same set of linguistic phenomena. Finally, in Sec. 9.4 we introduce new evaluation resource for the Italian language in order to test the understanding of textual connectives in real-usage sentences by the most recent NLMs.

9.2 What Makes my Model Perplexed?

In this study, we present an investigation aimed at studying how the linguistic structure of a sentence affects the perplexity of a NLM. Rather than studying the relation between the NLM's perplexity and its linguistic competences assessed on sentences undergoing controlled syntactic modifications, we focus on sentences representative of real usage. Our purpose indeed is to understand which linguistic phenomena of the input sentence may make perplexed a NLM and whether they can effectively predict the assigned perplexity score. To have a in-depth understanding of the relation between linguistic structure and perplexity, we relied on the set of linguistic features we have previously exploited to probe the competence of these models trough the probing tasks approach. As we also intend to evaluate the possible influence of the NLM architecture on this relation, in all our experiments we consider two of the most popular NLMs, a traditional unidirectional one, i.e. GPT-2, and a bidirectional model such as BERT.

The contributions of this study are as follows:

- we showed that a sentence-level likelihood computed by masking each word sequentially for the BERT model has a robust correlation with GPT-2's perplexity scores;
- we verified whether it is possible to predict NLMs' perplexities using a wide set of linguistic features extracted by a sentence;
- we identified the linguistic properties of a sentence that mostly cause perplexity, reporting differences and similarities between the two models.

9.2.1 Approach

We defined two sets of experiments. The first consists in investigating the relationship between BERT and GPT-2 sentence-level perplexity (*PPL*) scores. To do so, we first computed BERT and GPT-2 *PPL* scores for sentences contained in the English Universal Dependencies (UD) treebank [Nivre et al., 2016] and we assessed their correlation. In the second set of experiments, we studied whether a simple regression model that takes as input our set of linguistic features automatically extracted from each UD sentence is able to predict the two NLMs sentence-level perplexities.

To understand which linguistic phenomena contribute to the prediction of BERT and GPT-2 PPLs, and how these features differ between them, we performed an in-depth investigation training the regression model with one feature at a time.

Models and Data For our experiments, we rely on the pre-trained version of the two NLMs previously defined. We first computed GPT-2's sentence-level perplexities by dividing the sum of all sub-word conditional log-probabilities by the total number of words for each sentence in the UD dataset. On the other hand, since BERT masked language modeling task does not allow to compute well-formed probability distributions

Lengths	ρ score	# samples
All	0.63	22,505
n=10	0.66	847
n=15	0.60	793
n=20	0.64	643
n=25	0.53	422
n=30	0.54	277

Table 9.1: *Spearman correlations between BERT and GPT-2 perplexities computed for all UD sentences (All) and sentences with fixed-length n.*

over sentences, we measure BERT sentence-level likelihood by masking each word sequentially and computing the probability as follows:

$$p(S) \approx \prod_{i=1}^{k} p(w_i | w_1, ..., w_{i-1}, w_{i+1}, ..., w_k) \tag{9.1}$$

The perplexity is then computed as follows:

$$PPL_S = e^{(\frac{p(S)}{N})} \tag{9.2}$$

where *N* correspond to the length of sentence *S*. In order to uniform the terminology, in what follows we will refer to the BERT sentence-level likelihood as perplexity.

In order to evaluate our approach on gold annotated sentences, we relied on the three English Universal Dependencies (UD) treebanks (ParTUT, GUM corpus and EWT), for a total of 22,505 sentences, and we extracted a subset 78 linguistic features from those defined in [Brunato et al., 2020] for each sentence in the dataset.

9.2.2 A Linguistic Investigation on Perplexity

As a first step, we assessed whether there is a relationship between the perplexity of a traditional NLM and of a masked NLM. We thus calculated BERT and GPT-2 perplexity scores for each UD sentence and measured the correlation between them. Since *PPL* scores are highly affected by the length of the input sequence, we computed ρ correlation coefficients also considering groups of sentences with fixed length. Specifically, we relied on Spearman correlation because we were interested in measuring how the variations in perplexity scores relate each other, rather than focusing on the actual *PPL* values. Results are reported in Table 9.1. As we can notice, even considering samples with fixed length, the two NLMs' perplexities exhibit moderate to substantial correlation (with $p < 0.001$), thus showing that BERT an GPT-2 do not diverge excessively in their ability of predicting the likelihood of the input sentences. Moreover, this allows us to confirm that, although the deep bidirectional structure of BERT does not permit to compute a well-formed probability distribution over a sentence, this metric could be considered as a valid approximation of the perplexity computed with a unidirectional NLM.

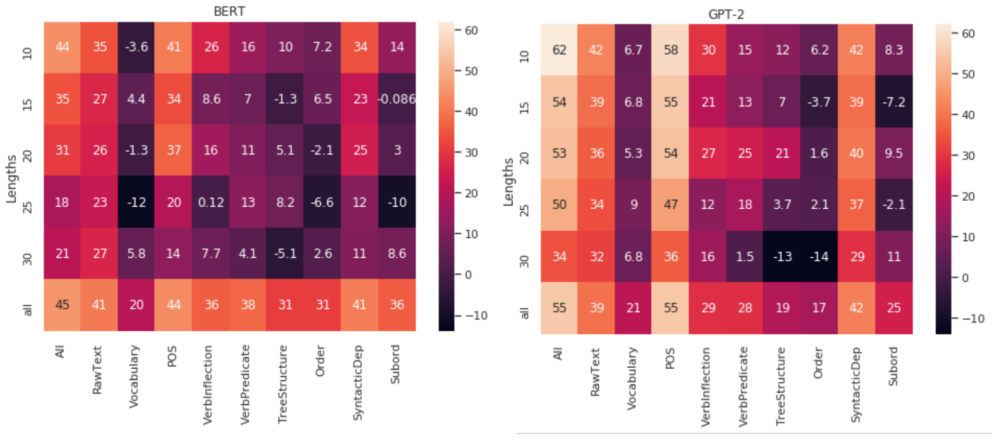


Figure 9.1: *BERT and GPT-2 ρ scores (multiplied by 100) obtained with the LinearSVR model using linguistic features, for the whole UD dataset and groups of sentences with fixed length.*

Once established the correlation between the perplexities of the two NLMs, we performed a second experiment to investigate (i) if the considered set of linguistic features plays a role in predicting their perplexity and (ii) which are the features that contribute more to the prediction task. To do so, we trained a LinearSVR model that predicts perplexity's scores using our set of linguistic properties as input features. Since most of them refer to syntactic properties of sentence that are strongly correlated with its length, we considered as a baseline a SVR model that takes sentence length as input and outputs BERT/GPT-2 sentence's perplexity. Regression results deriving by considering both the whole set (*All*) and each of the 9 groups of linguistic features separately are reported in Figure 9.1. As a general remark, for the whole UD dataset, we can observe that the results considering both all and the 9 groups of linguistic features outperform the results obtained by the baseline, i.e. ρ=0.38 for BERT and 0.22 for GPT-2 respectively. This demonstrates that the considered features are able to model aspects involved in NLM's perplexity that go beyond the simple length of sentence. This is particularly the case of GPT-2, suggesting that the probability assigned to a sentence by a traditional NLM is more explainable in terms of linguistic phenomena mainly affecting morpho-syntactic and syntactic structure. Consequently, the baseline score is higher for BERT. If we consider the scores obtained for each group of sentences with fixed length, we can see that higher scores are obtained for groups containing shorter sentences, for both NLMs. This is quite expected since in these sentences the possible output space is smaller for almost all features, thus making them more predictive. Also in this case, the impact of the linguistic features is always higher for the prediction of GPT-2's perplexity.

A more in-depth analysis of these results shows that the distribution of the morpho-syntactic characteristics of a sentence (*POS*) and of the syntactic dependency relations (*SyntacticDep*) are the two most predictive sources of linguistic information. As Figure 9.1 reports, this holds for the two NLM models and it remains constant throughout all the groups of sentences with fixed lengths. Interestingly, if we consider the whole set of

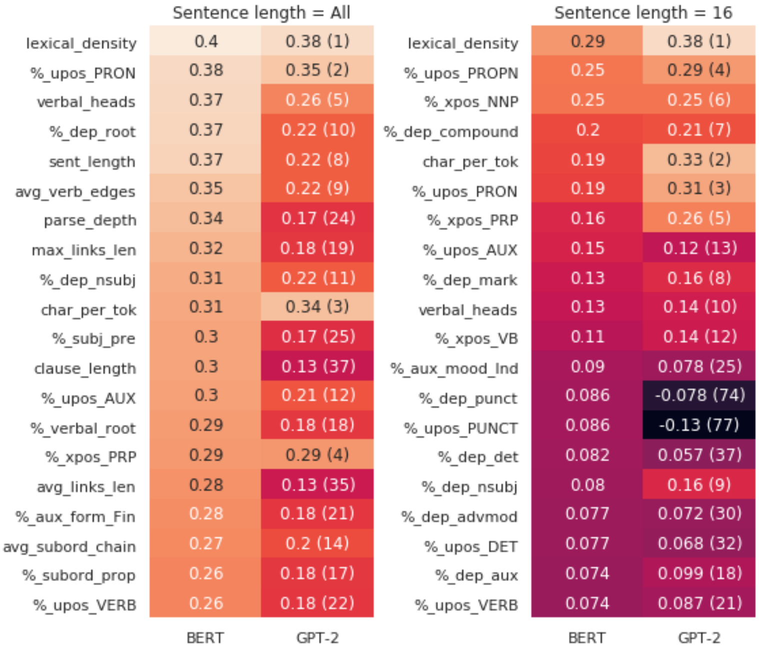


Figure 9.2: *BERT and GPT-2 ρ scores obtained with the LinearSVR model, for the whole UD dataset and 16 token-long sentences. Scores are reported for the 20 top-ranked features for BERT. Numbers in brackets correspond to the relative in the GPT-2 ranking.*

sentences, the effect of the morpho-syntactic information on the prediction of GPT-2's perplexity is exactly the same of that of the whole set of linguistic features. For some sentence lengths (15, 20, 30) the scores obtained using only this type of information outperform even those obtained considering the whole set of features. Note that this last remark is true also in the prediction of BERT's perplexity. As expected the other most predictive group is the one (*RawText*) that includes the length of sentence.

Focus on the contribution of individual features To investigate more in depth which linguistic phenomena are more involved in the perplexity of the two models, we trained the LinearSVR model using each individual feature at a time. This was done for both the whole dataset and the subset of sentences (i.e. 758 sentences) having a length of 16 tokens, which corresponds to the mean sentence length of the UD dataset. A subset of results is reported in Figure 9.2. As we can see in the left-side of the heatmap, the two models share many features in the first ten positions, thus showing that the two NLM architectures are made perplexed by similar linguistic characteristics of a sentence. In particular, for both of them, the two most predictive features correspond to the lexical density and the presence of pronouns confirming the highly predictive power of morpho-syntactic information. They are followed by features related to the presence of verbs and to their internal structure (i.e. *verbal_heads* and *avg_verb_edges*), and, as it was expected, by the length of the sentence. Despite these similarities, we can see that the scores obtained by the regression model to predict BERT's perplexity are on average higher than GPT-2's scores. Considering that we obtained higher scores using all (or groups of) features in the prediction of GPT-2' perplexity (see Figure 9.1), this latter result may suggest that the interaction among features is less relevant in the prediction of BERT's perplexity. Differences among the two models concern features that are highly sensitive to sentence length, which result to be more predictive of BERT's perplexity. This is the case of syntactic features capturing global and local aspects of sentence structure, i.e. the depth of the whole syntactic tree (*parse_depth*), the maximum length of dependency links (*max_links_len*) and the length of verbal clauses (*clause_length*). Also, the canonical order of nuclear sentence elements such as pre-verbal subjects contribute more to predict BERT's than GPT-2's perplexity. Instead, the distribution of proper nouns (*%_upos_PROPN*), in particular in their singular form (*%_xpos_NNP*), the length of token (*char_per_tok*) and vocabulary richness are more predictive of GPT-2's perplexity. Although we cannot say from ranking results whether features highly ranked are positively or negatively correlated with perplexity, we can hypothesize that knowing the distribution of tokens belonging to open lexical categories (e.g. proper nouns vs determiners) make the perplexity easier to identify.

The right-side heatmap shows the top-ranked features used to predict the two models perplexity for sentences 16-token long. As expected, when sentence length is controlled, the role of other features less related to length becomes predominant. In particular, morpho-syntactic information is still highly predictive for the two models, with lexical parts-of-speech showing to be relevant not only for GPT-2's but also of BERT's perplexity.

9.3 Is Neural Language Model Perplexity Related to Readability?

Once investigated the relationship between NLMs perplexity scores and linguistic generalization abilities, we decided to focus on a less investigated perspective, addressing the connection between perplexity and readability. Since by definition perplexity gives a good approximation of how well a model recognises an unseen piece of text as a plausible one, our intuition is that lower model perplexity should be assigned to easy-to-read sentences, while difficult-to-read ones should obtain higher perplexity. On the other hand, as we already seen in previous works and in our experiments, state-of-the-art NLMs trained on huge data have shown to implicitly learn a sophisticated knowledge of language phenomena, also with respect to complex syntactic properties of sentences [Tenney et al., 2019a, Jawahar et al., 2019, Miaschi et al., 2020a]. This could suggest that variations in terms of linguistic complexity, especially when related to subtle morpho–syntactic and syntactic features of sentence rather than lexical ones, could not impact on model perplexity to a great extent. This assumption seems to be confirmed by the results by [Martinc et al., 2021] which, to our knowledge, is the only one explicitly leveraging unsupervised neural language model predictions in the context of readability assessment. According to this study, a NLM is even less perplexed by articles addressed at adults than by documents conceived for a younger readership. From a relatively different perspective focused on the ability of automatic comprehension systems to solve cloze tests, [Benzahra and Yvon, 2019] showed that NLMs performance is not affected by the level of text complexity.

In order to test the validity of all these hypotheses, we rely on the perplexity score given by a state-of-the-art NLM for the Italian language to several datasets representative of different textual genres containing both easy– and complex–to–read sentences: ideally, such datasets should emphasise the correlation between perplexity and readability (if present) since the corpora are explicitly designed to contain both simple and difficult examples.

The contributions are as follows:

- we investigated if the perplexity of a NLM and the readability score of a set of sentences show a significant correlation;
- we studied whether the two metrics are equally affected by the same set of linguistic phenomena that occur in the sentence.

9.3.1 Approach

According to our research questions, we devised a set of experiments to study whether NLMs perplexity reflects the level of readability of a sentence and which are the linguistic phenomena mostly involved in each metric. For this purpose, we firstly investigated whether sentence-level perplexity scores computed with one of the most prominent NLM model correlate with the scores assigned to the same sentences by a supervised readability assessment tool. Secondly, we investigated which are the linguistic features of the considered sentences that correlate in a statistically significant way with the perplexity and readability score respectively. In order to verify whether correlations hold across different typology of texts, we tested our approach on five Italian datasets.

Models

Automatic readability (*ARA*) was assessed using READ-IT [Dell'Orletta et al., 2011a] the first readability assessment tool for Italian which combines traditional raw text features with lexical, morpho-syntactic and syntactic information extracted from automatically parsed documents. In READ-IT, analysis of readability is modelled as a binary classification task, based on Support Vector Machines using LIBSVM [Chang and Lin, 2001]. Training corpora are representative of two classes of texts, i.e. difficult– vs. easy–to-read ones, both containing newspaper articles. The set of features exploited for predicting readability has been proved to capture different aspects of sentence complexity. Thus, the assigned readability score ranges between 0 (easy-to-read) and 1 (difficult-to-read) referring to the percentage probability for unseen documents or sentences to belong to the class of difficult-to-read documents. For the purposes of our work, we carried out readability assessment at sentence level, making the analysis reliable for the comparison with sentence-based perplexity of a NLM.

Sentence-level perplexity scores were computed relying on the GePpeTto model [De Mattei et al., 2020]. The perplexity (PPL) of the model was computed as in the previous study (see Sec. 9.2).

Corpora

In order to test the reliability of our initial hypothesis, we chose four corpora containing different typologies of texts, i.e. web pages, educational materials, narrative texts, newspaper and scientific articles. Each corpus includes a balanced amount of difficult- and easy-to-read sentence. In addition, we also considered in the analysis the Italian Universal Dependency treebank. This is meant to verify whether the connection between sentence-level readability and perplexity also holds in a well-acknowledged benchmark corpus. For each of them, we excluded from our analysis short sentences, i.e. having less than 5 tokens.

PACCSS-IT We took into account 125,977 sentences belonging to PACCSS-IT[1] [Brunato et al., 2016], a corpus of complex-simple aligned sentences extracted from the ItWaC corpus. The resource was build using an automatic approach for acquiring large corpora of paired sentences able to intercept structural transformations (such as deletion, reordering, etc.). For example, the two following sentences represent a pair in the corpus, where a reordering operation occurs at phrase level (i.e. the subordinate clause proceeds vs. follows the main clause):

- Complex: *Ringraziandola per la sua cortese attenzione, resto in attesa di risposta.* [Lit: Thanking you for your kind attention, I look forward to your answer.]
- Simple: *Resto in attesa di una risposta e ringrazio vivamente per l'attenzione.* [Lit: I look forward to your answer and I thank you greatly for your attention.]

Terence and Teacher Two corpora of original and manually simplified texts aligned at sentence level[2] [Brunato et al., 2015]. *Terence* contains short Italian novels for children and their manually simplified version carried out by linguists and psycholinguists targeting children with text comprehension difficulties. *Teacher* is a corpus of pairs of documents belonging to different genres (e.g. literature, handbooks) used in educational settings manually simplified by teachers. We exploited 1,644 sentences belonging to these corpora.

Multi–Genre Multi–Type Italian corpus A collection of Italian texts representative of three traditional textual genres: Journalism, Scientific prose and Narrative. Each genre has been internally subdivided into two sub-corpora representative of an easy- vs difficult-to-read variety, which was defined according to the intended target audience for a given genre. The journalistic prose corpus includes articles automatically downloaded from the online versions of two general-purpose newspapers[3], while the "easy" sub-corpus contains articles from two easy-to-read newspapers[4] addressed to adults with low

[1] `http://www.italianlp.it/resources/paccss-it-parallel-corpus-of-complex-simple-sentences-for-italian/`

[2] `http://www.italianlp.it/resources/terence-and-teacher/`

[3] `www.repubblica.it` and `http://www.ilgiornale.it/`

[4] `www.dueparole.it` and `http://www.informazionefacile.it/`

Dataset	PPL	ARA
PACCSS-IT	3,905.83 (± 21,306.07)	0.55 (± 0.24)
Terence-Teacher	790.85 (± 5,002.62)	0.46 (± 0.27)
Multi-Genre Multi-Type	570.85 (± 4,820.12)	0.58 (± 0.31)
Italian-UD	436.75 (± 3,633.64)	0.61 (± 0.30)
Twitter-UD	986.28 (± 2,479.64)	0.59 (± 0.30)

Table 9.2: *Perplexity (PPL) and Readability (ARA) mean and standard deviation values for the 5 datasets.*

Dataset	PPL-ARA	Feats
PACCSS-IT	-0.031*	0.169*
Terence-Teacher	0.014	0.149
Multi-Genre Multi-Type	0.026*	0.184*
Italian-UD	-0.054*	0.332*
Twitter-UD	-0.038*	-0.037

Table 9.3: *Spearman's correlation coefficients between sentence-level perplexity and readability scores (PPL-ARA) and between rankings of linguistic features (Feats). Statistically significant correlations ($p < 0.05$) are marked with *.*

literacy skills or mild intellectual disabilities. The scientific prose collection consists of scholarly publications on linguistics and computational linguistics and Wikipedia pages downloaded from the portal "Linguistics", representative of the complex and easy variety respectively. For the narrative genre, we included long novels written by novelists of the last century and contemporary writers in the corpora of complex variety, while for the easy variety we collected short novels for children. The complete corpus contains 56,685 sentences.

Italian Universal Dependency Treebank The IUDT dataset already used in the probing experiments with the Italian Transformer models.

9.3.2 Sentence Perplexity and Readability

Our analysis starts from a comparison between the average perplexity and readability scores obtained for each sentence of the five considered datasets. As shown in Table 9.2, readability values (column *ARA*) are quite homogeneous across the datasets, with low standard deviation values. On the contrary, the range of perplexity scores is wider (column *PPL*), going from an average score of 3,905.83 of PACCSS-IT to 436.75 of the IUDT miscellaneous portion (Italian UD). These differences seem to provide a first evidence that perplexity and readability are not correlate to each other.

This intuition has been proved computing the Spearman's rank correlation coefficient between the perplexity and readability scores for each dataset. Results are reported in Table 9.3, column *PPL-ARA*. As it can be seen, all correlation rates are significant, except for the result obtained on the Terence and Teacher corpus, possibly due to the fact that the size of the corpus is too small to allow a significant comparison. Contrary to our expectations, no correlation was detected between the two metrics for all corpora, suggesting that perplexity and and readability are independent from each other.

To further investigate the reasons behind these scores and to deepen the analysis about the relationship between the two metrics, we investigated whether they capture the same (or similar) linguistic properties of the sentences. To this aim, we tested the presence and strength of the correlation between each of the two metrics and the whole set of linguistic features devised in [Brunato et al., 2020]. Column *Feats* of Table 9.3 illustrates the results of this analysis: we report the Spearman's correlation coefficients between the two rankings of linguistic features, each ordered by strength of correlation between feature value and perplexity score and readability score respectively. Once again we observe rather weak correlation values, with the only exception of Italian-UD which is the only one reporting a medium correlation (.332). Overall, these results corroborate our previous findings that the two metrics are not particularly related with each other, and they further suggest that the linguistic phenomena affecting the perplexity of NLM and the readability level of a sentence are very different. Consider for example the two following sentences:

> *Il furto è avvenuto giovedì notte.* [en. The theft has taken place Thursday night.]
>
> *Il comitato di bioetica: no all'eutanasia.* [en. The bioethics committee: no to euthanasia.]

While (1) is very easy-to-read, with a readability score of 0.25, but it has a quite high perplexity score, i.e. 40,737.81, (2) is quite difficult-to-read (ARA=1) but is has a very low perplexity score (PPL=11.24).

9.3.3 In-Depth Linguistic Investigation

To better explore the motivation behind these results, we performed an in-depth investigation aimed at understanding the relationship between our set of linguistic features and the two metrics taken into consideration. Since we noticed that for all datasets a higher number of features correlates with ARA than with PPL, we selected those that are significantly correlated with the two metrics. The number of shared features varies for each dataset, depending on their size. For example, for the two smallest ones, i.e. Terence and Teacher and the UD Twitter Treebank, we could only consider 34.65% (61) and 44.88% (79) of the whole set of features respectively, while for the larger corpora the sub-set is wider: 81.81% (144) in PACCSS-IT, 78.97% (139) for Multi-Genre Multi-Type and 84.65% (149) for the IUD Treebank.

Table 9.4 shows the top ten features for each dataset, i.e. those that obtained the strongest correlation with both PPL and ARA. As expected, correlations are generally stronger between linguistic features and readability scores, although they are lower than expected. This could be due to the fact that, even if the READ–IT classifier is trained with a similar set of features, the non-linear feature space makes it difficult to identify clear correlations with individual features. Similarly, our set of features seem to play only a marginal role on perplexity. However, this is not the case of the PACCSS-IT

PACCSS-IT			
PPL		ARA	
Feats	*Corr*	*Feats*	*Corr*
aux_num_pers_dist_Sing+3	0,53	xpos_dist_FF	0,34
dep_dist_cop	0,51	dep_dist_punct	0,32
avg_max_depth	0,50	upos_dist_PUNCT	0,32
upos_dist_ADP	0,50	ttr_form	0,29
xpos_dist_E	0,50	aux_mood_dist_Cnd	0,25
dep_dist_case	0,49	upos_dist_DET	0,25
n_tokens	0,48	dep_dist_det	0,25
dep_dist_root	0,48	ttr_lemma	0,22
xpos_dist_FS	0,48	upos_dist_NOUN	0,21
Terence and Teacher			
PPL		**ARA**	
Feats	*Corr*	*Feats*	*Corr*
xpos_dist_B	0,25	dep_dist_det	-0,39
verbs_num_pers_dist_Sing+3	0,23	upos_dist_DET	-0,38
lexical_density	0,22	upos_dist_NOUN	-0,37
dep_dist_advmod	0,21	xpos_dist_S	-0,37
upos_dist_ADV	0,21	xpos_dist_RD	-0,29
verbs_num_pers_dist_Plur+3	-0,16	upos_dist_ADV	0,27
xpos_dist_V	0,16	dep_dist_advmod	0,25
avg_token_per_clause	-0,16	xpos_dist_FF	0,25
upos_dist_VERB	0,14	avg_sub_chain_len	0,24
Multi-Genre Multi-Type			
PPL		**ARA**	
Feats	*Corr*	*Feats*	*Corr*
n_tokens	-0,19	principal_prop_dist	-0,42
dep_dist_root	0,19	ttr_form	-0,34
dep_dist_advmod	0,19	xpos_dist_FF	0,34
upos_dist_ADV	0,18	dep_dist_det	-0,33
n_prepositional_chains	-0,18	upos_dist_DET	-0,33
xpos_dist_B	0,18	upos_dist_PUNCT	0,33
upos_dist_ADP	-0,17	dep_dist_punct	0,33
xpos_dist_E	-0,17	xpos_dist_FB	0,31
ttr_lemma	0,16	sub_prop_dist	0,27
Italian UD Treebank			
PPL		**ARA**	
Feats	*Corr*	*Feats*	*Corr*
n_tokens	-0,27	principal_prop_dist	-0,53
dep_dist_root	0,27	sub_proposition_dist	0,40
n_prepositional_chains	-0,26	n_tokens	0,39
avg_max_depth	-0,24	dep_dist_root	-0,39
upos_dist_ADP	-0,24	ttr_form	-0,37
ttr_lemma	0,23	avg_max_depth	0,36
max_links_len	-0,23	avg_links_len	0,35
avg_max_links_len	-0,23	max_links_len	0,34
xpos_dist_E	-0,22	avg_max_links_len	0,34
Italian UD Twitter Treebank			
PPL		**ARA**	
Feats	*Corr*	*Feats*	*Corr*
upos_dist_SYM	0,38	upos_dist_PUNCT	0,30
avg_max_depth	-0,28	dep_dist_punct	0,30
xpos_dist_SYM	0,28	dep_dist_det	-0,29
in_dict	-0,24	upos_dist_DET	-0,29
dep_dist_vocative:mention	0,23	verbal_root_perc	-0,27
in_dict_types	-0,22	xpos_dist_RD	-0,27
ttr_lemma	0,21	avg_token_per_clause	-0,27
in_FO	-0,21	subj_pre	-0,27
verbal_head_per_sent	-0,19	obj_post	-0,24

Table 9.4: *Top 10 features along with their correlation scores between perplexity and readability.*

corpus, for which the set of considered linguistic features have an higher correlation with PPL. This can be possibly related to the partial overlap between the GePpeTto training data and the PACCSS-IT sentences, since the latter is drawn from the ItWac corpus which is included in the GePpeTto's training.

Inspecting these results, we can also observe that correlations between features and PPL seem to be more affected by genre–specific characteristics. This is particularly clear if we consider the Italian UD Twitter treebank, for which among the top ten most correlated features we find some of them characterising social media language, e.g. symbols (*upos-xpos_dist_SYM*) or the vocative relation, which marks a dialogue participant addressed in a text along with the specification, specifically used for Twitter @-mentions (*dep_dist_vocative:mention*).

9.4 On the role of Textual Connectives in NLMs Sentence Comprehension

To pursue the investigation of the inner competence learned by state-of-the art NLMs, it has become of pivotal importance the availability of challenging test sets, built to probe the sensitivity of a model to specific language phenomena. So far, most of the efforts have been focused on assessing the syntactic abilities encoded by NLMs by exploiting human curated benchmarks, which are usually proposed in the form of minimal sentence pairs, i.e. minimally different sentences exemplifying a wide array of linguistic contrasts (e.g. [Warstadt et al., 2020]). Differently from syntactic well-formedness, less explored is the sensitivity of these models to deeper linguistic dimensions involving semantics and discourse, such as textual cohesion, which are critical to language understanding. With this respect, one of the explicit devices that natural languages use to convey textual cohesion is represented by function words. As observed by [Kim et al., 2019], although these words plays a key role in compositional meaning as they introduce discourse referents or make explicit relations between them, they are still under-investigated in the literature on representation learning. To this end, the authors released a suite of nine challenge tasks for English aimed to test the NLMs' understanding of specific types of function word, e.g. coordinating conjunctions, quantifiers, definite articles.

Taking inspiration from this work, in this study we focus the attention on the role of textual connectives in the comprehension of a sentence and we introduce a new evaluation resource for Italian which, to our knowledge, is the first one for this language. The resource is articulated into two sections, each one corresponding to a distinct task aimed at probing, in a different format, to what extent current NLMs are able to properly encode the role of connectives in a sentence. A peculiarity of the dataset is that it contains sentences that were extracted and minimally modified from existing corpora so as to test the comprehension of connectives in the real use of language.

9.4.1 Corpus collection

Selecting connectives and extracting sentences As a first step, we defined the linguistic criteria for the selection of connectives to include in the corpus. By *connective* we mean specific words that have the function of drawing a relation between two or more clauses [Sanders and Noordman, 2000, Graesser and McNamara, 2011]. To this end, two resources were employed: the INVALSI reading comprehension and language reflection tests designed by the National Institute for the Evaluation of the Education System and the *Nuovo Vocabolario di Base* of Italian [De Mauro and Chiari, 2016]. Starting from the collection of the INVALSI tests proposed in the last six years for different grades, we extracted all words which were expressly called 'connective' in the tests or were involved in defining a logical relationship between two sentences. We thus obtained a first list of 46 elements, belonging to diverse morpho-syntactic categories (i.e. prepositions, conjunctions, adverbs), which was then integrated with other 19 connectives extracted from the *NVdB*. We then checked the distribution of the selected items in existing corpora and extracted the sentences in which these words were unambiguously used as sentence connectives. Three different sections of the Italian Universal Dependency Treebank (IUDT) were used: i.e. ISDT, PoSTWITA and TWITTIRò[5], the first one representative of standard language and the latter collecting Italian tweets. We employed PML TreeQuery[6] to query the treebanks and filter the sentences containing the connectives we were interested in. In particular, to exclude occurrences which do not have the role of phrasal connectives (e.g. the conjunction *e* joining two nouns), only sentences in which the connective was headed by a verb or a copula were taken into account. Given the overlapping of the frequency data in the three corpora and the potential non-standard use of connectives in treebanks representative of social media texts, also due to genre-specific features (e.g. hashtag, emoticons etc.), we decided to consider only the first 21 most frequent connectives occurring in ISDT. Further considerations on their distributions led us to the deletion of *per*, *così*, *ancora*, because of their ambiguous behavior as textual connectives (e.g. we noticed that the majority of the occurrences of *per* involves the presence of an infinite verb, a distribution which is far from the other connectives). The following 18 connectives were finally considered: *e*, *se*, *quando*, *come*, *ma*, *dove*, *o*, *anche*, *perché*, *poi*, *mentre*, *infatti*, *prima*, *però*, *invece*, *inoltre*, *tuttavia*, *quindi*.

Once established the final list, all sentences containing the selected connectives were extracted from ISDT and eventually modified following some patterns, to guarantee sentence comprehension. For example, in some cases two sentences occurring in the treebank in a subsequent order, but that were clearly extracted from the same text, were joined together to form a unique sentence, through the insertion of the appropriate punctuation. This happened e.g. when the connective appeared at the beginning of the second sentence joining this to the first one, which serves as the antecedent to comprehend the logical relationship. For the collection of the final dataset, we also tried to include sentences with different degrees of syntactic and lexical complexity, considering the number of subordinate clauses and the variety of the lexicon as related proxies.

The collected sentences were grouped in two sections aimed at testing the correct comprehension of connectives in a different format, i.e. through an acceptability assess-

[5] https://universaldependencies.org/treebanks/it-comparison.html

[6] https://ufal.mff.cuni.cz/pmltq

Section	Id	Sentence
Acceptability	e_11A	L'arte e la scienza sono libere e libero ne è l'insegnamento.
	e_11NA	L'arte e la scienza sono libere tuttavia libero ne è l'insegnamento.
	ma_64A	Paolo si muove con difficoltà, ma è sempre allegro e di buon umore.
	ma_64NA	Paolo si muove con difficoltà, perché è sempre allegro e di buon umore
Cloze test	se_23cl	Che cosa possiamo fare in estate ... vogliamo partire per le vacanze e abbiamo un cane o un gatto? [**se** *quando* perché dove come]
	mentre_162cl	Nelle botteghe artigianali della produzione di piastrelle la smaltatura è ancora tradizionale, ... i forni, come è naturale, oggi funzionano a gas. [**mentre** *invece* come dove perché]

Table 9.5: *Examples from the dataset. Sentences are indicated with the last part of id, which gives information about the target connective, the position of the sentence in the section and the label in each section (A='acceptable', NA='not acceptable'; cl='cloze test'). For the cloze task, the target connective is marked in bold and the plausible alternative in italics.*

ment task and a cloze test task. Table 9.5 provides an example of sentences/sentences pairs for each task.

Acceptability assessment section To design the acceptability assessment task, we selected 15 sentences per connective from the whole dataset. For each sentence, an unacceptable counterpart was created by replacing the original connective with another of the list. The replacement strategy was meant to obtain unacceptable sentences with contradictory or nonsensical meaning but preserving their grammaticality. Indeed those sentences should be the most challenging one for NLMs, which have been shown to be capable of detecting sentence grammaticality [Jawahar et al., 2019], but still struggle to track down unacceptable meanings and contradictions. Nevertheless, we were not always able to guarantee this constraint as for some specific contexts none of the available connective could be substituted without affecting the resulting grammaticality. This happened in 98 cases, which we decided to keep in the dataset but we signaled with the label 'no' in the field 'grammaticality'. A few sentences were also deleted due to ambiguity. The final section contains 518 sentence pairs, i.e. 259 acceptable and 259 unacceptable ones.

Cloze test section The second section was designed as a cloze test task and contains 270 sentences, 15 for connective. For every sentence the original connective was replaced by a blank space and 5 alternatives were proposed for completion: the target, a plausible alternative and three implausible options. For 'plausible alternative' we mean another connective of the list that could occupy the same linguistic contest of the target, yielding to an identical meaning or to a different, yet totally plausible, reading. As for the acceptability task, it turns out that for some connectives (e.g. *prima*) it was very challenging, if not impossible, to propose such a plausible connective. In those cases, that in truth are only a minority, it has been proposed an alternative that at least should guarantee the grammaticality.

Acceptability label	AvgIntScore	(StDev)
Acceptable	4.286	0.519
Unacceptable	1.822	0.451
Unacceptable (AG)	1.616	0.350

Table 9.6: *Average scores assigned by humans (with standard deviation) to the acceptable, unacceptable and unacceptable+ungrammatical sentences.*

9.4.2 Corpus Annotation

The two sections of the dataset were splitted into 9 surveys (5 for the acceptability assessment task and 4 for the cloze task) and submitted to human evaluation by recruiting Italian native speakers of different ages through the Prolific platform

In the **acceptability assessment task**, participants were asked to judge the acceptability of each sentence on a 5-grade Likert scale (from 1='totally unacceptable' to 5='totally acceptable'). Although this makes the dataset more challenging, we assume that acceptability is a gradual rather than binary notion as it is affected by many factors [Sorace and Keller, 2005, Sprouse, 2007]. To disambiguate the interpretation of sentence acceptability and orient annotators in giving their judgments, the survey guidelines encouraged them to think if they found the sentence natural in Italian and if they would have used it in a real conversation or any other communicative context.

For the **cloze test task**, participants were required to supply the missing element choosing among the proposed options plus the one "none of the previous options is suitable". Each survey was completed by 20 annotators on average. The number of annotations per sentence in the acceptability task ranges from 16 to 21 and for the cloze task from 18 to 21. To improve data quality, we discarded annotators who took less than 10 minutes to complete the test, considering the average threshold time for each survey. This led us to reject 5 annotators only for the acceptability task.

Table 9.6 reports the average human score and standard deviation obtained by the acceptable and unacceptable sentences. For the latter, we separately computed these scores for the subset of sentences which were also labeled as ungrammatical (see Section 9.4.1). As it can be seen, humans perform very well on the task assigning quite higher scores to the acceptable sentences with respect to the unacceptable ones, also with little variability. Within the unacceptable subset, the slightly smaller score received on average by ungrammatical sentences provides further evidence that humans are sensitive to this distinction.

Also for the **cloze test task** the human evaluation confirms the validity of the resource. Indeed, as shown in Table 9.7, the target connective was largely chosen by the majority of annotators as the most adequate one, although for $\sim$20% of sentences humans preferred the plausible candidate or the two options got half annotations each. The percentage of sentences for which the majority label was given to an implausible choice is largely negligible.

Cloze task choice	N. Items	(%)
Target	215	79.63
Plausible alt.	46	17.04
Implausible alt.	4	1.48
Target=Plausible alt.	5	1.85

Table 9.7: *Number and % of sentences for which the majority label was assigned to the target connective, to the plausible alternative, to an implausible alternative or equally balanced between the target and the plausible alternative.*

AcceptabilityLabel	AvgPPL	minPPL	maxPPL
Acceptable	42.512	2.059	455.961
NonAcceptable	78.280	3.534	390.824
NonAccept+Agr	98.992	9.933	1178.162

Table 9.8: *Average, minimum and maximum perplexity value given by the model to the acceptable, unacceptable and unacceptable+ungrammatical sentences.*

9.4.3 Testing the sensitivity of Neural Language Models to connectives

Once built the new evaluation dataset, we performed some preliminary analysis aimed at testing the performance of NLMs in the two tasks. Specifically, we performed two distinct evaluations. For the acceptability assessment task, we computed the perplexity (*PPL*) score assigned by the GePpeTto model to all sentences of the corresponding section. We assumed that higher *PPL* scores should be assigned to sentences labeled as unacceptable with respect to their original version. The sentence-level *PPL* was calculated using the formula reported in Sec. 9.2.

By inspecting the results in Table 9.8, we observed that the average *PPL* score assigned to the acceptable sentences is quite lower than the one assigned to the unacceptable ones (i.e. 42.512 vs 78.280).

As expected, for the subset of unacceptable sentences, perplexity was on average higher for the ones marked as ungrammatical (98.992), reflecting the model's capability of encoding syntactic phenomena. Interestingly, among unacceptable sentences, those obtaining lower *PPL* scores were perfectly well-formed but with an implausible meaning, as in the case of:

> *Il film 'Le chiavi di casa' ha partecipato al Festival del Cinema di Venezia di quest'anno,* ***perché*** *non ha vinto nessun premio* ($PPL = 13.892$).

To compare humans and model performance, we also computed the Spearman's rank correlation (ρ) between the average acceptability score given by annotators and the *PPL* score assigned by the model to the same sentences. Although limited to this analysis, the resulting very weak correlation (i.e. $\rho = -0.120$, p-value < 0.01) suggests

Predict.	**10_match**		**1st_match**	
Target	(85)	31.48%	(111)	41.11%
Pl. alt.	(12)	4.44%	(23)	8.52%
Target+Pl. alt.	(148)	54.81%	–	–
Other	(25)	9.26%	–	–

Table 9.9: *(Number) and % of BERT's completions in which only the target, only the plausible alternative, both of them or none of them (Other) occur in the first 10 predictions (10_match). (Number) and % of the completions in which the target and the plausible alternative were predicted with the highest probability are also reported (1st_match).*

that connectives differently impact on the ability of humans and models to assess the plausibility of a sentence.

As for the **cloze task** test, we relied on the pre-trained Italian version of the BERT model already used in the experiments of Sec. 8.5 and 8.6. We extracted the first ten completions provided by the model trough the Masked Language Modeling task (MLM) for each sentence, along with their probabilities. This allowed us to inspect whether and in how many cases either the target connective or the plausible alternative appear in the top-ranked predictions.

As shown in Table 9.9, for the large majority of cases BERT is able to infer in its first 10 predictions that the sentence should be completed with a correct connective. That happens in 86.29% of the sentences for the target, resulting from the sum of the cases where only the target occurs in the completions (31.48%) with the cases in which both the target and the plausible alternative were predicted (54.81%), and in 59.25% for the plausible alternative (that is 4.44% plus 54.81%). Focusing instead on the first completion for each sentence, we observe that in almost half of the sentences BERT assigns the highest probability to the original connective (41.11%) or to the plausible one (8.52%).

We are currently performing a more qualitative analysis to better investigate the cases in which the correct connective hasn't received a high probability score, as well as those in which neither of the two options appeared at all (i.e. *Other* cases in Table 9.9), in order to understand whether the other completions can still be considered as plausible ones. Preliminary findings showed that, among the *Other* cases, about 57 of the completions provided by BERT are unacceptable and 34 of them are dubious acceptable i.e. not clearly recognizable as acceptable[7], as in the case of the following sentence[8]:

> *Secondo gli esperti, in Italia i giovani leggono meno i giornali rispetto ai giovani di altri Paesi europei, ... rispetto agli anni passati i giovani tra i 14 e i 19 anni leggono più spesso i giornali.* [**perché** anche *però*].

[7] Note that in order to assign the acceptability label of each completion we refer to the usage of the Italian language as standard as possible.

[8] The unacceptable completion is marked in bold, the dubious acceptable one is reported in block and the original connective is indicated in italics.

Nevertheless, the majority of *Other*'s completions can be considered as acceptable ones. In fact, BERT predicted a word leading to the same meaning (or, at least, very similar) to the original sentence in more that 60 cases. Moreover, in most cases (i.e. 93) the completions provided are plausible ones, although in some of them the sentences acquire different meanings.

9.5 Discussion

The experiments devised in Sec. 9 allowed us to focus on the relationship between linguistic competence and the information learned by NLMs within their internal mechanisms without relying on task-oriented approaches. In Sec. 9.2 we proposed an investigation of the linguistic phenomena characterizing the perplexity of GPT-2 and a pseudo-perplexity metric computed for the BERT model. We first reported robust correlations between GPT-2's perplexity and the sentence-level likelihood computed with BERT. This is a quite prominent result, especially considering that these two metrics are differently computed as a consequence of the two NLMs architectures. Then, we found the effectiveness of our set of linguistic features in predicting the perplexity of the two NLMs, especially for shorter sentences. Despite similar trends, we observed some differences between the two NLMs both at the level of regression accuracy and in the rankings of the features exploited in the prediction of perplexity. GPT-2's perplexity is better captured by the considered features and it resulted to be more affected by lexical parts-of-speech and features capturing the vocabulary richness of a sentence. On the contrary, BERT's perplexity seems to be best predicted by syntactic features highly sensitive to sentence length.

In a follow-up study, we focused instead our analysis on the relationship between NLM perplexity and the scores assigned by a readability assessment tool to each sentence extracted from several datasets differing at the level of textual genre and language variety. Results showed that comparing the rankings obtained using the two metrics we cannot find any significant correlation, either between the scores of the two metrics or with respect to the set of linguistic features that mostly impact their values.

By introducing a new evaluation dataset for Italian designed to test the understanding of textual connectives in real-usage sentences, we showed that in several cases the NLMs are capable of distinguishing between acceptable and unacceptable sentences, thus suggesting their ability to encode sentence meaning within their internal mechanisms. However, it remains unclear to what extent these models rely on semantic acceptability features, since we observed cases in which they fail to recognize implausible meaning of perfectly grammatical sentences.

CHAPTER 10

Modeling Linguistic Abilities in Humans and NLMs

The experiments described in the previous chapter allowed us to investigate in detail the amount of linguistic competence encoded by the most recent NLMs. Exploiting different approaches, mostly based on a set of linguistic features that have proven to be highly predictive in tracking the evolution of L1 and L2 learners' linguistic competence across time, we showed that these models are able to implicitly encode a variety of language phenomena within their internal mechanisms. Nevertheless, there are still many open questions about their ability to learn linguistic properties. For instance, while the vast majority of previous studies focused on the inner working of NLMs testing their abilities to recognize specific linguistic phenomena (e.g. Subject-Verb agreement, negative polarity items) [Linzen et al., 2016, Wilcox et al., 2019, Kann et al., 2019, Warstadt et al., 2019] either on gold annotated [Liu et al., 2019a, Hewitt and Manning, 2019, de Vries et al., 2020] or on artificially created data [Yin et al., 2020, Li et al., 2021], relatively little work has been done in order to interpret NLMs linguistic knowledge considering authentic texts.

Starting from this premise, in this chapter we present a study we devised in order to test the robustness of the BERT model against non-standard forms emerging in authentic texts. In particular, relying on the errors manually annotated on the *CItA* corpus, we designed three sets of experiments with the aim of investigating: i) whether and in which layer BERT internal representations are able to discern the presence of a specific learner error; ii) how and to what extent BERT is robust to non-standard linguistic forms analyzing how its internal representations and attention heads; iii) how learner errors affect the ability of the model to implicitly encode linguistic knowledge.

10.1 Investigating NLM's Robustness to Non-standard Linguistic Forms

In this study, we decided to focus on the less considered typology of authentic texts by proposing an extensive interpretation study aimed at understanding the implicit behaviour of one of the most prominent NLM, BERT [Devlin et al., 2019], when dealing with a particular type of noisy texts, namely essays written by native language learners. Our idea is that this typology of authentic texts represents a very interesting and challenging testbed to deeply assess the robustness of NLMs since they contain both standard and non-standard productions, i.e. erroneous uses of standard linguistic forms. Although authentic learner corpora have been already used to interpret the working principles of NLMs, the interest has been largely focused on evaluating how these models solve specific downstream tasks, i.e. to improve Grammatical Error Detection (GED) [Bell et al., 2019, Kaneko and Komachi, 2019] and Grammatical Error Correction (GEC) systems [Grundkiewicz et al., 2019, Kaneko et al., 2020]. On the contrary, to the best of our knowledge, less attention has been paid to interpret the inner mechanisms and the robustness of these models before any fine-tuning on these data.

Our interest about the robustness of a NLM in the pre-training stage stems from the still open and largely discussed issue concerning the relationship between the information encoded in a representation and the information a model uses to solve specific downstream tasks. Therefore, we first provide a comprehensive analysis of how non-standard linguistic forms are encoded in the pre-trained model by inspecting its inner mechanisms from different perspectives with the aim of understanding whether, and to what extent, the internal representations diverge when the model is exposed to incorrect and correct forms. Based on the acquired evidence, we then try to assess the impact of errors on the model's linguistic competence. Our intuition is that investigating and quantifying in the pre-training stage the inner mechanisms of a NLM and its linguistic abilities on texts containing non-standard linguistic forms should be of extreme interest for future studies about NLMs' robustness in downstream tasks. Although this issue is still highly debated [Ravichander et al., 2021], it has been demonstrated that introducing linguistic information [Zhou et al., 2020, Bai et al., 2021], during the pre-training phase enhances model's performances. In addition, in our previous experiments (see Chapter 8) we showed that there is a significant correlation between the degree of linguistic knowledge a NLM implicitly acquires in the pre-training about a wide range of both local and structural phenomena specific of a sentence and its ability to solve correctly a downstream task where such linguistic knowledge is highly involved.

To investigate the robustness of the model, we test BERT on the CItA corpus. The choice of relying on this corpus has been explicitly driven by two main motivations. First, as showed in Sec. 5, CItA is supplied with the manual annotation of learner errors and their corrections, carried out according to a three-leveled annotation scheme where each label targets a specific of language competence, i.e. grammar, orthography and lexicon. The variety of errors contained in the corpus makes it particularly suitable to investigate whether, similarly to what happens for human readers, also for a NLM there is a ranking of robustness in coping with specific errors corresponding to different linguistic domains. As a large body of empirical studies in psycholinguistics suggests, some syntactic errors are not easily detected even by humans, such as those involving agreement attraction phenomena in the comprehension of long-distance constructions [Bock and Miller, 1991, Franck et al., 2002]. On the other hand, humans are extremely robust to cope with real input noise and they can derive a meaningful representation of a sentence in spite of the presence of spelling and grammatical errors, homophones replacement, newly introduced words or even an existing word used in an unfamiliar or a new context; instead, all these types of small input perturbations have been shown to negatively affect NLP systems, for instance, current Neural Machine Translation models [Belinkov and Bisk, 2018, Xu et al., 2021]. Second, authentic texts written by L1 learners are not affected by the interference of a pre-existing language and the peculiarities they exhibit can reflect both errors deriving from a still immature writing competence and less acceptable forms that could be admissible in informal spoken language but not in formal writing. From this point of view, these productions offer the possibility to test whether and to what extent NLMs show a sensitivity to linguistic errors which is comparable to the native speaker's one and if the presence of errors impacts on their ability to identify the correct structure of a sentence.

We believe that a comprehensive investigation of the robustness of a neural language model against noise data should be pursued accounting for more than one of the interpretation techniques defined in the literature. This represents a further innovative characteristic of our approach: in fact, the aim of all proposed experiments is not only to assess BERT's abilities to detect errors in L1 students written productions, but also to provide a better understanding of its internal mechanism by showing whether and how BERT's representations and distributional patterns of attention heads are affected by the presence of a specific error, how they change between the wrong and the corrected version of the same sentence and which typologies of error affect more the linguistic properties that the model has implicitly learnt of a sentence.

The contributions of our work are as follows:

1. we studied the behaviour of the pre-trained BERT model when dealing with authentic written productions by L1 learners containing non-standard linguistic forms (errors) and their corresponding correction;

2. differently from previous work focused on the interpretability of NLMs, we investigated how BERT perceives errors relying on multiple interpretation techniques, ranging from the definition of *probing tasks* to the analysis of word- and sentence-level representations and attention heads;

3. based on the robustness of the model in coping with non-standard linguistic forms, we studied whether it is possible to obtain a ranking of errors which corresponds to a specific area of language knowledge that a learner has to master (i.e. grammar, orthography and lexicon);

4. we studied the relationship between the presence of certain typologies of linguistic errors in a sentence and BERT's ability to correctly encode within its internal representations a set of linguistic phenomena characterising that sentence.

10.2 Neural Language Models and Noisy Text Data

Particularly relevant to the main focus of our study is the amount of previous studies focused on the analysis of state-of-the-art NLMs when dealing with noise in texts. In particular, these works are mainly focused i) on the analysis of how pre-trained NLMs perform in specific downstream tasks when fine-tuned on noisy text data [Sun et al., 2020a, Kumar et al., 2020] and ii) on the definition of new approaches to increase model robustness to such data [Belinkov and Bisk, 2018, Malykh, 2019, Namazifar et al., 2021]. For instance, [Kumar et al., 2020] tested BERT's performance on sentiment analysis and textual similarity by gradually introducing spelling mistakes and typos on the benchmark datasets, showing that noise clearly affects the performance of the model. It should be noted that the vast majority of these studies are focused on a single phenomenon (e.g. spelling mistakes) while few of them take into account how different sources of linguistic noise impact on these models. The study by [Yin et al., 2020], which is partially related to our, represents an exception in this context: the authors proposed an approach to automatically simulate various types of grammatical errors and analyzed how these different types affect downstream tasks. Specifically, by relying on a rule-based method to mimic eight of the most frequently occurring grammatical errors in the NUS Corpus of Learner English (NUCLE) [Dahlmeier et al., 2013], they: i) investigated NLMs robustness to noises by evaluating them on different downstream tasks; ii) quantified NLMs capacities of identifying grammatical errors by probing individual layers through a linguistic acceptability task; iii) studied how models capture the interaction between grammatical errors and context. Their experiments showed that the tested models (i.e. ELMo, BERT and RoBERTa) are influenced by ungrammatical inputs and that errors related to word choice and subject-verb agreement are the most harmful types. Furthermore, probing BERT's abilities in identifying grammatical errors, they demonstrated that middle layers are better in identifying errors than lower layers, although higher layers are better suited for locating errors related to long-range dependencies and verbs. More in line with previous studies aimed at testing linguistic knowledge of NLMs, but with a specific focus on the impact of specific linguistic anomalies on the model's inner mechanisms, [Li et al., 2021] proposed a different approach to test NLM's abilities in a grammaticality judgement task. They introduced a new probing tool based on a Gaussian model which is trained to fit distributions of embeddings at each layer of three transformer models. They aimed at understanding whether NLMs show different surprisals in their internal layers when exposed to linguistic anomalies. After evaluating their method on the BLiMP dataset [Warstadt et al., 2020], they studied whether NLMs exhibit different behaviour corresponding to different classes of anomalies and showed that morphosyntactic anomalies produce high surprisal already in the early layers of the models (from layers 3-4).

Even if we share with these previous studies the attention to noisy data, we differ in two main aspects: i) we relied on authentic texts, rather than on artificially created ones, to investigate how a NLM perceives erroneous linguistic forms, ii) we concentrated our analysis on the interpretability of a pre-trained model, rather than on a fine-tuned one, with the aim of providing evidences and insights about the behaviour of that model when exposed to different typologies of noisy data.

10.3 Neural Language Models and Learner Corpora

The choice of learner corpora as the testbed for our study was motivated by the peculiar and potentially insightful nature of this source of noise data. Indeed, much more than other typologies, these texts exhibit different types of errors or deviant uses from a given language norm that can be also observed across multiple types of real-word texts.

As previously stated, these corpora have been tested with Neural Language Models with the main purpose of improving the performance of GED and GEC systems rather than investigating the impact of specific error typologies on the hidden representations of these models. The only exceptions concern the analysis of the inner mechanisms of GEC/GED systems and their performance on specific error categories [Choe et al., 2019, Kaneko and Komachi, 2019]. For instance, [Kaneko et al., 2020] investigated the characteristics of the hidden representations of a pre-trained BERT model and a fine-tuned one in a grammatical error detection task. In particular, visualizing the hidden representations from the last layer of the two NLMs, they showed that the pre-trained model does not distinguish between correct and incorrect clusters. On the other hand, fine-tuned BERT generates a vector space that can separate correct and incorrect words.

Differently, [Misra et al., 2019] proposed an analysis based on the relatedness of polyglot [Al-Rfou' et al., 2013] and fasttext [Bojanowski et al., 2017] vector representations extracted from L2 (English) erroneous content words in order to investigate whether word embeddings models capture the interference of the first language when a subject processes words in a L2. Specifically, they introduced a new metric (i.e. *EPNO*) to quantify the semantic relatedness between correct-incorrect words in terms of their nearest neighbors in the vector space. Computing the Spearman's correlation between *EPNO* values of L2 words and the translations in the corresponding L1 (e.g. Catalan, French, German), the authors showed that the incorrect-correct word pairs that are highly overlapping with each other in a person's L1 are also highly overlapping in English, thus indicating equal strength between the similarities in L1 and L2.

Rather than focusing on the development of GED or GEC systems, in our study we investigated whether and how errors occurring within a learner corpus are perceived by a NLM and how their presence influences its ability to encode linguistic competence. A further main novelty of our approach with respect to the aforementioned related works is represented by the typology of learner corpora taken into account. Differently from previous studies, we do not rely on essays written by second language learners but by monolingual native speakers. Even if it is out of the scope of this paper to revisit the differences between first and second language learning, as we mentioned in the

Class of error	Type of Modification	Error Code	Frequency
	Grammar		
Verbs	**Erroneous use of tense**	111	505
	Erroneous use of mood	112	259
	Erroneous Subject-Verb agreement	113	208
	Omission	114	1
Prepositions	**Erroneous use**	121	251
	Omission/Redundancy	122	20
Pronouns	**Erroneous use**	131	193
	Omission	132	14
	Redundancy	133	28
	Erroneous use of relative pronoun	134	50
Nouns	**Erroneous gender agreement**	161	36
	Erroneous number agreement	162	48
Articles	**Erroneous use**	141	177
Conjunctions	Erroneous use	151	27
Other		100	177
Total			1,994
	Orthography		
Double consonants	Omission	211	226
	Redundancy	212	132
Use of *h*	Omission	221	97
	Redundancy	222	67
Monosyllables	**Erroneous use of monosyllabic words**	231	306
	po* and *pò* instead of *po’	232	72
Apostrophe	**Erroneous use**	241	190
Capital letter	**Erroneous use**	251	497
Other		200	1236
Total			2,823
	Lexicon		
Vocabulary	**Erroneous use**	311	289

Table 10.1: *Summary of the CItA Error annotation schema.*

introduction, we believe that authentic texts written by L1 learners can be particularly challenging as they are not affected by the interference of a pre-existing language. Indeed, if errors by L2 learners have been more codified in the literature [Heydari and Bagheri, 2012,Richards, 1971] and part of them can be somehow more predictable based on the similarities or differences between the source and the target language, we can expect that L1 learners would make a variety of errors reflecting both a still incomplete mastery of lexical, grammatical and spelling skills and less acceptable forms that could be admissible in informal spoken language but not in formal writing.

10.4 The CItA corpus

For the purpose of our study, we relied on the *CItA* corpus. As we discussed in the experiments we devised in Section 5, a main peculiarity of the corpus is represented by the annotation for writing errors made by students along with their corresponding corrections manually performed by a secondary school teacher. It is worth pointing out here that the annotation scheme devised to mark the writing errors was specifically meant to intercept authentic deviations from the Italian linguistic norm, as defined by the literature on the evaluation of written skills of Italian as first language. As shown in Table 10.1, the scheme covers different classes of errors corresponding to several areas of Italian language competence. This is the reason why we chose the *CItA* corpus to test how different learner errors are perceived by BERT. Specifically, the schema is articulated according to the following dimensions: (i) the macro-class of error, i.e. grammatical, orthographic and lexical; (ii) the class of error, i.e. the linguistic element involved (e.g. verbs, pronouns, monosyllables); (iii) the type of error (e.g. the erroneous use of tense or mood in verbs, the redundancy of pronouns), which corresponds to a specific error code.

As we can notice in Table 10.1, the macro-classes of errors have different distributions in the corpus with a main prevalence of the grammatical and orthographic ones, while the erroneous uses of lexicon are less represented. Specifically, as regards grammatical errors, the most common type of error concerns the incorrect use of verbal tenses, as in the following example[1]:

> Erroneous sentence: *Poco prima dell'inizio della scuola, mentre leggevo un libro,* ***viene*** *mia madre in camera mia [...]* [lit. Shortly before the beginning of school, while I was reading a book, my mother **comes** into my room [...]]
>
> Corrected sentence: *Poco prima dell'inizio della scuola, mentre leggevo un libro,* ***venne*** *mia madre in camera mia [...]* [Shortly before the beginning of school, while I was reading a book, my mother **came** into my room [...]]

In this case, the student erroneously used the present tense *viene* ('comes') after the imperfect tense *leggevo* ('was reading') thus violating the correct sequences of tenses.

Among the orthographic errors, the most frequent class is the *Other* one, which includes all the errors that not belong to any of the other classes of this macro-class. This is the case of the following sentence where the verb *lascerà* ('will leave') is erroneously misspelled with a redundant character:

> Erroneous sentence: *Durante la nostra crescita si possono perdere e conoscere nuovi amici, ognuno di loro ci* **lascierà** *qualcosa e ci arricchirà.* [lit. During our growth we can lose and meet new friends, each of them will **leeave** us something and enrich us]
>
> Corrected sentence: *Durante la nostra crescita si possono perdere e conoscere nuovi amici, ognuno di loro ci* **lascerà** *qualcosa e ci arricchirà.* [During our growth we can lose and meet new friends, each of them will **leave** us something and enrich us]

Note also that a single sentence of the corpus may contain more than one error, possibly belonging to different macro-classes and involving the erroneous use of different linguistic elements. This is for example the case of the following sentence, where one orthographic errors and two grammatical errors occur:

> Erroneous sentence: *Mi piacerebbe* **racogliere** *molti sassi e giocare con gli amici a palla a volo e mi piacerebbe andare ma mia mamma non* **gli** *va e non mi vuole mandare con la nonna.* [lit. I would love **to pik up** lots

[1] In all the examples, we present the erroneous sentence (*Erroneous sentence*) and the corresponding corrected version (*Corrected sentence*). The erroneous token and its corresponding corrected version are marked in bold. Whenever possible, we tried to provide a translation showing how the error would appear in English.

Table 10.2: *Number of minimal edit pairs in the CItA subset for the 18 typologies of errors.*

Error Code	Sentences	Error Code	Sentences
111	366	100	155
112	228	212	129
113	177	222	64
121	240	231	282
131	174	232	69
134	43	241	118
141	162	251	453
161	33	200	1,052
162	41	311	256

of rocks and play volleyball with friends and I would love to go but my mom **to him** doesn't feel like it and won't send me with grandma.]

Corrected sentence: *Mi piacerebbe* **raccogliere** *molti sassi e giocare con gli amici a palla a volo e mi piacerebbe andare ma* **a** *mia mamma non va e non mi vuole mandare con la nonna.* [I would love **to pick up** lots of rocks and play volleyball with friends and I would love to go but my mom doesn't feel like it and won't send me with grandma.]

The first error concerns the misspelling of the verb *raccogliere* ('to collect'), which requires a double consonant, the second one corresponds to the erroneous omission of the preposition *a* ('to'), while the last one involves a redundant clitic pronoun *gli* ('to her'), which should be omitted to have a fully grammatical sentence.

For the purpose of the experiments carried out in this study, we modified the original *CItA* corpus in order to pair each erroneous sentence with one or more corrected sentences, each containing only one single local edit. The resulting corpus thus consists in a collection of *minimal edit pairs*, one for each error type, as shown in the following example, where the original sentence ("*La parco cè tante persone*" [lit. The park there is many people]), which contains two different types of grammatical errors, i.e. the erroneous use of preposition (error type=121) and of subject-verb agreement (error type=113), has been included in two pairs, each corresponding to one of the two original errors:

Erroneous sentence (121): ***La*** *parco ci sono tante persone.* [lit. **The** park there are many people.)

Corrected sentence: ***Al*** *parco ci sono tante persone.* [**At** the park there are many people.]

Erroneous sentence (113): *Al parco* ***cè*** *tante persone.* [lit. At the park **there is** many people.)

Corrected sentence: *Al parco* ***ci sono*** *tante persone.* [At the park **there are** many people.]

We also excluded all the error types that are less represented in the corpus, i.e. those occurring in less than 20 sentences. Moreover, in order to avoid mismatches between pairs of original-corrected sentences, we excluded all errors with *Omission* as type of modification. At the end of this process, we ended up with a collection of 4,042 *minimal edit pairs* distributed in 18 different typologies of errors as showed Table 10.2.

10.5 Methodology

In order to inspect how the presence of an error affects the corresponding sentence representation encoded in the pre-trained model and to what extent the model is robust against errors, we designed the following three sets of experiments:

Probing for error detection abilities We studied whether and in which layer BERT internal representations are able to discern the presence of a specific learner error within our *CItA* subset of sentences. To this end, we devise an experiment aimed at verifying if a linear model trained with BERT's layer-wise internal representations can detect the erroneous sentence in a *minimal edit pair* (sentence-pair classification task).

Analysis of attention heads and internal representations We further investigated how and to what extent BERT is robust to non-standard linguistic forms analyzing how its internal representations and attention heads behave when exposed to the 18 different typologies of learner errors and their corresponding revisions.

Probing for linguistic competence We analyzed how learner errors affect the ability of BERT to implicitly encode the linguistic knowledge of a set of linguistic features characterising a sentence. For this purpose, we trained a probing model on sentences extracted from the *CItA* corpus that did not contain errors and tested it on our 18 sets of minimal edit pairs, each corresponding to a different type of errors.

As regards the contextualized NLM, we choose the pre-trained Italian version of BERT already used in the experiments on the Italian models described in Chapter 8.

10.6 Is BERT able to recognize learners errors?

For the purpose of probing the model's ability to discriminate erroneous vs. corrected sentences, we devised a sentence-pair probing task[2]. In particular, we relied on a Linear Support Vector Classifier (LinearSVC) that, for each type of error, takes as input layer-wise BERT internal representations[3] of each sentence in a minimal edit pair and predicts which sentence of the pair is the erroneous one. We devised three sets of experiments: i) in the first one we built a classifier using all sentences in the *CItA* corpus, without distinguishing them into distinct subsets of error typology, ii) in the second one, we built three binary classifiers, one for each of the three macro-categories of Grammar, Orthography and Lexicon; ii) in the third one, we train a classifier for each error type. In all cases, the classifier is provided with: 50% of the samples in the following order $(s_{erroneous}, s_{corrected})$ and 50% in the reverse order $(s_{corrected}, s_{erroneous})$. Moreover, since the amount of sentences varies according to the type of error involved, we designed

[2] We also tested BERT's competence to identify sentences with an error relying on a single-sentence probing task observing poor performance (i.e. accuracy scores below a random baseline), thus suggesting that pre-trained BERT is not able to implicitly identify learner errors by looking only at the internal representations extracted from individual sentences.

[3] We relied on the activation of the first input token *[CLS]*, that has been shown to summarize the information encoded in an input sequence [Jawahar et al., 2019].

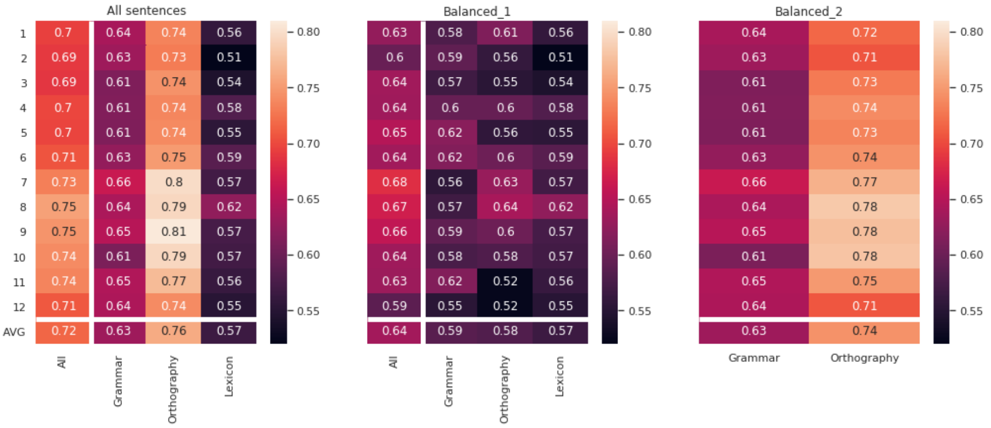


Figure 10.1: *Sentence-pairs probing scores (in terms of accuracy) for each BERT's layer (rows) and according to: i) all sentences in the dataset (All); ii) sentences belonging to the three macro-categories of Grammar, Orthography and Lexicon. Experiments were performed considering all the sentences available for each macro-category (All sentences) and balancing each macro-category according to a fixed number of sentences (Balanced_1 and Balanced_2). Average accuracy scores are also reported (rows AVG).*

two different experimental settings: a first one where all the set of sentences of each macro-category of error were used and a second one where we balanced all the datasets.

Figure 10.1 reports the classification results considering the first two experiments. In particular, the *All sentences* heatmap contains the results obtained using all sentences, the *Balanced_1* heatmap shows the accuracy scores obtained by undersampling the datasets to the smallest macro-category among the three ones, i.e. the Lexicon category that includes 256 sentences. The *Balanced_2* setting differs with respect to the macro-category of errors considered for the undersampling. In this second case, we considered only the two most numerous macro-categories, i.e. Grammar and Orthography, and we reduced the datasets to the smallest one represented by the Grammar category with 1,502 sentences. The evaluation was performed using a 10-fold cross validation and accuracy as evaluation metric.

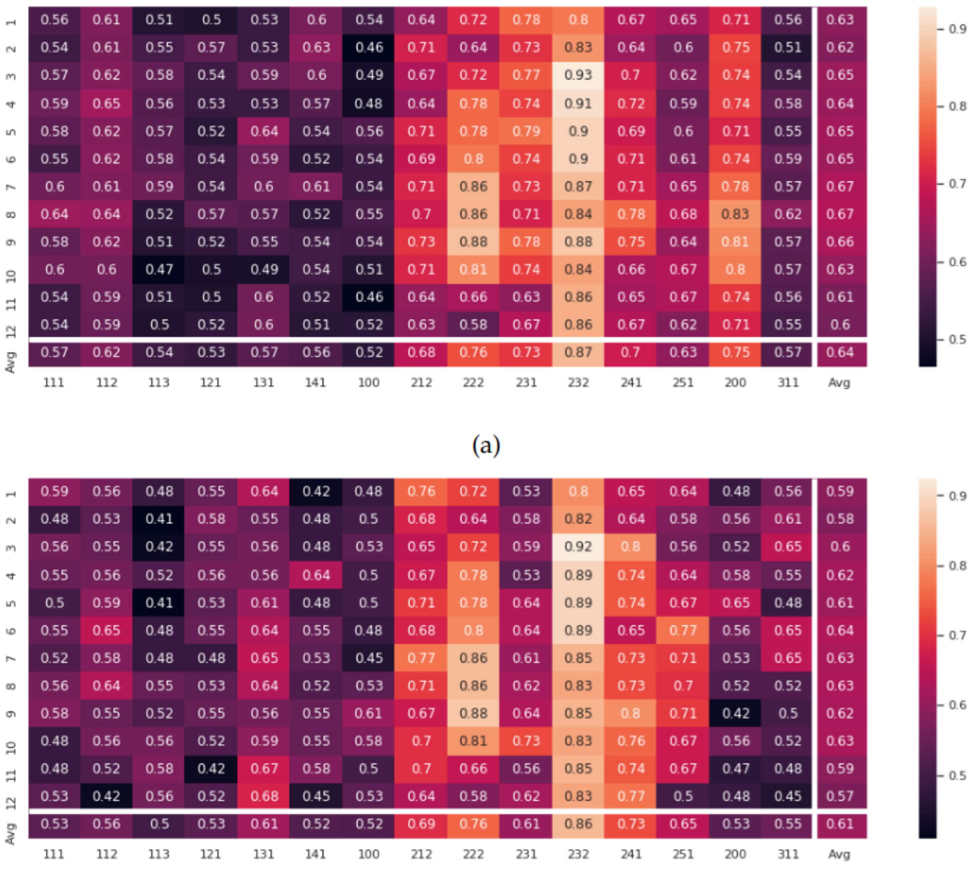
As we can see, classification results of the first experiment (columns *All* in the heatmaps) are all above a majority baseline (0.50) thus showing that BERT is able to distinguish an erroneous from a corrected sentence despite the typology of error, at least to a moderate extent. As expected, the accuracy scores when the classifier uses BERT's representations of the balanced datasets are slightly lower. In line with what observed by [Yin et al., 2020] in their study on the robustness of NLMs in an Grammatical Error Detection task, in both settings, the best scores are achieved in the middle layers, specifically between layer 7 and 9. The overall accuracy tends to decrease as far as the last layers are approached.

If we consider the results of the second experiment, i.e. the one taken into account the tree macro-typology of errors separately, we observe a clear distinction between the accuracy scores reported in the unbalanced and balanced setting. Specifically, in the first one when all sentences of each category were used (*All sentences*), we can see that orthographic errors are the best identified category while errors related to grammar and lexicon are the most difficult to recognize. This seems to suggest that the representations of erroneous and corrected sentences with grammatical and lexical errors and their corresponding revisions are less distinguishable than those related to orthographic errors. If we focus instead on the same results achieved in three-fold classification but with the balanced datasets, we can observe that in the *Balanced_1* configuration the accuracy scores tend to decrease significantly, although they all remain above the majority baseline. Moreover, we note that the distinction into three macro-categories is no longer appreciable, since the scores obtained by the three probing classifiers are more or less similar to each other. On the contrary, the ranking among the macro-categories of errors in terms of classification accuracy is maintained when we balance the datasets with respect to the two biggest macro-categories (*Balanced_2*). Namely, sentences containing orthographic errors are, on average, easier to distinguish than those containing grammatical ones. This seems confirming the intuition that the drop of accuracy obtained with the *Balanced_1* setting is possibly due to the very different size of the Lexicon macro-category with respect to the categories including the other types of errors. In addition, this outcome suggests that a more fine-grained investigation is needed in order to investigate whether there is a ranking of BERT's robustness in coping with individual types of errors.

Thus, in order to deepen the investigation, Figure 10.2 reports the results of the third experiment consisting in a binary classification task for each error type. To achieve reliable results, we decided to consider only those types occurring in at least 64 sentences in the input corpus, for a total of 15 types (see Table 10.2). In particular, Figure 10.2 (a) shows the accuracy scores obtained considering all the sentences available for each of the 15 types, while Figure 10.2 (b) presents the results achieved considering datasets that contain the same amount of sentences (i.e. 64) for each typology. Despite the different amount of input data, we can observe a quite similar trend of accuracy. The scores achieved considering all the amount of sentences available for a specific error typology are in line with those achieved with input data of smaller and comparable size, thus suggesting that there are specific non-standard linguistic forms that BERT is always more able to distinguish from the standard ones. This is confirmed by the rankings reported in Table 10.3, where we compared the ordering of the considered types of error by decreasing mean accuracy score, obtained considering the two experimental settings.

Similarly to what we observed in the previous experiments, in both rankings all types of error belonging to the orthographic macro-category are ranked in the top list, while the grammatical ones occupy the lower positions. Specifically, the top-ranked ones are represented by the incorrect use of the adverbial monosyllabic word *po'* (lit. 'a little') (error code=232) and by the redundancy in the use of *h* (error code=222). For what concerns the former, in the *CItA* corpus *po'* was erroneously written by learners without the required apostrophe as in the following example:

Erroneous sentence (232): *Dopo un* ***po*** *la volpe pensò: "Dato che il leone è indifeso me lo posso mangiare con*



(a)

(b)

Figure 10.2: *Sentence-pairs probing scores for each BERT's layer (rows) and according to the 15 considered types of error (columns). Average accuracy scores are also reported (rows and columns AVG). Figure 10.2 (a) reports the results obtained considering all the sentences available for each specific error typology, while Figure 10.2 (b) shows the results achieved with the balanced datasets, i.e. containing a fixed number of sentences (64) for each error typology.*

calma. [After a **a bit** the fox thought: "As the lion is helpless I can eat it with calm."]

Corrected sentence: *Dopo un **po'** la volpe pensò: "Dato che il leone è indifeso me lo posso mangiare con calma.* [After a **a bit** the fox thought: "As the lion is helpless I can eat it with calm."]

For what concerns the latter, it should be noted that a redundant *h* may turn the affected word into a different, but still existent, word of the Italian lexicon, which can also belong to a different grammatical categories. This is exemplifies by the following pair where the preposition *a* ('at') was erroneously written with a redundant *h*, which corresponds to the third singular person of the auxiliary verb *avere* ('to have'):

Erroneous sentence (222): ***Ha** casa abito con mia madre, mio padre e con mio fratello più piccolo di me.* [**Has** home I live with my mom, my dad and my brother who is younger than me.]

Corrected sentence: ***A** casa abito con mia madre, mio padre e con mio fratello più piccolo di me.* [**At** home I live with my mom, my dad and my brother who is younger than me.]

All sentences		Balanced datasets	
Error code	**Mean score**	**Error code**	**Mean score**
232	0.87	232	0.86
222	0.76	222	0.76
200	0.75	241	0.73
231	0.73	212	0.69
241	0.70	251	0.65
212	0.68	131	0.61
251	0.63	231	0.61
112	0.62	112	0.56
111	0.57	311	0.55
131	0.57	111	0.53
311	0.57	200	0.53
141	0.56	121	0.53
113	0.54	141	0.52
121	0.53	100	0.52
100	0.52	113	0.50

Table 10.3: *Rankings of the best predicted error typologies according to the experiments performed with all the sentences available for each error typology (column All sentences with the balanced datasets (column Balanced datasets). The rankings are obtained averaging probing layer-wise scores, here reported in column Mean score.*

Quite high average classification scores are also achieved in the recognition of two types of orthographic errors that are related to the *po'* error, i.e. the erroneous use of the apostrophe (error code=241) and of other monosyllabic words (error code=231). Both of them are exemplified by the following pairs:

Erroneous sentence (241): *C'è una ragazza che mi piace **d'avvero**.* [lit. There's a girl I **r'eally** like.)

Corrected sentence: *C'è una ragazza che mi piace **davvero**.* [There's a girl I **really** like.]

Erroneous sentence (231): *La volpe molto felice disse di **si**.* [lit. The very happy fox said **itself**.)

Corrected sentence: *La volpe molto felice disse di **sì**.* [The very happy fox said **yes**.]

On the one hand, the high performance shown by BERT in recognizing erroneous sentences containing a misspelled *po'* is probably related to the fact that this classification scenario is potentially easier than other ones. In fact, the corresponding subset for this error typology contains pairs of sentences where the error affects always the same word and has only one possible correction, independently from the surrounding contextual words. The same is true, although with a slightly higher variability, for other types of monosyllabic words, for which the majority of examples in the corresponding subset involves few words (such as *e* for *è*, i.e. the third singular form of the verb "to be"). On the other hand, it is interesting to note that the correct spelling of *po'* and of other monosyllabic words reflect an aspect of the Italian orthographic competence that is quite difficult to master for high-school learners and seem to be persistent even at higher level of education [Cignetti, 2011], as shown by studies on new forms of digital writing [Antonelli, 2012]. Thus, the NLM's abilities in discriminating these typologies of linguistic forms that deviate from the orthographic norm may suggest a distance between BERT's training set, mostly representative of formal writing, and student writings.

If we inspect the lower part of the two rankings we can find all errors belonging to the grammatical macro-category. In particular, the erroneous subject-verb agreement (error code=113), the incorrect use of articles (error code=141) and of prepositions (error code=121) are the errors recognised with the lowest accuracy. On the contrary, the incorrect use of verbal mood (error code=112) is the grammatical error that BERT is most able to distinguish.

Despite the similar trend between the two rankings, there are two main exceptions represented by the orthographic errors that do not belong to any specific class (error code=200) and by the erroneous use of pronouns (error code=131). Namely, when the classification is performed relying on datasets of comparable size for each type of errors (*Balanced datasets*), BERT is less able to distinguish the erroneous/corrected the miscellaneous type of orthographic errors, while is more able to distinguish erroneous pronouns.

Similarly to what observed for the classification results reported in Figure 10.1, Figure 10.2 (b) shows that the best scores are generally obtained around layers 7 and 9, while the average accuracy tends to decrease as far as the last layers are approached. However, there are specific types of error for which top performances are not achieved in the middle layer but rather in the first ones. This is for example the case of the erroneous use of *po'* (error code=232) for which the peak of accuracy is reached at layer 3, in both experimental settings. Finally, we can observe that the drop of classification accuracy in the last layers is particularly evident for the redundant use of *h* (error code=222).

10.7 How does BERT perceive learner errors?

Having observed that BERT, at least to a moderate extent, can implicitly detect the presence of learner errors, although with differences among categories, we proceeded to explore how the model encodes and represents these errors in its components. In particular, we investigated how: (i) attention heads behave according to different typologies of learner errors; (ii) internal representations of erroneous and corrected sentences differ from each other. For these experiments, we performed our analysis considering all the 18 types of error, also including the types that were previously excluded in the classification tasks.

10.7.1 Investigating Attention Heads

The analysis of the attention heads was carried out comparing how they individually target erroneous and corrected tokens contained in the 18 sets of minimal edit pairs. Specifically, for each sentence and for all the 12 heads and layers of the BERT model, we measured the average attention attended to a token with a given category of error:

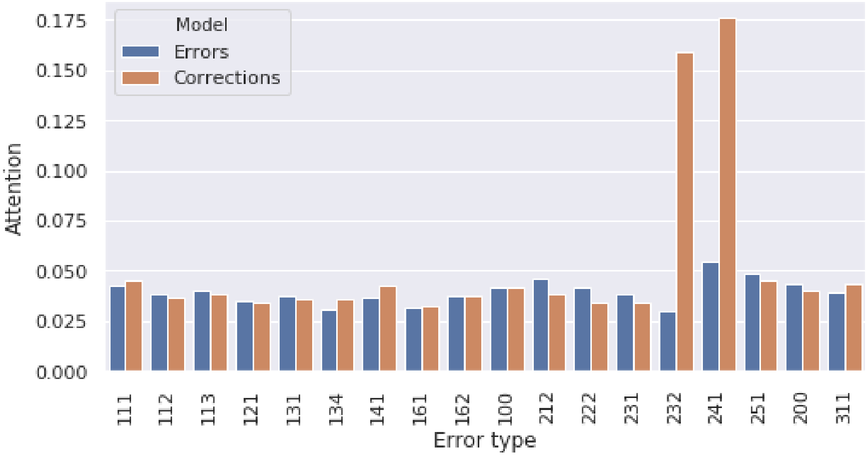


Figure 10.3: *Average attention values considering the pairs of erroneous (Errors) and corrected (Corrections) sentences for each type of error.*

$$A_a(e) = \frac{\sum_{i=1}^{|s|} \sum_{j=1}^{i} a_{i,j}(s)\, [ctg(s_j) = e]}{|s|} \tag{10.1}$$

where *s* is a sentence in the *CItA* subset, $a_{i,j}$ is the attention from s_i to s_j for a specific head/layer and *ctg(*s_j*)* is the error category of the token s_j (e.g. *111* (*Use of tense*)). We then computed average attentions by taking the mean value over the whole dataset.

Figure 10.3 reports average attention values on different error types both considering the set of corrected and erroneous sentences. Average scores were computed by aggregating all heads and all layers in the model. As we can see, BERT's attention behaves quite similarly across different error typologies, with few differences between the attention values of the corrected and erroneous sentences. However, two main exceptions are clearly visible and concern two orthographic errors, i.e. the erroneous use of *po'* (error code=232) and of the apostrophe (error code=241). In these two cases, BERT tends to pay more attention to such tokens when observing their corrected form than the erroneous one. As already observed in the classification task described in the previous section, these are also two errors that BERT is particularly able to recognize. It should be note that, even if they are considered two different types of errors according to the *CItA*'s error annotation schema, they identify two aspects pointing to the same area of the orthographic competence, which concerns the proper use of the apostrophe.

In order to deepen our analysis and to understand how BERT's attention on learner errors changes across layers, we reported in Figure 10.4 the average attention that each BERT's head attends to corrected (*Corrections*) and erroneous tokens (*Errors*) along its 12 layers. We found that the differences between attention heads that focus on *Errors* and *Corrections* become more pronounced in the last layers of the model. In particular, BERT starts to pay more attention to corrected tokens from layer 7. It is also interesting

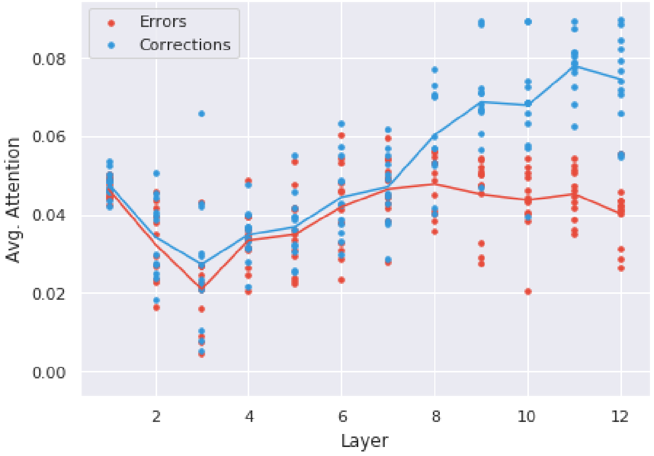


Figure 10.4: *Average attention that each BERT's head (points in the image) puts toward tokens containing errors (Errors) or their corresponding corrections (Corrections) across its 12 layers.*

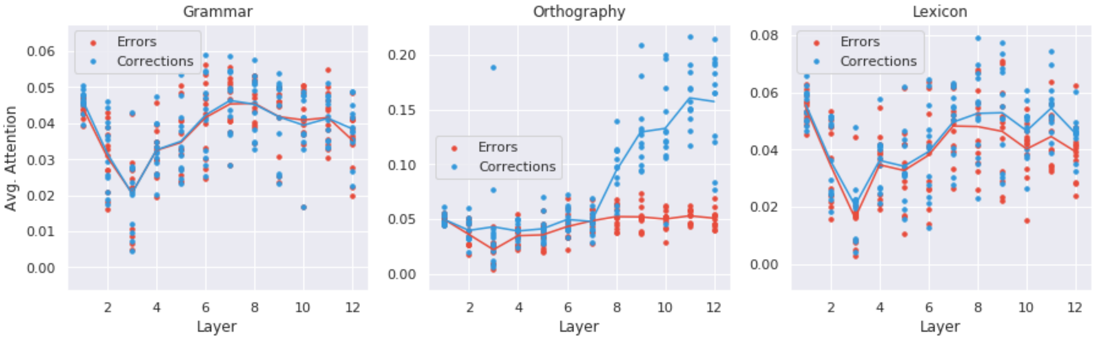


Figure 10.5: *Layerwise attention a particular BERT head puts toward corrected (Errors) and erroneous tokens (Corrections) containing grammatical, orthographic and lexical errors.*

to notice that, although the difference between erroneous and corrected tokens is less sharp in the early heads, layers 1–3 attend less to learner errors and attention tends to decrease for both erroneous and corrected tokens. This seems to be in line with [Clark et al., 2019], where the authors showed that early (and middle) heads tend to attend more to special tokens (*[CLS]* and *[SEP]* than to any other type of token, being it erroneous or correct. If we focus on the impact that the different types of errors have on how BERT's attention changes across layers (see Figure 10.5), we can observe that the variation between attention values in sentences containing an erroneous and corrected token is primarily due to orthographic errors. Our intuition is that this mostly concerns the error types which were distinguished with the highest accuracy in the classification experiments described in Section 10.6, i.e. the erroneous use of the monosyllable *po'* and of the apostrophe, as well as the redundancy in the use of *h* (see Table 10.3). In addition, as the Figure shows, there are some attention heads of corrected tokens that have average values similar to the ones obtained for erroneous tokens and others that

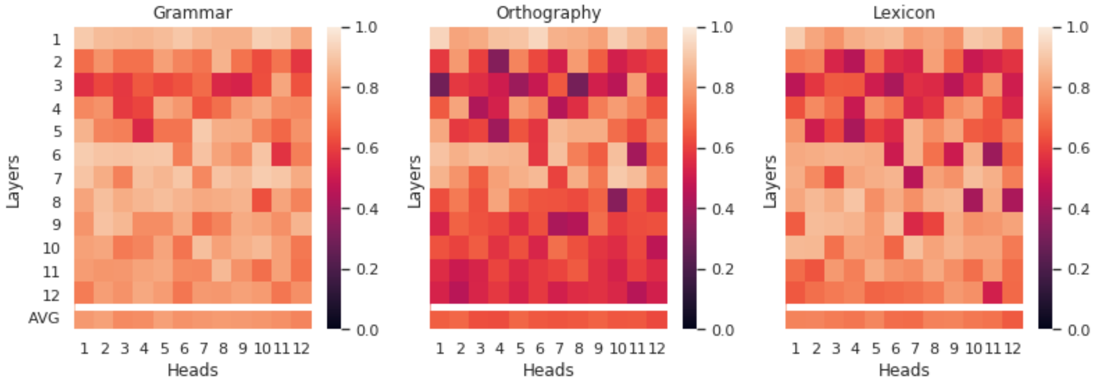


Figure 10.6: *Layer-wise (rows) and head-wise (columns) ρ correlation scores between tokens with errors and their corresponding corrected ones. Scores are reported by averaging ρ scores of tokens containing grammatical, orthographic and lexical errors/corrections. AVG row reports average correlations scores between layers. All correlations are statistically significant (p-value < 0.05).*

Error Code	Spearman Corr.	Std	Error Code	Spearmanr Corr.	Std
161	0.84	0.06	111	0.72	0.11
162	0.80	0.11	112	0.70	0.17
141	0.80	0.10	311	0.65	0.12
113	0.80	0.07	232	0.64	0.21
251	0.76	0.07	212	0.63	0.12
121	0.75	0.10	222	0.63	0.14
100	0.74	0.09	200	0.62	0.12
231	0.73	0.07	134	0.56	0.15
131	0.72	0.12	241	0.34	0.34

Table 10.4: *Spearman correlation scores between attention heads of tokens with errors and their corresponding revisions for each error category. Scores are reported by averaging all attention heads extracted from BERT's output layer. All correlations are statistically significant (p-value < 0.05).*

show a very different behaviour.

In order to investigate if there are attention heads and layers more involved in such a variation, we computed Spearman correlation between the attention values of corrected and erroneous tokens. Figure 10.6 shows layer-wise and head-wise Spearman ρ scores between attention values of tokens containing errors and corrections for the three macro-classes of grammar, orthography and lexicon. As quite expected since in line with our previous findings, the highest correlation scores mainly concern the *Grammar* macro-category, thus suggesting that BERT's attention mechanism does not perceive big differences between tokens containing grammatical errors and those with their corresponding corrections. On the other hand, lower correlation scores on orthographic errors confirm that differences between BERT heads focusing on spelling errors/corrections are more pronounced, especially for what concerns attention values extracted from middle to output layers. Interestingly, we found that low ρ values are also observed in the early layers of the model (layers 2–4), regardless of the macro-class

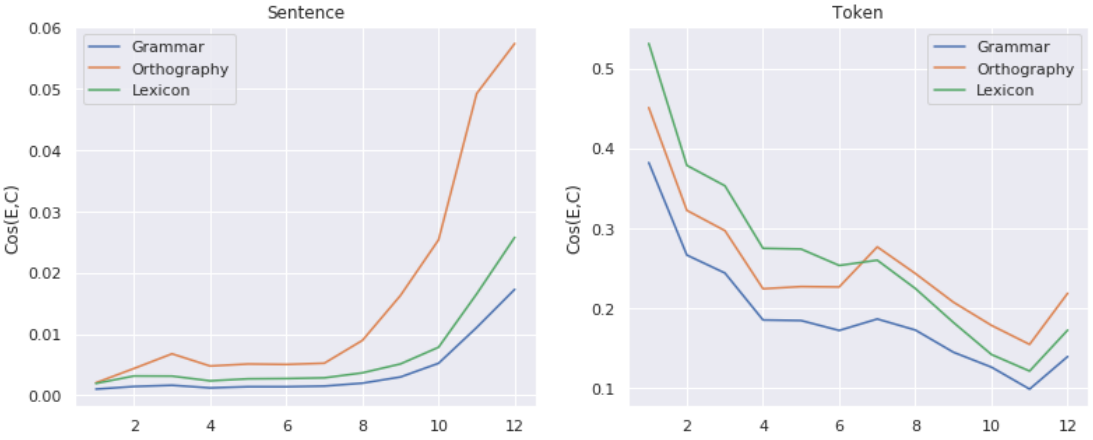


Figure 10.7: *Cosine distances between BERT internal representations of corrected and erroneous sentences and tokens across layers.*

taken into consideration. Taking a closer look at the differences between attention values, we found that there are some heads that are more affected by the presence of errors than others and that this applies indiscriminately to grammatical, orthographic and lexical errors. For instance, head 6 in layer 6 and head 10 in layer 8 exhibit lower average correlations than the other heads extracted from the same layers of the model. We can see this result in line with the findings of previous works [Clark et al., 2019], where it has been observed that specific heads specialize to specific aspect of language, and therefore not all attention heads are equally influenced by a specific phenomenon, such as the occurrence of an erroneous token.

Results reported in Table 10.4 allow focusing this analysis on specific types of error. Namely, the table reports the Spearman correlation scores computed between the attention values of corrected and erroneous tokens, only for the output layer. As expected, variations of attention heads seem to generally reflect the distinction in the three macro-categories that we noticed in the previous experiments. In fact, with the exception of the incorrect use of relative pronouns (error code=134), of capital letter (error code=251) and of monosyllabic words (error code=231), the typologies that exhibit lower correlations are those related to orthographic and lexical errors, while grammatical ones show strong Spearman scores (≥ 0.70). In fact, we can observe that tokens containing an incorrect use of the apostrophe and their corresponding revision (error code=241) are those most differently attended ($\rho = 0.34$), while BERT tends to give the most similar attention to corrected and erroneous tokens implying four grammatical errors, i.e. the incorrect gender and number agreement of nouns (error code=161 and 162), the incorrect Subject-Verb agreement (error code=113) and the incorrect use of articles (error code=141).

10.7.2 Investigating representations

As a further interpretation methodology, we studied how BERT encodes learner errors within its internal representations and across its layers both at sentence and token level. To do so, we computed the cosine distances between the internal representations extracted from corrected and erroneous sentences (*cos(E, C)*) according to the 18 different typologies of errors. As in the probing experiments (see Section 10.6), we relied on the activation of the *[CLS]* token to obtain sentence-level representations. Moreover, to verify the actual impact of specific errors on BERT representations, we measured also the cosine distance between embeddings of the erroneous and corrected tokens. Figure 10.7 reports layer-wise cosine distances between sentence- and token-level internal representations extracted from corrected and erroneous sentences/tokens. The scores were computed by averaging the cosine distances of the three macro-classes of grammatical, orthographic and lexical errors. As regards distances computed considering sentence-level representations, we can observe that they become more marked as we approach the output of the model (from layer 7 to 12). This is in line with the higher scores obtained in the sentence-pair classification task (see Section 10.6), where we observed that BERT's abilities to distinguish an erroneous sentence from its corresponding correction increases in the intermediate layers. Nevertheless, average distance scores are quite low, probably because the embeddings extracted using the activation of the *[CLS]* token tend to mitigate the effect of the error within BERT's internal representations. In contrast, cosine distances computed with the token-level embeddings are more pronounced (from 0.10 to 0.53). This is quite expected, since differences between representations are assessed relying on the exact tokens that tend to vary in each sentence. It is also interesting to notice that the variation of the token-based cosine distances across layers shows an opposite trend with respect to the sentence-based representations extracted using the *[CLS]* token. That is, differences between corrected and erroneous tokens are more pronounced in the early layers of the model and then decrease toward the output, although another peak is reached at layer 7. This trend is similar to that observed with differences between attention heads in the early layers of the model (Figure 10.5). On the contrary, from layer 7 onward the two trends tend to diverge. Specifically, heads of corrected and uncorrected tokens became more pronounced in the last layers of BERT while token-level representations tend to be more similar. Our intuition is that, despite the presence of an error, in the last layers BERT tends to capture in its internal representations the meaning of a word more accurately as the output layer is approached. Thus, the encoding of a token with or without a specific learner error in the last layers of the model will not significantly alter the associated representation. As regards differences between the three macro-categories of errors, we observe that orthographic errors are, in general, those that contribute more to the variation between representations. Despite this, we can notice that cosine distances calculated between the embeddings at token level in the first 6 layers are slightly more pronounced for tokens belonging to the Lexicon category. This might represent a further evidence that BERT already has the linguistic competence to model the semantics of

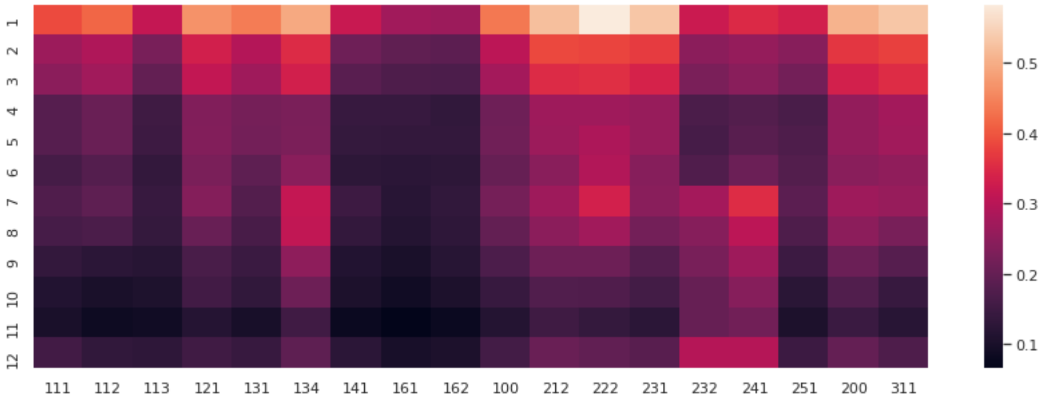


Figure 10.8: *Cosine distances between BERT internal representations of erroneous and corrected tokens for all error categories across layers.*

Code	Slope	r-value	Code	Slope	r-value
311 *	-0.029	-0.92	134 *	-0.019	-0.76
222 *	-0.028	-0.86	111 *	-0.018	-0.84
231 *	-0.027	-0.89	113 *	-0.014	-0.84
212 *	-0.025	-0.88	251 *	-0.014	-0.84
121 *	-0.024	-0.91	141 *	-0.014	-0.80
200 *	-0.024	-0.87	161 *	-0.014	-0.89
112 *	-0.023	-0.90	162 *	-0.011	-0.85
131 *	-0.023	-0.89	232	-0.001	-0.06
100 *	-0.021	-0.87	241	0.000	-0.01

Table 10.5: *Ranking of the error types according to decreasing slope of the regression lines. Correlation coefficients are also reported (r-value). Statistically significant scores (p-value > 0.05) are marked with *.*

tokens also when misused.

The last part of this investigation about BERT's internal representations goes more in detail and it is devoted to assess the impact of the different types of considered errors. Specifically, Figure 10.8 shows cosine distances between representations of corrected and erroneous tokens for all the 18 error typologies. As we can observe, despite the increase of similarity between word representations across BERT's 12 layers, there are some typologies of error that exhibit different trends. This is the case, for instance, of the erroneous use of *po'* (error code=232) and of the apostrophe (error code=241), which both show an increase in cosine distance between representations from seventh to output layer. There are also some types of error that follow the main trend but whose distance between erroneous and corrected representations is extremely high at the first layer. Nevertheless, the distance narrows in the output layer becoming similar to that of the other types. This is for example the case of two orthographic errors, e.g. the redundant use of *h* (error code=222) and of double consonants (error code=212), and of the erroneous use of lexicon (error code=311).

	cos(E,C)	***cos(E,C)***
Code	**(Sentence)**	**(Token)**
241	0.14 ± 0.11	0.30 ± 0.10
232	0.13 ± 0.11	0.30 ± 0.05
111	0.04 ± 0.05	0.15 ± 0.05
200	0.03 ± 0.05	0.20 ± 0.09
251	0.03 ± 0.04	0.15 ± 0.09
311	0.03 ± 0.06	0.17 ± 0.08
212	0.03 ± 0.05	0.21 ± 0.09
231	0.03 ± 0.04	0.18 ± 0.09
141	0.02 ± 0.07	0.13 ± 0.07
222	0.02 ± 0.03	0.20 ± 0.07
100	0.02 ± 0.04	0.16 ± 0.07
131	0.02 ± 0.04	0.14 ± 0.06
113	0.02 ± 0.04	0.13 ± 0.06
112	0.01 ± 0.03	0.13 ± 0.07
134	0.01 ± 0.01	0.19 ± 0.06
121	0.01 ± 0.02	0.15 ± 0.05
162	0.01 ± 0.01	0.11 ± 0.06
161	0.01 ± 0.01	0.10 ± 0.04

Table 10.6: *Cosine distances (cos(E,C)) between representations of sentences/tokens with errors and their corresponding corrected versions. Cosine distances are reported for the internal representations extracted from the output layer (layer 12).*

To further investigate these trends, we computed the slopes of a linear regression line between BERT layers and the cosine distances of corrected and erroneous token representations for each error type. Table 10.5 reports the ranking of error typologies according to decreasing slope values. As it can be noted, error types whose erroneous/corrected representation distances are higher in the first layer show higher values. This suggests that at the beginning BERT is highly sensitive to the difference between the non-standard and standard form, but it rapidly becomes able to cope with the non-standard form thus generating token-level representations more similar the standard ones. In the top part of the ranking we can find the erroneous use of lexicon (error code=311) and three orthographic errors (i.e. the redundant use of *h*, the erroneous use of monosyllabic words and the redundant use of double consonants), while the erroneous use of prepositions (error code=121) is the grammatical error whose representations become more rapidly similar to the corresponding standard ones. On the contrary, error types characterised by a small distance between the erroneous and corrected representations in the first layers are located in the lower part of the ranking. This is the case of three grammatical errors that BERT perceives more similar to the corresponding standard forms, i.e. the erroneous number and gender agreement of nouns (error code=162 and 161) and the use of articles (error code=141), while the orthographic error type with the lowest slope value concerns the use of the capital letter (error code=251). Interestingly enough, the two orthographic errors BERT is more able to classify, i.e. the erroneous spelling of the monosyllable *po'* (error code=231) and the erroneous use of the apostrophe (error code=241), show not significant values.

If we focus the analysis on the output layer, we can observe that the impact of the different types of errors changes according to the level (sentence or token) from which the representations are extracted. The results of this investigation are shown in Table 10.6 that reports the cosine distances between sentence- and token-level internal representations of corrected and erroneous sentences (*cos(E, C)*) for each of the different 18 error typologies computed on the output layer, and ± the standard deviation.

As expected by inspecting the trend reported in Figure 10.7, distances between erroneous/corrected representations at token level are higher than the distances between representations extracted at sentence level. However, it can be observed that most error types are similarly ordered by decreasing cosine distance: for example, the representations of the erroneous use of the apostrophe (error code=241) and of the erroneous spelling of the monosyllable *po'* (error code=232) have the main cosine distance both at the sentence and at token level, while the representations of the erroneous number and gender agreement of nouns (error code=162 and 161) are the most similar ones in both cases. The main exceptions of this ranking concern the incorrect use of *h* (error code=222), of relative pronouns (error code=134) and of prepositions (error code=121), for which high cosine distances obtained at token-level representations do not correspond to the same high distance at sentence level. On the contrary, an opposite trend can be observed for the incorrect us of capital letters (error code=251), of articles (error code=141) and verb tenses (error code=111). In these cases, sentence-level representations of erroneous and corrected forms are more highly ranked than those that those at token-level, suggesting that BERT is more able to perceive as similar the standard and non-standard form when the single token is considered instead of the whole sentence.

If we move to a qualitative analysis and we look more closely at individual occurrences of learner errors within our *CItA* subset, we can notice that, regardless of the category taken into account, there are specific cases that affect more strongly BERT's internal representations extracted at sentence level. For instance, considering the erroneous use of the apostrophe, the cosine distance between the two sentences in the following pair is 0.50:

Erroneous sentence (241): *Perché **c'è** l'abbiamo.* [lit. Because **there is** we have.]

Corrected sentence: *Perché **ce** l'abbiamo* [Because we have **it**.]

This is significantly higher than the distance between the representations of the two sentences contained in the following minimal edit pair, i.e. 0.08:

Erroneous sentence (241): *Vedo **un** ombra gigante.* [lit. I see **an** giant shadow.]

Corrected sentence: *Vedo **un'** ombra gigante.* [I see **a** giant shadow.]

This is probably due to the fact that the erroneous use of the apostrophe can lead, in many cases, to a complete distortion of the original meaning of the sentence. In particular, the incorrect use of the clitic pronoun "*c'*" erroneously written in conjunction with the verb "*è*" ('is') in an existential construction may have led BERT to consider it as the main verb of the sentence instead of "*abbiamo*", thus distorting the whole sentence representation. Similarly, the erroneous use of the article (error code=141) in the following sentence pair yields to cosine distance of 0.11 between the erroneous and corrected sentence:

Erroneous sentence (141): *Oltre tutto **l'**industrie vengono costruite anche in zone dove distruggono l' equilibrio naturale.* [lit. Besides all, **th'** industries are built in areas where they destroy the natural balance.]

Corrected sentence: *Oltre tutto **le** industrie vengono costruite anche in zone dove distruggono l' equilibrio naturale.* [Besides all, **the** industries are built in areas where they destroy the natural balance.]

On the contrary, the distance between the sentence-level representations of the two following sentences containing the same error type is dramatically lower, i.e. 0.0003:

Erroneous sentence (141): *Le gambe mi tremano come* ***dei*** *stuzzichini.* [lit. My legs are shaking like **some** sticks.]

Corrected sentence: *Le gambe mi tremano come* ***degli*** *stuzzichini.* [My legs are shaking like **some** sticks.]

In this case, the erroneous use of the partitive article *dei* in *dei stuzzichini* is more acceptable, since *dei* is a variant of the required form *degli* in agreement with masculine nouns. It follows that this erroneous form may have been frequently observed by BERT during the training phase. On the contrary, the elided form of the feminine article (*l'* instead of *le*) is allowed only in agreement with singular nouns.

The examples reported in this short qualitative analysis suggest that there are specific instances of error that significantly alter BERT's internal representations of the whole sentence. Thus, it could be the case that specific instances have a negative impact on the linguistic competence of the model, such as we hypothesized in the example concerning the erroneous use of the apostrophe in the spelling of the clitic pronoun *ce*. Our intuition is that in this case the model is not able to recognise that, in this specific context, the form *c'è* is a clitic pronoun but in a misspelled form with the apostrophe, since the erroneous form is an homograph of the verb "*è*" ('is') in the existential construction. This might have yielded a representation of the sentence containing the erroneous form quite distant from the representation of the corrected sentence. In order to investigate whether and to which extent the considered types of error negatively affect BERT's ability to encode the linguistic information of a sentence, we carried out further probing tasks described in the following section.

10.8 How do learner errors affect BERT's linguistic knowledge?

In this last section, we investigate how BERT's ability to encode the linguistic information of a sentence changes between sentences containing standard and non-standard linguistic forms. For this purpose, we defined a probing model (LinearSVR) that takes as input layer-wise BERT sentence-level representations (i.e. *[CLS]* token) and outputs the actual value of a specific linguistic feature of the sentence. We relied on 15 different probing features, which were acquired from raw and morpho-syntactic levels of linguistic annotation. In particular, we tested BERT's ability to encode sentence length (*sentence_length*) and average word length (*char_per_tok*), as well as the distribution of the main Parts-of-Speech occurring in our *CItA* subset (e.g. *NOUN*, *VERB*, *PUNCT*, *DET*, etc.). Linguistic annotation were performed using Stanza[4] [Qi et al., 2020]. In order to verify the impact of specific typologies of errors on BERT's linguistic competence, we trained the LinearSVR model on 15.116 sentences of the *CItA* corpus without learner errors and then tested it on the 18 datasets of *minimal edit pairs* previously defined, i.e. datasets containing only one error typology at a time, and its corresponding corrected counterpart. The hypothesis we want to test is that if the model representations extracted from the erroneous sentence are similar to those extracted from the corrected one, the values of the linguistic features should be similar as well. For example, our intuition

[4] https://stanfordnlp.github.io/stanza/

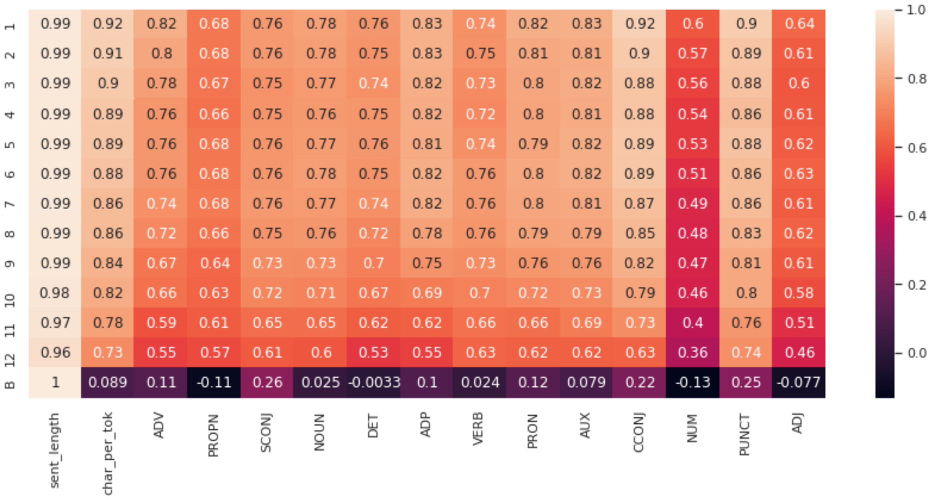


Figure 10.9: *Layer-wise probing scores (ρ correlations) for each probing feature obtained with LinearSVR tested on [CLS] representations of all the corrected sentences. Baseline scores are also reported (row B).*

is that as far as BERT is robust to non-standard forms, it should be able to recognize that the form *ce* is a clitic pronoun given the context in which it appears, even when it is erroneously written with a redundant apostrophe, i.e. *c'è*. As evaluation metric, we used Spearman correlation (ρ) between the values of linguistic features extracted from corrected sentences and the values predicted using BERT's representations extracted from the erroneous and corrected sentences, respectively.

First, in order to inspect BERT's competence in encoding our set of linguistic features, we report in Figure 10.9 the probing results (in terms of Spearman correlation) obtained for each linguistic task relying on the LinearSVR model tested on the representations of all the corrected sentences. We report also the results obtained with a baseline corresponding to a LinearSVR model trained using only sentence length as input feature (row *B* in the Figure). Apart from *sent_length*, we can clearly observe that the scores obtained with BERT internal representations greatly outperform the ones obtained with the sentence length baseline, thus suggesting that the model is capable of implicitly encoding our set of linguistic features. Moreover, we notice that BERT's linguistic competence tends to decrease across its 12 layers, as we already noticed in the experiments devised in Chapter 8.

In Figure 10.10 we compare these probing scores with those extracted from the erroneous sentences. Specifically, we report layer-wise scores obtained for all the linguistic features according to the three macro-classes of errors: Grammar, Orthography and Lexicon. As we can notice, the most noticeable differences are related to the presence of grammatical and orthographic errors, while lexical errors do not seem to affect BERT's ability to correctly encode our probed linguistic features. This is quite expected since a lexical error, differently from an orthographic and especially a grammatical one, is expected to have a minor impact on the overall sentence structure

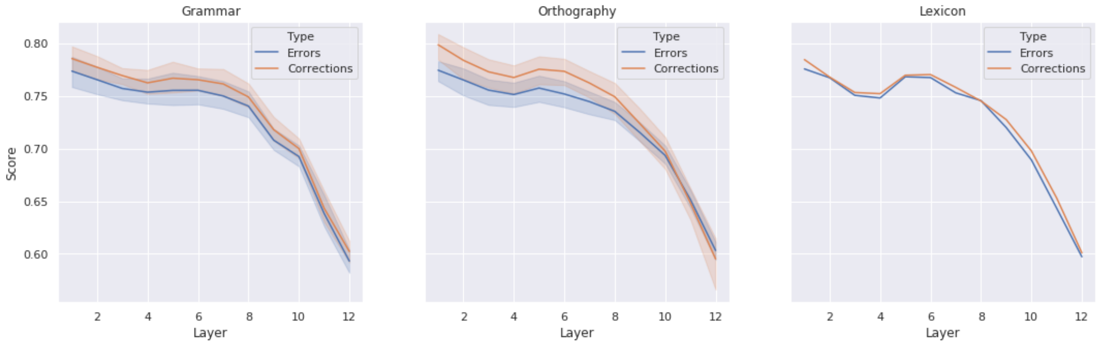


Figure 10.10: *Layer-wise probing scores (ρ correlations) for all linguistic tasks according to the three macro-groups of Grammar, Orthography and Lexicon.*

and, consequently, on the prediction of the average score of our probing features. This is shown, for instance, by the following sentence pair, where the lexical error concerns the use of the adjective *costante* (constant, invariant), which is improperly used by the learner with the meaning of another word, yet still an adjective, i.e. *aderente* (compliant, appropriate).

> Erroneous sentence (311): *Devi cercare di rimanere il più* ***costante*** *possibile alla traccia.* [lit. You should try to be as much **constant** as possible with the prompt topic.]
>
> Corrected sentence: *Devi cercare di rimanere il più* ***aderente*** *possibile alla traccia.* [You should try to be as much **compliant** as possible with the prompt topic.]

Interestingly, as the last layers are approached (layers 10-12), the differences between probing scores obtained with corrected and erroneous sentence representations become much less pronounced. This shows that, regardless of the ability to solve a specific probing task, BERT tends to assimilate these representations possibly becoming more robust to students' errors.

Focusing more specifically our analysis on the impact of standard and non-standard forms on BERT's linguistic competence, we report in Figure 10.11 the average differences between probing scores obtained with *[CLS]* representations of corrected and erroneous sentences for the 11 classes of errors previously defined (see column 1 in Table 10.1, Sec. 10.4) and characterized by the linguistic units involved in the errors. As we can see, the two classes for which in the output layer the impact of non-standard forms is higher are represented by two orthographic errors, i.e. the erroneous misspelling of monosyllables and the erroneous use of the apostrophe. It means that BERT, using the representations extracted from the erroneous sentence, is less able to encode the raw and morpho-syntactic information of the corrected version of the sentence. On the contrary, the erroneous use of vocabulary and of prepositions, pronouns and articles are the classes of errors which impact less on the probing scores. For what concerns lexical errors, our intuition is that the inappropriate use of a single word does not negatively affect the accurate recognition of all the considered linguistic features, possibly with the exception of the correct prediction of word length. For what concerns grammatical errors due to the incorrect use of words belonging to closed lexical categories, i.e. prepositions, pronouns

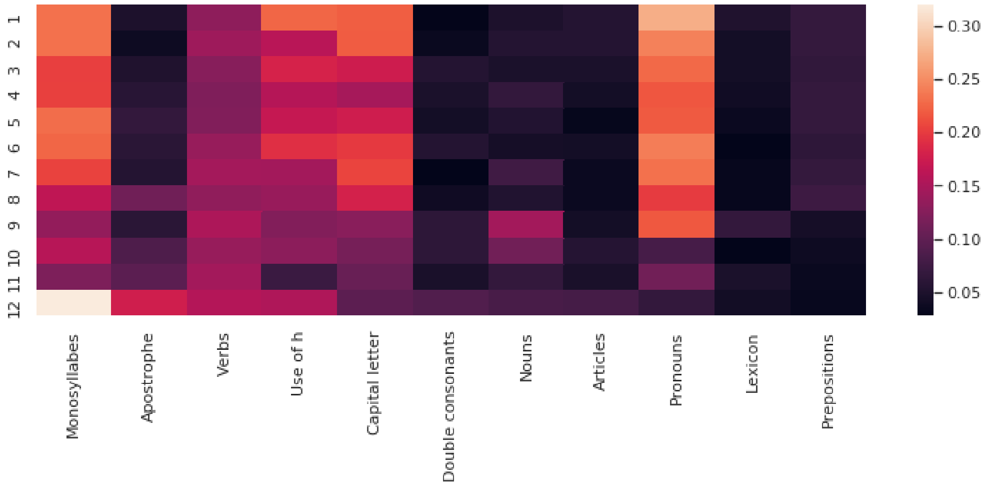


Figure 10.11: *Average differences (absolute values) between probing scores obtained with [CLS] representations of corrected and erroneous sentences for the 11 classes of errors. The 11 classes are ranked in terms of decreasing values of differences between probing scores obtained in the output layer (layer 12).*

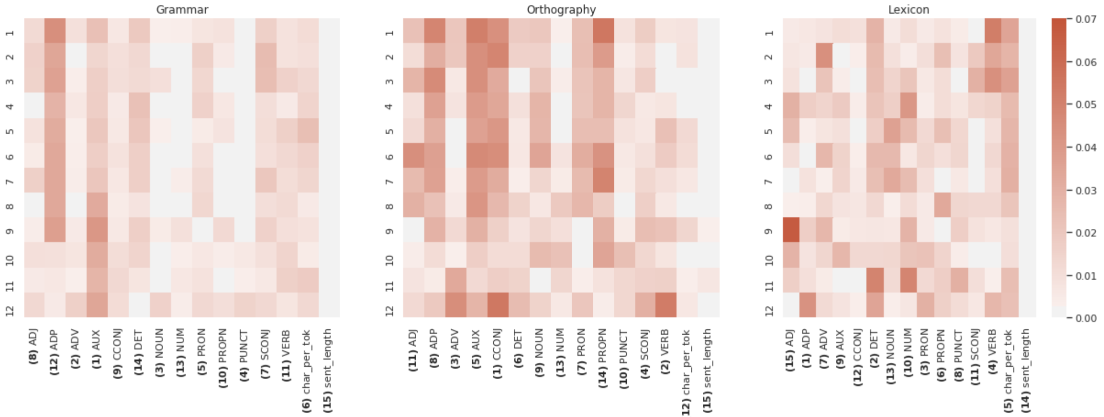


Figure 10.12: *Average differences (absolute values) between probing scores obtained with [CLS] representations of corrected and erroneous sentences for the 15 linguistic features. Scores are reported by averaging the results of the three macro-categories of errors. Bold numbers in parentheses correspond to the ranking of each linguistic feature in terms of decreasing differences computed on the output layer (layer 12).*

and articles, it may be the case that these categories do not have much impact on *[CLS]* representations. However, when the results within internal layers are considered, we noticed an exception among errors involving a closed lexical category. Namely, the incorrect use of pronouns shows a quite peculiar trend, in that the differences between probing scores obtained using BERT's representation extracted from the first layer up to the ninth one are much higher that the differences in the last layers. This may suggest that the erroneous use of a pronoun may distort BERT's representation of the whole

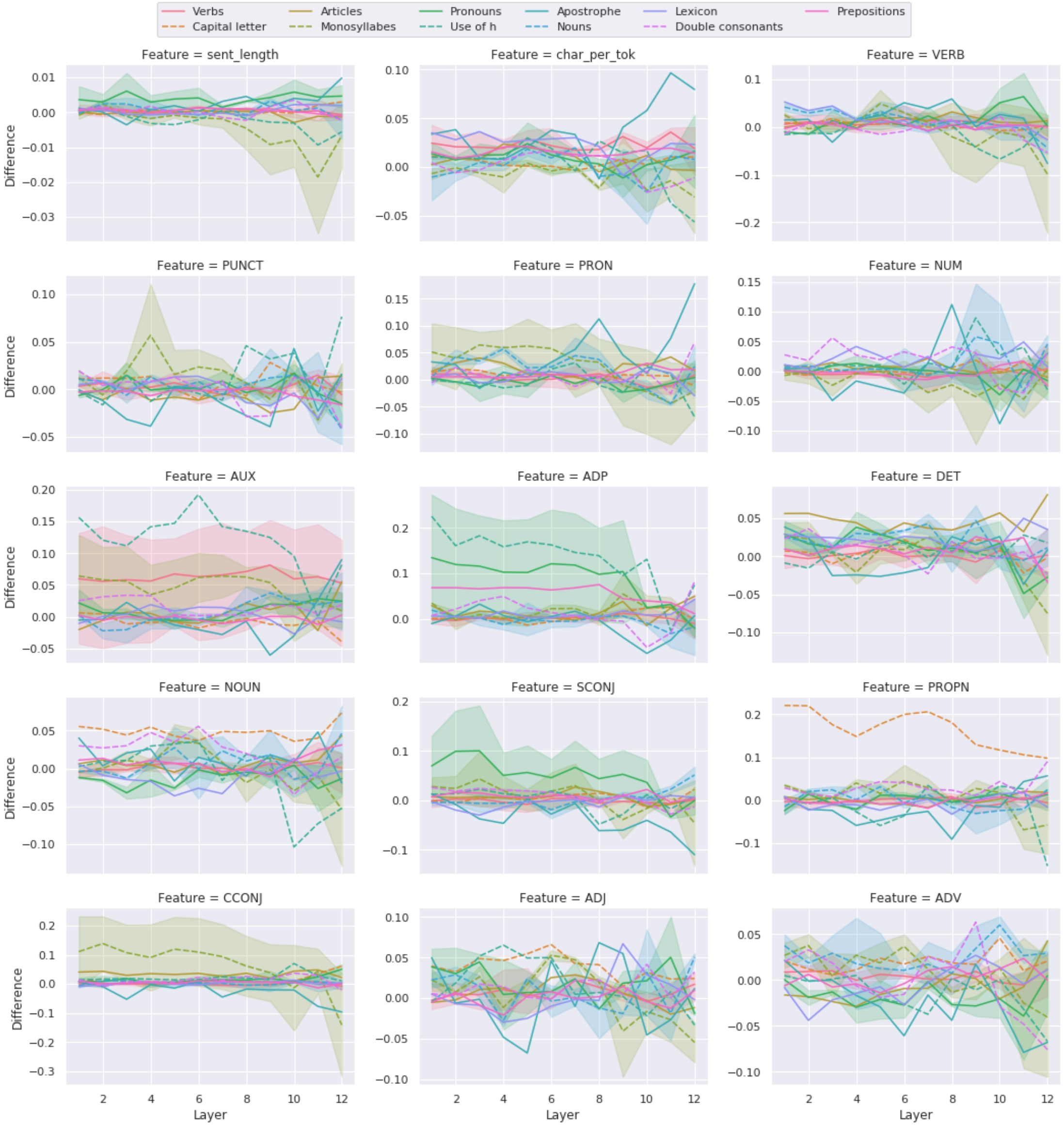


Figure 10.13: *Differences between probing scores obtained with sentence representations with and without specific learner errors.*

sentence and the model needs to achieve a deep knowledge of the sentence in order to correctly predict the underlying linguistic information.

Figure 10.12 reports the results of a complementary investigation. It shows the average differences between probing scores obtained with *[CLS]* representations extracted from the corrected and erroneous sentences for the 15 linguistic features. As we can observe by inspecting the different rankings of features across the three macro-categories of grammatical, orthographic and lexical errors, *sent_length* represents the only probing feature that does not vary also across layers. It means that BERT keeps this competence regardless of the presence of a non-standard form in the sentence. On the contrary, the impact of errors on all the other linguistic features depends on the macro-category taken into account. As a general remark, it appears that orthographic errors represent the macro-category that mostly influences BERT's linguistic abilities. This is in line with what we observed in Section 10.7.2, where we showed how the cosine distance between the representations extracted from the erroneous and corrected sentences is higher as far as the orthographic errors are concerned, both at sentence and token levels, especially in the last layers (see Figure 10.8). In fact, the last layer of the *Orthography* heatmap is characterised by higher average differences between probing scores which correspond to an higher number of darker cells in the heatmap. In particular, coordinating conjunctions (*CCONJ*), verbs (*VERB*) and adverbs (*ADV*) are the linguistic features whose recognition is mostly affected by orthographic errors. On the contrary, grammatical errors mostly affect the correct identification of auxiliary verbs (*AUX*), adverbs and nouns. For what concerns the non-standard use of lexicon, BERT's linguistic abilities decreases in particular in the identification of adpositions (*ADP*), determiners (*DET*) and pronouns (*PRON*).

In Figure 10.13 we report for each linguistic probing task the layer-wise differences between probing scores obtained with the representations extracted from corrected and erroneous sentences for the 11 classes of error. As we can observe, differences between probing scores are mostly positive, thus suggesting that the identification of the linguistic features starting from the representations extracted from corrected sentences is more accurate than those extracted from the erroneous sentences. Nevertheless, these differences tend to become less pronounced as the output layer is approached. This seems to confirm the fact that, regardless of the error typology taken into consideration, BERT becomes progressively more robust to cope with non-standard linguistic forms. For instance, although the incorrect use of *h* or of verbs (either in terms of tense, mood or subject-verb agreement) compromises BERT's ability to correctly predict the distribution of auxiliaries (*AUX*) and prepositions (*ADP*) in the early layers, the presence of these specific error typologies becomes less important in the output layers of the model. For example, in the following sentence:

> Erroneous sentence (222): *Carlo era bravo* ***ha*** *nascondersi infatti nessuno riusciva a vederlo.* [lit. Carlo was good **has** hiding, in fact no one could see him.]
>
> Corrected sentence: *Carlo era bravo* ***a*** *nascondersi infatti nessuno riusciva a vederlo.* [Carlo was good **at** hiding, in fact no one could see him.]

we observe that the use of the *h* before the *a* preposition (lit. *has*) led BERT to incorrectly identify the token as an auxiliary verb (*have*). Nevertheless, this behaviour is more pronounced in the early layers compared to the output ones (*err_diff*[5] is 0.063, 0.048 and 0.006 in layers 1, 6 and 12 respectively.) As expected, the erroneous use of *h* has a major effect also on the correct prediction of the average distribution of prepositions (*ADP*). In fact, we can note that it is among the top error for which there is a great difference, in terms of *err_diff*, between the representation extracted from the

[5] *err_diff* has been computed as *corrected_diff - erroneous_diff*, where *corrected_diff* and *erroneous_diff* correspond to the difference between the feature value (e.g. distribution of auxiliaries) and the value predicted by the probing model tested on the corrected and erroneous sentences respectively.

Features	Class of errors	Slope	r-value
CCONJ	Monosyllabes	-0.017	-0.80
ADP	Use of h	-0.014	-0.81
ADP	Pronouns	-0.011	-0.82
PROPN	Capital letter	-0.010	-0.82
PRON	Monosyllabes	-0.009	-0.85
SCONJ	Pronouns	-0.008	-0.81
NOUN	Use of h	-0.008	-0.65
AUX	Use of h	-0.008	-0.58
SCONJ	Apostrophe	-0.007	-0.77
VERB	Monosyllabes	-0.007	-0.69
PROPN	Monosyllabes	-0.006	-0.63
ADV	Apostrophe	-0.006	-0.66
CCONJ	Apostrophe	-0.006	-0.64
SCONJ	Monosyllabes	-0.006	-0.74
ADJ	Monosyllabes	-0.006	-0.60
VERB	Nouns	-0.006	-0.87
VERB	Lexicon	-0.006	-0.87
ADV	Monosyllabes	-0.005	-0.78
DET	Monosyllabes	-0.005	-0.61
DET	Pronouns	-0.005	-0.72

Table 10.7: *Ranking of the top 20 probing features and corresponding class of error according to decreasing slope of the regression lines. Correlation coefficients are also reported (r-value).*

corrected and erroneous sentence. Once again, this effect becomes less pronounced in the output layer.

Other interesting examples concern the incorrect use of the apostrophe and of the capital letter, which have an effect, especially in the early layers, on the prediction of the average distribution of pronouns (*PRON*) and proper nouns (*PROPN*), respectively. For what concerns the former, the following pair shows how a missing apostrophe, which is required between the elided form of the clitic pronoun *ci* and the imperfect tense of the verb *essere* (to be), affects BERT's ability to predict the average distribution of pronouns in the sentence.

Erroneous sentence (241): *Non **cerano** più i telefoni, non **cerano** gli orologi che chiamavano e facevano gli ologrammi.* [lit. **[They don't] wax** anymore phones, **[they don't] wax** anymore watches calling and they were making holograms.]

Corrected sentence: *Non **c'erano** più i telefoni, non **c'erano** gli orologi che chiamavano e facevano gli ologrammi.*[There **were** no more phones, there **were** no watches calling and they were making holograms.]

For what concerns the latter, the incorrect use of the capital letter plays an important role in the prediction of proper nouns' distribution (*PROPN*), as in:

Erroneous sentence (251): *La bambina si chiama **sara**.* [lit. The child's name is **sara**.]

Corrected sentence: *La bambina si chiama **Sara**.* [The child's name is **Sara**.]

Focusing more specifically on the variation across layers, we can observe that the differences between probing scores obtained with corrected and erroneous representations tend to decrease for most of the tested probing features. Specifically, of a total of 165 probing classifiers, about 58% show a decrease in the difference between accuracy scores achieved in the first and last layers of BERT, thus again suggesting its robustness against non-standard linguistic forms. Moreover, we notice that there are some cases for which the increase of robustness by BERT is constant across its 12 layers. In order to deeply investigate these trends, we computed the slopes of a linear regression line between BERT layers and the differences between the corrected and erroneous probing scores for each linguistic feature and class of error. Table 10.7 reports top 20 probing features and classes of error BERT becomes more rapidly able to cope with across layers, ordered according to decreasing slope values. In most cases, they correspond to the linguistic competences that, when a specific class of error appears in a sentence, the model tends to lose in the first layers but that it acquires more rapidly and constantly across layers tending toward the linguistic knowledge that it has on the corrected sentences. This is for example the case of the incorrect use of *h* and of pronouns for which the differences between the probing scores obtained from erroneous and corrected representations are higher in the first layers but tend to rapidly decrease, as we already noted in Figure 10.11. A quite odd trend, instead, is represented by the erroneous spelling of monosyllables (they appear 8 times in Table 10.7), for which BERT's linguistic competences are negatively affected in the first layers, then rapidly increase across layers, but in the last layer this class of error goes back to negatively affect the model's competence. Interestingly, we already observed the same trend in Figure 10.8, when we compared the internal BERT's representations in terms of cosine distance.

Despite this trend, there are still some exceptions. A visible one is represented by the trend of the class of errors involving the misuse of verbs. As Figures 10.11 and 10.13 show, the differences concerning this type of errors tend to be quite stable across the 12 layers, thus demonstrating that the linguistic knowledge about verbs has the same impact on the layer-wise representations.

10.9 Discussion

One of the most lively research field in current NLP work is devoted to analyzing and interpreting the underlying mechanisms of deep networks taking insights from interdisciplinary perspectives going from machine learning, psychology, linguistics, and neuroscience. The in-depth linguistic investigation presented in this paper goes in this direction and has pursued this objective by testing the robustness of one of the most prominent NLM, i.e. BERT, against non-standard forms emerging in authentic texts. We deeply inspected BERT's behaviour through distinct interpretation strategies which, all together, have revealed the existence of regularities in how BERT handles linguistic errors despite the inherent 'black box' nature of the model. First of all, we observed that the presence of an error in the sentence has always an effect, which emerges in a classification scenario, as well as when the model's attention heads and internal representations were considered. Nevertheless, we also noticed that not all errors impact in the same way. In particular, BERT appears to be more sensitive to orthographic errors with respect to grammatical ones and, even less, to errors affecting the use of lexicon. This was shown in Section 10.6, where we reported a higher classification performance in discriminating pairs of sentences containing orthographic errors; in Section 10.7.1, where we observed that the variation between average attention values in sentences containing an erroneous and corrected token is primarily due to orthographic errors, and in Section 10.7.2, where we showed that orthographic errors are, in general, those that contribute more to the variation, measured in terms of cosine distances between representations extracted from sentences containing a corrected or an erroneous linguistic form.

By exploiting the internal subdivision of errors for each macro-category available in the CItA corpus, we were able to provide a more accurate picture of the effect of errors of different nature on the model's robustness. For instance, if in general orthographic errors turned out to be the ones with the strongest impact, not all errors of this class play the same role. In particular, our data emphasize the existence of an internal hierarchy, with some errors such as the misspelling of the monosyllabic word *po'* and the incorrect use of the apostrophe occurring in the top-ranked positions. Interestingly, this kind of errors represent an area of the Italian written competence which is quite difficult to master not only for younger learners but also for adult writers [Serianni, 1989].

However, when the model's internal representations are considered, we observed that the degree of variation between representations extracted from correct and incorrect sentences is different according to token-based or sentence-based representations. Namely, the former are more pronounced especially in the earlier layers but they also tend to converge rapidly. This trend may explain also why the best layers for identifying the presence of a learner error are the ones between layers 7 and 9, as showed in the classification results of Section 10.6. This is also in line with results reported by [Yin et al., 2020] in their study. Again, the type of error differently impacts on this process: errors that give rise to a greater distance between the erroneous and corrected representations in the first layers are also the ones for which the model generates a representation converging to the correct standard one in more rapid way. It is the case of errors concerning the use of lexicon but also of some orthographic errors such as the redundant use of *h* and of double consonants and the erroneous use of monosyllabic words. On the contrary, sentence-level representations based on the activation of the *[CLS]* token do not diverge too much, suggesting that the model is able to incorporate the error when generating a representation of the whole sentence.

In spite of the reported small variations between correct and incorrect sentence-level representations, when BERT's linguistic competences are probed using these representations, the impact of errors is visible. Also from this perspective, the model's robustness turns out to be differently affected by the specific type of error occurring in the sentence and by the specific layer from which the sentence-level representation is extracted. Generally speaking, as shown in Figure 10.10, BERT's ability to properly encode a set of properties related to superficial and morpho-syntactic information of a sentence is more affected by the presence of orthographic and grammatical errors, while lexical errors are rather harmless. However, as the output layer is approached, the model becomes progressively more robust to cope with non-standard linguistic forms regardless of the error typology occurring in the sentence.

As for all our previous experiments, also in the evaluation of BERT's linguistic competence, the fine-grained analysis focused on the specific type of error within each macro-category allowed us to find interesting trends. For instance, comparing the results obtained using the output layer representations (Figure 10.11), the misspelling of monosyllables and the erroneous use of apostrophe turned out to be the two classes of orthographic errors for which the impact of non-standard forms is greater. This confirms our expectations that errors that impact more are those that may potentially affect the overall sentence structure, such as the misuse of an apostrophe, which not only alters the spelling of the affected token but may also change the corresponding POS, with a possible propagation of the error to other tokens of the sentence. On the contrary, at grammatical level, we found a distinction between errors affecting tokens belonging to open and closed lexical categories. In particular, the misuse of verb (in terms of a wrong use of tense and number features) has a quite prominent effect that holds across all layers, thus showing that BERT is scarcely robust to cope with this type of error. On the contrary, errors related to the closed lexical categories (i.e. prepositions, pronouns and articles) show a weak impact although with some different behavior. For instance, while errors involving articles and prepositions do not have impact from the first layers, errors concerning pronouns are the only closed class category showing an adverse effect from the first up to (at least) the ninth layer, suggesting that the model needs to achieve a deep knowledge of the sentence in order to mitigate the effect of a wrong pronoun and correctly predict the underlying linguistic information. This behaviour might also explain the reason why while articles and prepositions show high correlation values between attention heads of corrected and erroneous tokens (see Table 10.4), on the contrary for what concerns the erroneous use of relative pronouns we observe very low correlations. Our intuition is that BERT has to attend to other tokens in the sentence in order to cope with the presence of this specific error.

With findings reported in Figures 10.12 and 10.13 we tried to achieve a deeper understanding of the relationship between each type of error and the specific phenomenon of BERT's linguistic competence affected. In particular, we showed that the presence of orthographic and grammatical errors negatively influences the correct prediction of all our probing features. Specifically, the former has an effect on coordinating conjunctions, verbs and adverbs, while the latter on auxiliary verbs, adverbs and nouns. Vocabulary-based errors instead tend to affect only functional POS, such as prepositions, determiners and pronouns. Once again, this overall effect becomes less pronounced as we approach the output layer.

CHAPTER 11

Conclusions

In this thesis, we have investigated different approaches in order to interpret the inner mechanisms of state-of-the-art NLMs and to understand the amount of linguistic knowledge implicitly encoded by Transformer-based models. In particular, we verified whether exploiting a profiling approach to study human linguistic competence and, more specifically, the process of written language evolution could provide important insights about the linguistic knowledge encoded (and used) by these neural models.

In the first part of the thesis, we defined an NLP-based stylometric approach to model the evolution of written language competence in L1 and L2 learners. In particular, relying on a wide set of linguistically motivated features extracted from the students' essays contained in two longitudinal corpora of Italian L1 and Spanish L2 learners, we showed that it is possible to automatically predict the chronological order of two essays written by the same student, especially at more distant temporal spans. Moreover, we have highlighted that our set of linguistic features can be also exploited to investigate the typologies of language phenomena that contribute more to the prediction task and how they change according to different temporal spans. In fact, experiments devised on the *CItA* corpus showed that morpho-syntactic properties, and especially those related to grammatical categories and to the inflectional properties of verbs, play an important role in the classification task as the temporal span between two essays increases. Similarly, features related to the errors made by the students become more important when larger temporal intervals are taken into account. For what concerns instead the experiments devised on the COWS-L2H corpus, we found that the linguistic features that are most important in the predicting task often reflect the explicit instructions that students receive during each course (e.g. features related to verbal morphology for the introductory courses).

Starting from the assumption that our profiling approach provided us with important insights about the process of evolution of written language competence and, more importantly, about the most relevant morphosyntactic and syntactic properties at different time intervals, in the second part of the thesis we decided to apply a similar methodology in order to study the implicit linguistic competence of recent Transformer-based models. We first proposed a suite of probing tasks based on our set of linguistic features in order to investigate the amount of linguistic knowledge stored in several pre-trained NLMs, mostly based on the BERT architecture. Results showed that English and Italian pre-trained models are able to encode a wide amount of linguistic proprieties across their layers, although this implicit competence tends to decrease as the output layer (i.e. the layer that is more close to the pre-training objective) is approached. Focusing exclusively on the BERT architecture, we showed that, differently from a non-contextual model (word2vec), BERT sentence-level representations perform better at encoding features related to the raw and syntactic structure of a sentence and that this information is also preserved in the embeddings of individual words. Moreover, fine-tuning BERT on the Native Language Identification task we found that, despite the model tends to lose its precision in encoding our set of features, the ability to solve the downstream task is related to its ability in storing linguistic knowledge.

Then, we further investigated the linguistic competence of pre-trained Transformer models proposing two complementary studies aimed at understanding the relationship between linguistic generalization abilities and perplexity scores. In our first work, the relationship between BERT and GPT-2 PPLs and our set of linguistic features was comparatively assessed. Specifically, training a linear regression model to predict PPL scores, we showed that our features are able to model aspects involved in NLM's perplexity, and that this is true especially for GPT-2. Moreover, we found that the properties that are more involved in the PPL of the two models correspond to the lexical density, the presence of pronouns and verbs. In the follow-up study, we examined whether PPL is affected by the same linguistic phenomena used to automatically assess sentence readability and if there is a correlation between the two metrics. Our findings suggested that this correlation is actually quite weak and the two metrics are affected by different linguistic phenomena.

Moving instead on the framework of studies related to the assessment of NLMs ability on targeted diagnostic tests, we built a new evaluation resource for Italian aimed at assessing the role of textual connectives in the comprehension of the meaning of a sentence. The resource was arranged in two sections (acceptability assessment and cloze test), each one corresponding to a distinct challenge task conceived to test how subtle modifications involving connectives in real usage sentences influence the perceived acceptability of the sentence by native speakers NLMs. Preliminary findings showed that BERT and GPT-2 often are capable of distinguishing between acceptable and unacceptable sentences, thus suggesting their ability to understand sentence meaning within their internal mechanisms.

In the last part of the thesis, we introduced a study to test the robustness of a pre-trained Italian BERT against non-standard forms emerging in authentic texts. In particular, we proposed an extensive analysis on the behaviour of BERT when dealing with the learner errors derived from the *CItA* corpus. For this purpose, we relied on several interpretation techniques, ranging from the definition of probing tasks for inferring the linguistic competence of the model to the analysis of word- and sentence-level representations and attention heads. The results showed that the model is more sensitive to orthographic errors and especially to the erroneous use of the monosyllabic word *po'* and of the apostrophe. On the other hand, grammatical and lexical errors seem to have a smaller impact on BERT representations and attention heads. Interestingly, we observed that, regardless of the methodology devised, the presence of a learner error starts to play a important role mainly from the intermediate layers of the model. Probing the linguistic knowledge of BERT relying on 15 features derived from the raw structure and the distribution of POS tags in a sentence, we noticed that, once again, orthographic errors are generally those for which the impact of error is most pronounced when solving the probing tasks, especially for what concerns properties related to the distribution of coordinating conjunctions, verbs and adverbs. Nevertheless, as the output layer is approached, the impact of learner errors on probing performances becomes less pronounced, thus suggesting that the model is getting progressively more robust to cope with non-standard linguistic forms regardless of the error typology taken into account.

11.1 Future Work

There are several improvements and advancements that can be introduced to extend this research and that we leave as future work.

Since the approach proposed in Part II could lead the way to further comparative studies, it would be interesting, along the lines of studies such as [Chi et al., 2020], to deepen the analysis of the linguistic knowledge encoded in multilingual models and in Transformer models pre-trained on different languages. Furthermore, it could be worth exploring whether and how the linguistic phenomena that most characterise the evolution of different L1 and L2 learners' writing skills are reflected in the implicit abilities of these models, especially during a pre-training phase.

One of the fundamental aspects regarding the interpretability of NLMs that we have left out concerns the way in which the linguistic knowledge is implicitly learned during the training process. For this reason, in future work we also plan to study how this knowledge arise during the pre-training phase and how it changes when dealing with different (and more linguistically motivated) training objectives. A possible outcome of this study would be the investigation of new strategies to maximize the linguistic competence of a NLM. In fact, despite it has been demonstrated that the introduction of linguistic information during pre-training enhances the performance of these models [Zhou et al., 2020, Bai et al., 2021], this improvement has not yet been investigated in the light of shifts of linguistic competence during the training process.

The experiments devised in the last part of the thesis can be a starting point for further studies on the impact of noisy data on NLMs linguistic competence and downstream performances. In fact, as it was shown by previous studies (e.g. [Sun et al., 2020a] and [Kumar et al., 2020]), noisy input data adversely affect BERT's performance in real-word scenarios, such as sentiment analysis, question answering and sentiment similarity. Although our study has been conducted on a learner corpus, some of the most representative classes of errors and non-standard forms that it contains are persistent at higher level of education and widely spread on informal writing, such as social media texts. In light of this, we hope that our findings could support researchers working on improving the robustness of NLMs in multiple real-word applications by also providing reliable explanations of the model's behaviour at prediction time. For instance, they could be used to evaluate which is the best model for a specific downstream task or to define new strategies (e.g. selecting input data appropriately during the pre-training phase) to develop more robust systems by strengthening their implicit linguistic competence.

Bibliography

[Al-Rfou' et al., 2013] Al-Rfou', R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

[Alain and Bengio, 2016] Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *ArXiv Pre-print*, 1610.01644.

[Antonelli, 2012] Antonelli, G. (2012). L'e-taliano: una nuova realtà tra le varietà linguistiche dell'italiano? *E. Garavelli, E. SuomelaHarma (curated by), Dal manoscritto al web: canali e modalità di trasmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua. Atti del XII Congresso SILFI, Società Internazionale di Linguistica e Filologia Italiana*, II:549–551.

[Argamon et al., 2003] Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text & Talk*, 23(3):321–346.

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Bai et al., 2021] Bai, J., Wang, Y., Chen, Y., Yang, Y., Bai, J., Yu, J., and Tong, Y. (2021). Syntax-BERT: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3011–3020, Online. Association for Computational Linguistics.

[Ballier et al., 2019] Ballier, N., Gaillat, T., Simpkin, A., Stearns, B., Bouyé, M., and Zarrouk, M. (2019). A supervised learning model for the automatic assessment of language levels based on learner errors. In *European Conference on Technology Enhanced Learning*, pages 308–320. Springer.

[Barbagli, 2016] Barbagli, A. (2016). *Quanto e come si impara a scrivere nel corso del primo biennio della scuola secondaria di primo grado*. Nuova Cultura.

[Barbagli et al., 2016] Barbagli, A., Lucisano, P., Dell'Orletta, F., Montemagni, S., and Venturi, G. (2016). CItA: an L1 Italian learners corpus to study the development of writing competence. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 88–95, Portorož, Slovenia. European Language Resources Association (ELRA).

[Baroni et al., 2009] Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

[Basile et al., 2018] Basile, V., Lai, M., and Sanguinetti, M. (2018). Long-term social media data collection at the university of turin. In *Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 1–6. CEUR-WS.

Bibliography

[Belinkov and Bisk, 2018] Belinkov, Y. and Bisk, Y. (2018). Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

[Belinkov and Glass, 2019] Belinkov, Y. and Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

[Belinkov et al., 2017] Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., and Glass, J. (2017). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10.

[Bell et al., 2019] Bell, S., Yannakoudakis, H., and Rei, M. (2019). Context is key: Grammatical error detection with contextual word representations. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 103–115, Florence, Italy. Association for Computational Linguistics.

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155.

[Bentivogli et al., 2016] Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas. Association for Computational Linguistics.

[Benzahra and Yvon, 2019] Benzahra, M. and Yvon, F. (2019). Measuring text readability with machine comprehension: a pilot study. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 412–422, Florence, Italy. Association for Computational Linguistics.

[Bergsma et al., 2012] Bergsma, S., Post, M., and Yarowsky, D. (2012). Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337.

[Berman, 2004] Berman, R. A. (2004). *Between emergence and mastery. The long development route of language acquisition.* Trends in language acquisition research vol. 3. Benjamins, Amsterdam Philadelphia.

[Berruto, 1987] Berruto, G. (1987). *Sociolinguistica dell'italiano contemporaneo*, volume 33. Carocci.

[Berzak et al., 2016] Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., and Katz, B. (2016). Universal dependencies for learner english.

[Bestgen and Granger, 2018] Bestgen, Y. and Granger, S. (2018). Tracking l2 writers' phraseological development using collgrams: Evidence from a longitudinal efl corpus. In *Corpora and lexis*, pages 277–301. Brill.

[Biber, 1993] Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2):219–242.

[Blanchard et al., 2013] Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. (2013). Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

[Blevins et al., 2018] Blevins, T., Levy, O., and Zettlemoyer, L. (2018). Deep rnns encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19.

[Bock and Miller, 1991] Bock, K. and Miller, C. (1991). Broken agreement. *Cognitive Psychology*, 23 (1):45–93.

[Bojanowski et al., 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[Bosco et al., 2013] Bosco, C., Montemagni, S., and Simi, M. (2013). Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*.

[Boyd et al., 2014] Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Stindlová, B., and Vettori, C. (2014). The merlin corpus: Learner language and the cefr. In *LREC*, pages 1281–1288.

[Brooke and Hirst, 2012] Brooke, J. and Hirst, G. (2012). Measuring interlanguage: Native language identification with l1-influence metrics. In *LREC*, pages 779–784.

[Brunato et al., 2020] Brunato, D., Cimino, A., Dell'Orletta, F., Montemagni, S., and Venturi, G. (2020). Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France. European Language Resources Association (ELRA).

[Brunato et al., 2016] Brunato, D., Cimino, A., Dell'Orletta, F., and Venturi, G. (2016). PaCCSS-IT: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361, Austin, Texas. Association for Computational Linguistics.

[Brunato et al., 2018] Brunato, D., De Mattei, L., Dell'Orletta, F., Iavarone, B., and Venturi, G. (2018). Is this sentence difficult? do you agree? In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*.

[Brunato et al., 2015] Brunato, D., Dell'Orletta, F., Venturi, G., and Montemagni, S. (2015). Design and annotation of the first italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41.

[Brunner et al., 2020] Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., and Wattenhofer, R. (2020). On identifiability in transformers. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

[Brysbaert et al., 2011] Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., and Böhl, A. (2011). The word frequency effect. *Experimental psychology*.

[Bulté and Housen, 2012] Bulté, B. and Housen, A. (2012). Defining and operationalising L2 complexity. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, pages 23–46.

[Bulté and Housen, 2014] Bulté, B. and Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of second language writing*, 26:42–65.

[Caines and Buttery, 2020] Caines, A. and Buttery, P. (2020). REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5614–5623, Marseille, France. European Language Resources Association.

[Chalkidis et al., 2020] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

[Chang and Lin, 2001] Chang, C.-C. and Lin, C.-J. (2001). LIBSVM: a library for support vector machines.

[Chi et al., 2020] Chi, E. A., Hewitt, J., and Manning, C. D. (2020). Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

[Chipere et al., 2001] Chipere, N., Malvern, D. D., Richards, B. J., and Durán, P. (2001). Using a corpus of school children's writing to investigate the development of lexical diversity. In *P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (Eds.), Proceedings of the Corpus Linguistics 2001 Conference (pp. 126–133), Lancaster, UK: University Centre for Computer Corpus Research on Language.*

[Cho et al., 2014] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

[Choe et al., 2019] Choe, Y. J., Ham, J., Park, K., and Yoon, Y. (2019). A neural grammatical error correction system built on better pre-training and sequential transfer learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–227, Florence, Italy. Association for Computational Linguistics.

[Chrupała and Alishahi, 2019] Chrupała, G. and Alishahi, A. (2019). Correlating neural and symbolic representations of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962, Florence, Italy. Association for Computational Linguistics.

[Cignarella et al., 2019] Cignarella, A. T., Bosco, C., and Rosso, P. (2019). Presenting TWITTIRÒ-UD: An italian twitter treebank in universal dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*.

[Cignetti, 2011] Cignetti, L. (2011). Paragrafematici, segni. *In: R. Simone (curated by), Enciclopedia dell'italiano, Roma: Istituto dell'Enciclopedia Italiana G. Treccani.*, II:1033–1034.

Bibliography

[Cimino et al., 2018] Cimino, A., Dell'Orletta, F., Brunato, D., and Venturi, G. (2018). Sentences and Documents in Native Language Identification. In *Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-it)*, pages 1–6, Turin.

[Clark et al., 2019] Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

[Clark et al., 2020] Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

[Colombo, 2010] Colombo, A. (2010). *A me mi. Dubbi, errori, correzioni nell'italiano scritto: Dubbi, errori, correzioni nell'italiano scritto*. FrancoAngeli.

[Conneau et al., 2018] Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

[Covington et al., 2006] Covington, M. A., He, C., Brown, C. adn Naci, L., and Brown. (2006). How complex is that sentence? a proposed revision of the rosenberg and abbeduto d-level scale. *CASPR Research Report 2006-01, Athens, GA: The University of Georgia, Artificial Intelligence Center*.

[Crossley et al., 2011a] Crossley, A. S., Weston, J., McLain Sullivan, S., and McNamara, D. S. (2011a). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, Written Communication, vol. 28, no. 3, pp. 282–311*.

[Crossley, 2020] Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*.

[Crossley and McNamara, 2012] Crossley, S. A. and McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115–135.

[Crossley et al., 2011b] Crossley, S. A., Salsbury, T., McNamara, D. S., and Jarvis, S. (2011b). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, 28(4):561–580.

[Daelemans, 2013] Daelemans, W. (2013). Explanation in computational stylometry. *Computational Linguistics and Intelligent Text Processing*, 7817:451–462.

[Dahlmeier et al., 2013] Dahlmeier, D., Ng, H. T., and 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*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

[Dai et al., 2019] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

[Danks and London, 2017] Danks, D. and London, A. J. (2017). Regulating autonomous systems: Beyond standards. *IEEE Intelligent Systems*, 32(1):88–91.

[Davidson et al., 2019] Davidson, S., Yamada, A., Carando, A., Sagae, K., and Sánchez Gutiérrez, C. (2019). Word use and lexical diversity in second language learners and heritage speakers of spanish: A corpus study. In *American Association for Applied Linguistics*.

[Davidson et al., 2020] Davidson, S., Yamada, A., Fernandez Mira, P., Carando, A., Sanchez Gutierrez, C. H., and Sagae, K. (2020). Developing nlp tools with a new corpus of learner spanish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7240–7245, Marseille, France. European Language Resources Association.

[De Mattei et al., 2020] De Mattei, L., Cafagna, M., Dell'Orletta, F., Nissim, M., and Guerini, M. (2020). Geppetto carves italian into a language model. In *Italian Conference on Computational Linguistics 2020*. CEUR-WS. org.

[De Mauro, 1983] De Mauro, T. (1983). Per una nuova alfabetizzazione. *Gensini S., Vedovelli M.(edited by). Teoria e pratica del glotto-kit. Una carta d'identità per l'educazione linguistica.*

[De Mauro and Chiari, 2016] De Mauro, T. and Chiari, I. (2016). Il nuovo vocabolario di base della lingua italiana. *Internazionale.[28/11/2020]. https://www. internazionale. it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana.*

[de Vries et al., 2020] de Vries, W., van Cranenburgh, A., and Nissim, M. (2020). What's so special about BERT's layers? a closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, Online. Association for Computational Linguistics.

[Dell'Orletta et al., 2011a] Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011a). READ-IT: assessing readability of Italian texts with a view to text simplification. In *Proceedings of Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, UK.

[Dell'Orletta et al., 2011b] Dell'Orletta, F., Venturi, G., and Montemagni, S. (2011b). Ulisse: an unsupervised algorithm for detecting reliable dependency parses. In *CoNLL*.

[Dell'Orletta et al., 2014] Dell'Orletta, F., Wieling, M., Venturi, G., Cimino, A., and Montemagni, S. (2014). Assessing the readability of sentences: Which corpora and features? In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173, Baltimore, Maryland. Association for Computational Linguistics.

[Delmonte et al., 2007] Delmonte, R., Bristot, A., and Tonelli, S. (2007). VIT - Venice Italian Treebank: Syntactic and quantitative features. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Díaz-Villanueva et al., 2010] Díaz-Villanueva, W., Ferri, F. J., and Cerverón, V. (2010). Learning improved feature rankings through decremental input pruning for support vector based drug activity prediction. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 653–661. Springer.

[Dyer et al., 2016] Dyer, C., Kuncoro, A., Ballesteros, M., and Smith, N. A. (2016). Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209, San Diego, California. Association for Computational Linguistics.

[Ettinger, 2020] Ettinger, A. (2020). What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

[Farzindar and Inkpen, 2015] Farzindar, A. and Inkpen, D. (2015). *Natural Language Processing for Social Media*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool.

[Fayyaz et al., 2021] Fayyaz, M., Aghazadeh, E., Modarressi, A., Mohebbi, H., and Pilehvar, M. T. (2021). Not all models localize linguistic knowledge in the same place: A layer-wise probing on bertoids' representations. In *Proceedings of the 2021 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

[Franck et al., 2002] Franck, J., Vigliocco, G., and Nicol, J. (2002). Subject-verb agreement errors in french and english: the role of syntactic hierarchy. *Language and Cognitive Processes*, 17:371–404.

[Geertzen et al., 2013] Geertzen, J., Alexopoulou, T., Korhonen, A., et al. (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. Somerville, MA: Cascadilla Proceedings Project*, pages 240–254. Citeseer.

[Gibson et al., 2000] Gibson, E. et al. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

[Gildea, 2001] Gildea, D. (2001). Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

[Goldberg, 2019] Goldberg, Y. (2019). Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Bibliography

[Graesser and McNamara, 2011] Graesser, A. C. and McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in cognitive science*, 3(2):371–398.

[Granger, 2003] Granger, S. (2003). Error-tagged learner corpora and call: A promising synergy. *CALICO journal*, pages 465–480.

[Grundkiewicz et al., 2019] Grundkiewicz, R., Junczys-Dowmunt, M., and Heafield, K. (2019). Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, Florence, Italy. Association for Computational Linguistics.

[Guidotti et al., 2018] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93.

[Gulordava et al., 2018] Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

[Hall Maudslay et al., 2020] Hall Maudslay, R., Valvoda, J., Pimentel, T., Williams, A., and Cotterell, R. (2020). A tale of a probe and a parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7389–7395, Online. Association for Computational Linguistics.

[Hancke and Meurers, 2013] Hancke, J. and Meurers, D. (2013). Exploring CEFR classification for German based on rich linguistic modeling. In *Proceedings of the Learner Corpus Research (LCR) conference*.

[Hewitt and Liang, 2019] Hewitt, J. and Liang, P. (2019). Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

[Hewitt and Manning, 2019] Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. 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)*, pages 4129–4138.

[Heydari and Bagheri, 2012] Heydari, P. and Bagheri, M. (2012). Error analysis: Sources of l2 learners' errors. *Theory and Practice in Language Studies*, 2:1583–1589.

[Htut et al., 2019] Htut, P. M., Phang, J., Bordia, S., and Bowman, S. R. (2019). Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

[Hu et al., 2020] Hu, J., Gauthier, J., Qian, P., Wilcox, E., and Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

[Jain and Wallace, 2019] Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. 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)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

[Jawahar et al., 2019] Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

[Jolliffe and Cadima, 2016] Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

[Jozefowicz et al., 2016] Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

[Kaneko and Komachi, 2019] Kaneko, M. and Komachi, M. (2019). Multi-head multi-layer attention to deep language representations for grammatical error detection. *Computación y Sistemas*, 23(3).

[Kaneko et al., 2020] Kaneko, M., Mita, M., Kiyono, S., Suzuki, J., and Inui, K. (2020). Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4248–4254, Online. Association for Computational Linguistics.

[Kann et al., 2019] Kann, K., Warstadt, A., Williams, A., and Bowman, S. R. (2019). Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

[Karmiloff-Smith, 1986] Karmiloff-Smith, A. (1986). Some fundamental aspects of language development after age 5. *Language Acquisition: Studies in First Language Development. Cambridge: Cambridge University Press*.

[Kerz et al., 2020] Kerz, E., Qiao, Y., Wiechmann, D., and Ströbel, M. (2020). Becoming linguistically mature: Modeling English and German children's writing development across school grades. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 65–74, Seattle, WA, USA → Online. Association for Computational Linguistics.

[Kerz et al., 2021] Kerz, E., Wiechmann, D., Qiao, Y., Tseng, E., and Ströbel, M. (2021). Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–209, Online. Association for Computational Linguistics.

[Kim et al., 2019] Kim, N., Patel, R., Poliak, A., Wang, A., Xia, P., McCoy, R. T., Tenney, I., Ross, A., Linzen, T., Durme, B. V., Bowman, S. R., and Pavlick, E. (2019). Probing what different nlp tasks teach machines about function word comprehension. In **SEMEVAL*.

[Kingston, 2016] Kingston, J. K. (2016). Artificial intelligence and legal liability. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 269–279. Springer.

[Kiperwasser and Goldberg, 2016] Kiperwasser, E. and Goldberg, Y. (2016). Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

[Kobayashi et al., 2020] Kobayashi, G., Kuribayashi, T., Yokoi, S., and Inui, K. (2020). Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

[Kobayashi et al., 2021] Kobayashi, G., Kuribayashi, T., Yokoi, S., and Inui, K. (2021). Incorporating residual and normalization layers into analysis of masked language models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*. Association for Computational Linguistics.

[Kovaleva et al., 2019] Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (2019). Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

[Kriegeskorte et al., 2008] Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

[Kroll et al., 2016] Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2016). Accountable algorithms. *U. Pa. L. Rev.*, 165:633.

[Kumar et al., 2020] Kumar, A., Makhija, P., and Gupta, A. (2020). Noisy text data: Achilles' heel of BERT. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 16–21, Online. Association for Computational Linguistics.

[Kuncoro et al., 2019] Kuncoro, A., Dyer, C., Rimell, L., Clark, S., and Blunsom, P. (2019). Scalable syntax-aware language models using knowledge distillation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3472–3484, Florence, Italy. Association for Computational Linguistics.

[Kyle and Crossley, 2015] Kyle, K. and Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49(4):757–786.

[Lavalley et al., 2015] Lavalley, R., Berkling, K., and Stüker, S. (2015). Preparing children's writing database for automated processing. In *LTLT@ SLaTE*, pages 9–15.

[Lei and Wen, 2020] Lei, L. and Wen, J. (2020). Is dependency distance experiencing a process of minimization? a diachronic study based on the state of the union addresses. *Lingua*, 239:102762.

[Li et al., 2021] Li, B., Zhu, Z., Thomas, G., Xu, Y., and Rudzicz, F. (2021). How is bert surprised? layerwise detection of linguistic anomalies. *arXiv preprint arXiv:2105.07452*.

Bibliography

[Li et al., 2019] Li, F., Jin, Y., Liu, W., Rawat, B. P. S., Cai, P., and Yu, H. (2019). Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: an empirical study. *JMIR medical informatics*, 7(3):e14830.

[Li et al., 2020] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. (2020). A survey on text classification: From shallow to deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11).

[Lin et al., 2021] Lin, T., Wang, Y., Liu, X., and Qiu, X. (2021). A survey of transformers. *arXiv preprint arXiv:2106.04554*.

[Lin et al., 2019] Lin, Y., Tan, Y. C., and Frank, R. (2019). Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

[Linzen et al., 2016] Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

[Liu et al., 2019a] Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. (2019a). Linguistic knowledge and transferability of contextual representations. 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)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

[Liu et al., 2019b] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[Liu et al., 2020] Liu, Z., Lin, Y., and Sun, M. (2020). *Representation learning for natural language processing*. Springer Nature.

[Lu, 2009] Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. In *International Journal of Corpus Linguistics, 14(1):3–28*.

[Lu, 2011] 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.

[Lubetich and Sagae, 2014] Lubetich, S. and Sagae, K. (2014). Data-driven measurement of child language development with simple syntactic templates. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2151–2160.

[Lucisano, 1984] Lucisano, P. (1984). L'indagine iea sulla produzione scritta. *Ricerca educativa*, 5:41–61.

[Luhn, 1957] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317.

[MacWhinney, 2000] MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* J: Lawrence Erlbaum Associates.

[Malmasi et al., 2017] Malmasi, S., Evanini, K., Cahill, A., Tetreault, J., Pugh, R., Hamill, C., Napolitano, D., and Qian, Y. (2017). A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75, Copenhagen, Denmark. Association for Computational Linguistics.

[Malykh, 2019] Malykh, V. (2019). Robust to noise models in natural language processing tasks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 10–16, Florence, Italy. Association for Computational Linguistics.

[Manevitz and Yousef, 2001] Manevitz, L. M. and Yousef, M. (2001). One-class svms for document classification. *Journal of machine Learning research*, 2(Dec):139–154.

[Marconi, 1994] Marconi, L. (1994). *Lessico elementare: Dati statistici sull'italiano scritto e letto dai bambini delle elementari*. Zanichelli.

[Martinc et al., 2021] Martinc, M., Pollak, S., and Robnik-Šikonja, M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

[Marvin and Linzen, 2018a] Marvin, R. and Linzen, T. (2018a). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

[Marvin and Linzen, 2018b] Marvin, R. and Linzen, T. (2018b). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

[McCann et al., 2017] McCann, B., Bradbury, J., Xiong, C., and Socher, R. (2017). Learned in translation: Contextualized word vectors. *Advances in Neural Information Processing Systems*, 30.

[McNamara et al., 2010] McNamara, D. S., Crossley, S. A., and McCarthy, P. M. (2010). Linguistic features of writing quality. *Written communication*, 27(1):57–86.

[McNamara et al., 2015] McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., and Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59.

[Melamud et al., 2016] Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

[Miaschi et al., 2020a] Miaschi, A., Brunato, D., Dell'Orletta, F., and Venturi, G. (2020a). Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[Miaschi et al., 2020b] Miaschi, A., Davidson, S., Brunato, D., Dell'Orletta, F., Sagae, K., Sanchez-Gutierrez, C. H., and Venturi, G. (2020b). Tracking the evolution of written language competence in l2 spanish learners. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

[Mikolov et al., 2010] Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Misra et al., 2019] Misra, K., Devarapalli, H., and Rayz, J. T. (2019). Measuring the influence of l1 on learner english errors in content words within word embedding models. In *Proceedings of the 7th International Conference on Cognitive Modeling*.

[Namazifar et al., 2021] Namazifar, M., Tur, G., and Hakkani-Tür, D. (2021). Warped language models for noise robust language understanding. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 981–988. IEEE.

[Neelakantan et al., 2014] Neelakantan, A., Shankar, J., Passos, A., and McCallum, A. (2014). Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar. Association for Computational Linguistics.

[Ng et al., 2013] Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., and Tetreault, J. (2013). The conll-2013 shared task on grammatical error correction. *CoNLL-2013*, page 1.

[Nguyen et al., 2016] Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational Sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

[Nivre et al., 2016] Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

[Norris and Ortega, 2009] Norris, J. and Ortega, L. (2009). Measurement for understanding: An organic approach to investigating complexity, accuracy, and fluency in SLA. *Applied Linguistics*, 30(4):555–578.

[Ortega, 2003] 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.

[Otterbacher, 2010] Otterbacher, J. (2010). Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 369–378.

Bibliography

[Pascual y Cabo and Soler, 2015] Pascual y Cabo, D. and Soler, I. (2015). Preposition Stranding in Spanish as a Heritage Language. *Heritage Language Journal*, 12:186–209.

[Pennebaker et al., 2007] Pennebaker, J., Booth, R., and Francis, M. (2007). Linguistic inquiry and word count (liwc2007).

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

[Peters et al., 2018] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

[Peters et al., 2017] Peters, M. E., Ammar, W., Bhagavatula, C., and Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.

[Pimentel et al., 2020a] Pimentel, T., Valvoda, J., Hall Maudslay, R., Zmigrod, R., Williams, A., and Cotterell, R. (2020a). Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

[Pimentel et al., 2020b] Pimentel, T., Valvoda, J., Maudslay, R. H., Zmigrod, R., Williams, A., and Cotterell, R. (2020b). Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.

[Polignano et al., 2019] Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., and Basile, V. (2019). Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it)*.

[Qi et al., 2020] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

[Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

[Raganato and Tiedemann, 2018] Raganato, A. and Tiedemann, J. (2018). An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

[Raghu et al., 2017] Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *NIPS*.

[Rajpurkar et al., 2016] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

[Ravichander et al., 2021] Ravichander, A., Belinkov, Y., and Hovy, E. (2021). Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

[Richards, 1971] Richards, J. C. (1971). A non-contrastive approach to error analysis. *English Language Teaching Journal*, 25:204–219.

[Richter et al., 2015] Richter, S., Cimino, A., Dell'Orletta, F., and Venturi, G. (2015). Tracking the evolution of written language competence: an nlp–based approach. In *Proceedings of 2nd Italian Conference on Computational Linguistics (CLiC-it)*.

[Robertson and Jones, 1976] Robertson, S. E. and Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146.

[Rogers et al., 2020] Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

[Sagae, 2021] Sagae, K. (2021). Tracking child language development with neural network language models. *Frontiers in Psychology*, 12.

[Sagae et al., 2005] Sagae, K., Lavie, A., and MacWhinney, B. (2005). Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 197–204. Association for Computational Linguistics.

[Sahakian and Snyder, 2012] Sahakian, S. and Snyder, B. (2012). Automatically learning measures of child language development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 95–99. Association for Computational Linguistics.

[Sanders and Noordman, 2000] Sanders, T. J. and Noordman, L. G. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse processes*, 29(1):37–60.

[Sanguinetti and Bosco, 2015a] Sanguinetti, M. and Bosco, C. (2015a). Parttut: The turin university parallel treebank. In *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, pages 51–69. Springer.

[Sanguinetti and Bosco, 2015b] Sanguinetti, M. and Bosco, C. (2015b). PartTUT: The turin university parallel treebank. In et al., R. B., editor, *Harmonization and Development of Re- sources and Tools for Italian Natural Language Processing within the PARLI Project*, page 51–69. Springer.

[Sanguinetti et al., 2018] Sanguinetti, M., Bosco, C., Lavelli, A., Mazzei, A., and Tamburini, F. (2018). PoSTWITA-UD: an Italian Twitter Treebank in universal dependencies. In *Proceedings of the Eleventh Language Resources and Evaluation Conference (LREC 2018).*

[Saphra and Lopez, 2019] Saphra, N. and Lopez, A. (2019). Understanding learning dynamics of language models with SVCCA. 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)*, pages 3257–3267, Minneapolis, Minnesota. Association for Computational Linguistics.

[Scarborough, 1990] Scarborough, H. S. (1990). Index of productive syntax. *Applied psycholinguistics*, 11(1):1–22.

[Serianni, 1989] Serianni, L. (1989). *Grammatica italiana. Italiano comune e lingua letteraria*. UTET, Torino.

[Serrano and Smith, 2019] Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

[Shen et al., 2018] Shen, Y., Tan, S., Sordoni, A., and Courville, A. (2018). Ordered neurons: Integrating tree structures into recurrent neural networks. In *International Conference on Learning Representations*.

[Silveira et al., 2014] Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. D. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014).*

[Slater et al., 2017] Slater, S., Ocumpaugh, J., Baker, R., Almeda, M. V., Allen, L., and Heffernan, N. (2017). Using natural language processing tools to develop complex models of student engagement. In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*, pages 542–547. IEEE.

[Sorace and Keller, 2005] Sorace, A. and Keller, F. (2005). Gradience in linguistic data. *Lingua*, 115(11):1497–1524.

[Sprouse, 2007] Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental synta. *Biolinguistics*, pages 1123–134.

[Straka et al., 2016] Straka, M., Hajic, J., and Strakova, J. (2016). UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC).*

[Straka and Straková, 2017] Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

[Ströbel, 2014] Ströbel, M. (2014). Tracking complexity of l2 academic texts: A sliding-window approach. *Unpublished master's thesis). RWTH Aachen University, Germany*.

[Suárez et al., 2019] Suárez, P. J. O., Sagot, B., and Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *Challenges in the Management of Large Corpora (CMLC-7) 2019*, page 9.

[Sun et al., 2019] Sun, C., Huang, L., and Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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)*, pages 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.

[Sun et al., 2020a] Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P., and Xiong, C. (2020a). Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. *arXiv preprint arXiv:2003.04985*.

[Sun et al., 2020b] Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., and Wang, H. (2020b). Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[Tenney et al., 2019a] Tenney, I., Das, D., and Pavlick, E. (2019a). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

[Tenney et al., 2019b] Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., et al. (2019b). What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[Tiedemann and Nygaard, 2004] Tiedemann, J. and Nygaard, L. (2004). The opus corpus-parallel and free: http://logos. uio. no/opus. Citeseer.

[Vajjala and Lõo, 2014] Vajjala, S. and Lõo, K. (2014). Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 113–127, Uppsala, Sweden. LiU Electronic Press.

[Vajjala and Lėo, 2014] Vajjala, S. and Lėo, K. (2014). Automatic CEFR Level Prediction for Estonian Learner Text. In *NEALT Proceedings Series Vol. 22, pages 113–127.*

[Vajjala and Rama, 2018] Vajjala, S. and Rama, T. (2018). Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

[van Halteren, 2004] van Halteren, H. (2004). Linguistic profiling for author recognition and verification. In *Proceedings of the Association for Computational Linguistics*, pages 200–207.

[van Schijndel et al., 2019] van Schijndel, M., Mueller, A., and Linzen, T. (2019). Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837, Hong Kong, China. Association for Computational Linguistics.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

[Vig and Belinkov, 2019] Vig, J. and Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

[Visalberghi and Costa, 1995] Visalberghi, A. and Costa, M. C. (1995). *Misurare e valutare le competenze linguistiche: guida scientifico-pratica per gli insegnanti*. La nuova Italia.

[Voita and Titov, 2020] Voita, E. and Titov, I. (2020). Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196.

[Volodina et al., 2016] Volodina, E., Pilán, I., Alfter, D., et al. (2016). Classification of swedish learner essays by cefr levels. *CALL communities and culture–short papers from EUROCALL*, 2016:456–461.

[Wang et al., 2018] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

[Wang et al., 2019] Wang, Y., Che, W., Guo, J., Liu, Y., and Liu, T. (2019). Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.

[Warstadt et al., 2019] Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., Alsop, A., Bordia, S., Liu, H., Parrish, A., Wang, S.-F., Phang, J., Mohananey, A., Htut, P. M., Jeretic, P., and Bowman, S. R. (2019). Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

[Warstadt et al., 2020] Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

[Weiss and Meurers, 2019] Weiss, Z. and Meurers, D. (2019). Analyzing linguistic complexity and accuracy in academic language development of german across elementary and secondary school. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA) at ACL*, pages 380–393.

[Wen et al., 2005] Wen, Q., Wang, L., and Liang, M. (2005). Spoken and written english corpus of chinese learners. *Foreign Language Teaching and Research Press*.

[Wiegreffe and Pinter, 2019] Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

[Wieting et al., 2016] Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2016). Charagram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515, Austin, Texas. Association for Computational Linguistics.

[Wilcox et al., 2019] Wilcox, E., Qian, P., Futrell, R., Ballesteros, M., and Levy, R. (2019). Structural supervision improves learning of non-local grammatical dependencies. 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)*, pages 3302–3312, Minneapolis, Minnesota. Association for Computational Linguistics.

[Wilson et al., 2017] Wilson, J., Roscoe, R., and Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing*, 34:16–36.

[Witten et al., 1999] Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., and Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with java implementations.

[Xu et al., 2012] Xu, B., Guo, X., Ye, Y., and Cheng, J. (2012). An improved random forest classifier for text categorization. *J. Comput.*, 7(12):2913–2920.

[Xu et al., 2021] Xu, W., Aw, A. T., Ding, Y., Wu, K., and Joty, S. (2021). Addressing the vulnerability of NMT in input perturbations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 80–88, Online. Association for Computational Linguistics.

[Yang et al., 2015] Yang, W., Lu, X., and Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28:53–67.

[Yang et al., 2019] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

[Yin et al., 2020] Yin, F., Long, Q., Meng, T., and Chang, K.-W. (2020). On the robustness of language encoders against grammatical errors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3386–3403, Online. Association for Computational Linguistics.

[Yoon, 2017] Yoon, H.-J. (2017). Linguistic complexity in l2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, 66:130–141.

[Zeldes, 2017] Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

[Zeman et al., 2019] Zeman, D., Nivre, J., Abrams, M., and et al. (2019). Universal dependencies 2.5. In *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL)*.

[Zhang and Bowman, 2018] Zhang, K. and Bowman, S. (2018). Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

[Zhou et al., 2020] Zhou, J., Zhang, Z., Zhao, H., and Zhang, S. (2020). LIMIT-BERT : Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461, Online. Association for Computational Linguistics.