Complex Word Identification for Italian Language: a dictionary-based approach

Laura Occhipinti
University of Bologna, Italy
laura.occhipinti3@unibo.it

Abstract

Assessing word complexity in Italian poses significant challenges, particularly due to the absence of a standardized dataset. This study introduces the first automatic model designed to identify word complexity for native Italian speakers. A dictionary of simple and complex words was constructed, and various configurations of linguistic features were explored to find the best statistical classifier based on Random Forest algorithm. Considering the probabilities of a word to belong to a class, a comparison between the models' predictions and human assessments derived from a dataset annotated for complexity perception was made. Finally, the degree of accord between the model predictions and the human inter-annotator agreement was analyzed using Spearman correlation. Our findings indicate that a model incorporating both linguistic features and word embeddings performed better than other simpler models, also showing a value of correlation with the human judgements similar to the inter-annotator agreement. This study demonstrates the feasibility of an automatic system for detecting complexity in the Italian language with good performances and comparable effectiveness to humans in this subjective task.

Keywords: complex word identification, Italian language, lexical complexity.

1 Introduction

Identifying the complexity of a word is a very challenging process that requires a series of linguistic reflections intertwined with the concept of complexity itself (Pallotti, 2015). While humans can intuitively perceive word simplicity, translating this intuition into quantitative parameters for automatic systems is challenging.

The task of Complex Word Identification (CWI) aims to pinpoint those words that may pose decoding challenges for certain readers due to a variety of linguistic features (Shardlow, 2013). The concept of linguistic complexity indeed is closely intertwined with the readability and accessibility of texts (Chen and Meurers, 2019). Recognizing complex words is crucial, not only for readers with learning difficulties, such as dyslexia or aphasia (Stajner, 2021; De Hertog and Tack, 2018), but also for native speakers, since understanding word meanings is fundamental for comprehension (Carroll et al., 1998). Studies related to CWI have seen a significant increase in recent years, either as a part of lexical simplification systems (Saggion and Hirst, 2017), or as an independent task, promoted by several shared tasks (Paetzold and Specia, 2016; Yimam et al., 2018; Shardlow et al., 2021). In the latter case, it is very useful for the development of systems aiming at facilitating foreign language acquisition, creating reading tools for individuals with limited linguistic skills, and enhancing accessibility for native speakers (Gooding and Kochmar, 2018, 2019). Despite the importance of CWI, the development of such systems has been limited to a few languages, mainly due to the scarcity of necessary linguistic resources and the high costs associated with their development (Štajner et al., 2022).

To the best of our knowledge, there have been no studies directly addressing the CWI in the Italian language, even though research has focused on text simplification (Brunato et al., 2022). The absence of requisite databases classifies Italian as a 'low-resource language' for this specific task. The main contribution of this article is to propose the first automatic system to identify lexical complexity specifically designed for native Italian speakers, motivated by educational concerns (ISTAT, 2021) and the need to understand perceived complexity under typical conditions. We created a dataset of individual lexical entries, labelled as simple or complex (3.1) and selected various linguistic features (3.2), through which a classifier system could be trained in a supervised setting (3.3). Our approach

is dictionary-based and context agnostic (Billami et al., 2018; Baeza-Yates et al., 2015). We considered the probability of each item belonging to a certain class as the prediction of word complexity. Finally, the system was validated against a dataset containing human judgements regarding the perceived complexity of selected words (3.3).

2 Related Work

Recent investigation in CWI focused on the development of statistical classifiers that can accurately assign lexical items to specific complexity classes based on labelled data (Paetzold and Specia, 2016; Yimam et al., 2018). Classification systems typically utilize feature-based approaches or neural networks with word embeddings to enhance prediction accuracy (Aroyehun et al., 2018). Most CWI studies classify word complexity in two primary ways: **binary classification**, labeling words as either simple or complex (0-1), and **continuous classification**, where words receive a complexity score on a continuum from very simple to very complex. In recent years, it has become more common to use Lexical complexity prediction name to refer to the latter (North et al., 2023). Among the statistical classifiers, Support Vector Machines, Decision Trees, Random Forests, Logistic Regression, and Recurrent Neural Networks have been prominently used (Yimam et al., 2018; Shardlow et al., 2021).

For the Italian language, the few studies concerning lexical simplification (Tonelli et al., 2016; Brunato et al., 2015) have overlooked this task deemed crucial for the proper execution of simplification (Shardlow, 2014). The words to be simplified were selected exclusively on the basis of the frequency parameter (Brunato et al., 2022) relying on *Nuovo Vocabolario di base* (De Mauro and Chiari, 2016), which is a fundamental lexicon for the Italian language, comprising approximately 7,000 selected words. This approach poses significant limitations, as words outside this vocabulary are often prematurely considered complex, and potential substitutions are restricted to those within the same lexicon. This approach does not take into account the nuanced and multifaceted nature of linguistic complexity, so relying on a single measure such as frequency can lead to oversimplification (Bott et al., 2012): frequency is strongly linked to the reference corpus used to calculate it.

3 Methods

In this study, we developed a binary classification system using a dataset of isolated words, created due to the absence of comprehensive resources for the CWI task. Recognizing that word complexity is intrinsically context-dependent and that complexity itself is a gradient, our approach was constrained by resource limitations. The creation of a large-scale dataset capable of training models with nuanced human judgments in context would require significant time and resources. Consequently, we opted for a more manageable solution by employing a word list, which is computationally 'small and easily tractable' (Kilgarriff et al., 2014: 124). This choice was motivated by its ability to provide representative data for our purpose. The decision to adopt binary classification reflects not only these practical constraints but also avoids the subjectivity inherent in gradual classifications without extensive contextual data. This pragmatic approach aims to establish a foundational methodology that can be expanded as more comprehensive data become available.

We selected various linguistic features to characterize the complexity of our items. From these features, our model learned to predict the complexity classification and the likelihood of a word belonging to a particular class. While it is recognized that the probability of a target word being classified as simple or complex does not directly predict its degree of complexity (North et al., 2023), the continuous probabilistic values generated by our model provide valuable insights into the nuanced nature of word complexity. These probabilistic values reflect the uncertainty inherent in the classification process: higher probabilities indicate a stronger likelihood that a word has intricate linguistic properties, whereas lower probabilities suggest simpler linguistic structures. These predicted values can then be compared with complexity assessments derived from human judgments, which serve as our gold standard. This methodology helps bridge the gap between objective classification and subjective perception, enhancing our understanding of lexical complexity.

3.1 Dataset

Recognizing the importance of context in the domain of complexity perception, our approach was limited by the absence of an available dataset for the CWI task. As a result, we opted to build a list

of words for the purpose of training an automatic complexity classification system. The word selection was not based on frequency parameter, but on a series of heuristics aimed at minimising personal bias in the selection of lexical items. Considering the challenges in defining complexity (Miestamo et al., 2008), we decided to classify as simple all words that should be known or learned by Italian L2 learners, as outlined in levels A1-B1 of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). The selection of these words was made by exploring various linguistic resources developed and/or used for teaching Italian to non-native speakers.

Merlin Corpus (Wisniewski et al., 2013; Boyd et al., 2014) This corpus is a linguistic resource aimed at exploring texts produced by L2 students for Italian, Czech, and German languages. The Italian section includes 813 texts, each associated with specific CEFR levels by professional evaluators and featuring metadata related to various linguistic levels. We specifically selected texts with metadata corresponding to the 'Vocabulary Range' from levels A1 to B1. We extracted all forms and manually corrected any orthographic errors, recognizing that this corpus also serves as a representation of the errors made by the writers. Despite these graphical errors, we opted to include 'wrong' words as they are undoubtedly familiar to the writers, who employ them in their writing.

Kelly (Kokkinakis and Volodina, 2011) This resource was developed as part of the European Kelly project, aimed at creating vocabularies for nine languages, including Italian. The Kelly word list reflects modern usage and captures the core vocabulary of each language, selected through an objective process based on corpus analysis and pedagogical criteria. Words were categorized across levels based on daily themes deemed essential by the CEFR. This categorization guided their inclusion in our study due to their alignment with established language proficiency standards.

ELI: vocabolario illustrato junior (ELI Publishing Group, 2020) This dictionary is designed for a target audience of young students, presenting basic vocabulary ranging from levels A1 to A2, using graphical representations to link images with words effectively. It organizes 936 words into 45 themes relevant to everyday contexts. We chose this tool because it targets a beginner audience, suggesting that the included words are widely recognized within the native speaking population.

Word lists identified by University for Foreigners of Perugia [1] The University for Foreigners of Perugia offers a range of open-access resources essential for teaching Italian to foreign learners. We focused on the section relating to the lexical lists for each level from A1 to B2, developed through extensive validation by linguistic and pedagogical experts. After downloading these lists, we removed additional details such as word index numbers and grammatical descriptions.

To these resources, purely related to L2 teaching, we added

Varless (Burani et al., 2001; Barca et al., 2002) This resource includes a list of simple Italian nouns accompanied by various lexical and sub-lexical variables such as age of acquisition, familiarity, concreteness, and frequency metrics. These variables significantly affect how words are perceived by speakers, with early-acquired words being recognized and named more rapidly and accurately. We included this resource because the words are classified as simple based on their acquisition and familiarity profiles.

The integration of these resources involved the exclusion of common vocabulary, multi-word expressions (which are not within the scope of this paper), and the normalization of word forms to their respective reference lemma. This process yielded a consolidated list comprising 5,382 lemmas.

It was not feasible to apply the same criterion in selecting complex words, as digital resources available for levels B2-C2 are limited and primarily focus on pragmatic aspects of the language. Therefore, for complex words, a dictionary containing words defined as difficult or truly difficult in the Italian language was utilized.

Dizionario delle parole difficili e difficilissime (Vallardi, 2016) This dictionary comprises Italian words that are arcane, remote, or enigmatic, and seldom used in colloquial, television, or journalistic contexts. It spans various domains such as literature, science, and technology, serving as a repository of linguistic richness and cultural heritage. From its approximately 13,000 lemmas, we carefully selected about 8,000 terms for our dataset to ensure sample balance and integrity.

[1] https://www.unistrapg.it.

The final dataset consists of a list of words labelled as simple (0) or complex (1), comprising 13,319 lemmas distributed between the two categories as follows: 5,382 simple lemmas; 7,937 complex lemmas.

3.2 Features

Defining the linguistic features for identifying lexical complexity is critical, involving several interrelated aspects (Collins-Thompson, 2014). The selection of features is based on their strong psycholinguistic evidence, which significantly impacts the perception of complexity. These features are calculated using both the word form and its lemma, with lemmatization performed using the Italian SpaCy model[2]. Given their general applicability and the robust psycholinguistic backing, these measures are particularly suited for our target population of native Italian speakers.

Frequency Frequency appears to be the predominant and essential parameter in all approaches to CWI, supported by various pieces of psycholinguistic evidence (Segui et al., 1982). For instance, frequency is significant for gauging familiarity with a term. We used two reference corpora to calculate frequency, aiming to reduce bias from corpus composition. The first corpus we considered is the **ItWac corpus** (Baroni et al., 2009), that is a 2 billions word corpus, created from the web. The other is **Subtlex-it** (Crepaldi et al., 2015), a word frequency list based on movie and tv show subtitles for approximately 520,000 Italian word-forms. For both, we calculated the row frequency for each lemma, representing the number of occurrences within the corpus. The two frequencies are treated separately and the values were converted into base 10 logarithmic scales, returning 0 if before normalization frequency value was 0.

3.2.1 Surface Features

We considered some surface linguistic parameters that are crucial from a psycolinguistic standpoint (Perfetti et al., 2001) because they significantly affect reading and decoding times:

Word lenght The number of characters in the word.

Syllable count The number of syllables of the word calculated using `Pyphen`[3].

Vowels count The number of vowels presented in the word. This feature was determined by iterating through each character in the word and checking if it corresponds to any vowel, including accented characters. Notably, we included vowels typical of the Italian language in our analysis.

3.2.2 Linguistic features

In addition to the superficial characteristics of words, it is necessary to carry out deeper analyses concerning the types of words and the meanings attached to them.

Stop words Recognizing whether a word is a stopword is crucial for determining its complexity. Stopwords, such as articles, prepositions, and conjunctions, are frequently encountered and widely understood by readers. Therefore, identifying whether a word is a stopword provides insights into its familiarity and ease of comprehension. This measure was computed using SpaCy.

Number of senses We assessed the number of senses for each lemma using the ItalWordNet (Roventini et al., 2000). This analysis helps clarify the semantic complexity of words by revealing how many different meanings a word can have, indicating its potential to cause decoding ambiguities for readers.

3.2.3 Morphological Features

We selected features related to word morphology, crucial for defining lexical complexity. Most of the morphosyntactic information we have for Italian language from existing corpora or from readability measures concerns the class to which words belong. Beyond this, we incorporated details about internal structure of the word (Baerman et al., 2015).

POS-tag We categorized the lemma into predefined POS labels, assessing the presence or absence of each label using a list. The provided method iterates through the lemma, assigning a value of 1 to the corresponding POS label if matched, otherwise 0. Using SpaCy, we predicted the POS labels while consolidating certain subcategories into broader groups to simplify analysis. We merged 'VERB' and 'AUX' into a category 'VERB', 'NOUN' and 'PROPN' into 'NOUN', and 'CCONJ' and 'SCONJ' into 'CONJ'.

Number of morphemes We calculated the number of morphemes, the smallest units of meaning,

[2] https://spacy.io/models/it#it_core_news_sm.

[3] https://pyphen.org/.

that composed the word. In this way, we can provide indications about the amount of information readers must decode to understand the term they are facing (Brezina and Pallotti, 2019). Italian is an inflected language (Grandi, 2011) that employs inflection, derivation, and composition to modify words. The number and type of morphemes in a word are crucial indicators of its complexity; for instance, a derived word is more complex than a simple one, as it contains more elements to decode (Rastle and Davis, 2003).

felice is simpler than **infelice**

The adding of the prefix **in-** to the base form leads us to decode the meaning of **felice** (happy) to which a negation is added. For this reason, we could argue that the word *infelice* (unhappy) is marked compared to *felice* and increases its degree of complexity. To calculate the morphological composition of a word, we used a Convolutional Neural Model[4] trained on an Italian hand-checked dataset[5] to obtain an automatic morphological segmentation.

Morphological Density This measure quantifies morphological complexity at the word level (Sandra, 1994; Manova et al., 2020), defined as the ratio of the number of morphemes to word length. It helps analyze a word's structural complexity, indicating how densely packed it is with meaningful units. A higher morphological density suggests a more complex word, with many meaningful units condensed into a shorter length, possibly making it more challenging to comprehend. Lower density, conversely, implies simpler and potentially easier-to-understand words.

Frequency of lexical morpheme We determined the frequency of the lexical morpheme that most conveys the meaning of the word (Amenta and Crepaldi, 2012). Employing our morphological segmentator on the ItWac corpus, enabled us to dissect the word into segments and aggregate the frequencies of individual morphemes. The use of lexical morpheme frequency as a complexity indicator is based on the idea that even if a word is unfamiliar as a whole, its component morphemes may be common in the language and more recognizable (Colé et al., 1997). Such words are inherently more relatable to familiar concepts due to the frequent occurrence of their constituent morphemes. Leveraging the familiarity of these morphemes enhances the transparency and interpretability of the word's meaning. We adopted the longest splitting morpheme as the lexical one, as this heuristic aligns with many cases in Italian, acknowledging that there are exceptions to this rule. Additionally, the frequency values have been logarithmized to facilitate analysis.

Word Embedding We utilized pre-trained word embeddings from FastText for Italian (Joulin et al., 2016, 2017), which provides word vector representations with 300 dimensions. The model used in our study was trained on Wikipedia and Common Crawl datasets[6]. These embeddings provided vector representations for each word in our dataset, primarily comprising isolated items, allowing us to incorporate contextual features into our analysis.

3.3 Models

We evaluated the performance of a classifier built using Random Forest (Breiman, 2001) implemented with the `scikit-learn library` (Pedregosa et al., 2011). Specifically, we utilized the `RandomForestClassifier module` provided by scikit-learn[7]. To assess the classifier's performance, we employed 12-fold cross-validation over the training data. We selected different configuration of features to understand which model is the best in prediction [8] after training and to comprove that only one frequency value is not enough for an efficient prediction:

1. Frequency model, that utilizes the two parameters related to Frequency in 3.2.3.
2. Feature-based Model, that leverages the eleven linguistic features discussed above (frequencies, surface, linguistic and morphological features presented in 3.2.3).
3. Embedding Model, that utilizes only pre-trained word embeddings (Word embedding paragraph in 3.2.3).

[4] https://github.com/AlexeySorokin/NeuralMorphemeSegmentation/tree/master.

[5] The details of the implementation of this system and the database used will be discussed in a forthcoming paper currently in preparation.

[6] https://fasttext.cc/docs/en/crawl-vectors.html.

[7] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[8] For further practical details concerning the best performing model, the source code, and the resources used, interested parties are encouraged to directly contact the author via email.

4. Total model, that integrates both feature-based and embedding-based features.

These models were trained and evaluated on the dataset presented in Section 3.1, consisting of 13,319 words. To establish a robust evaluation, we employed the `train_test_split function` from `scikit-learn` to partition the dataset into training and testing sets. The split allocated 70% of the data for training, amounting to approximately 9,000 words, while the remaining 30% (about 4,000 words) was reserved for testing. We shuffled the data before splitting to mitigate any bias, and subsequently instantiated a Random Forest Classifier model with the random state set to 42 for reproducibility. For performance evaluation, each model underwent rigorous assessment on four key parameters: Accuracy, Precision, Recall, and F1 Score on the test set. These metrics are commonly employed to evaluate the performance of classification systems (North et al., 2023) (results in Section 4.1).

3.4 External validation

After validating our models on the original test set, we extended our evaluation by testing the models on an external resource. This dataset[9] consists of 600 sentences, in each of which a target word was identified. For each word, we gathered a minimum of 10 human judgements regarding the complexity level of the target for a generic native speaker of Italian. The data were annotated exclusively by native speakers, that had the task of assigning a level of complexity to each target word, using a Likert scale ranging from 1 to 5:

- 1: very easy - Words which are very familiar
- 2: easy - Words which are mostly familiar
- 3: neutral - When the word is neither difficult or easy
- 4: difficult - Words which you are unclear of the meaning, but may be able to infer from the context
- 5: very difficult - Words that are very unclear.

This dataset was built as a resource of lexical complexity prediction; for information on how the dataset was constructed and annotated, please refer to (Shardlow et al., 2024). This resource represents our gold standard. For each target word, the average score between annotations was used a single human-derived complexity value that was compared with our models predictions. We transformed the values from a range of 1 to 5 to a scale of 0 to 1 using the min-max normalization (Abdi, 2007). This normalisation aligns the data with our model's output range (between 0 and 1), facilitating effective analysis and consistent evaluation of model performance. The validation metrics normally used to evaluate lexical prediction system performance (North et al., 2023) are Pearson Correlation, Spearman's Rank, mean absolute error, and mean squared error (Hastie et al., 2009). We calculated these measures evaluating the relations between our predictive model outputs (we excluded the model with the lowest performance) and the aggregated human judgements (results in Section 4.2). We conducted a further analysis by comparing the predictions of the best model with the level of inter-annotator agreement (Artstein, 2017) observed in our resource. While the initial comparison provided valuable insights into the model's performance against a consolidated human judgement, assessing its agreement with multiple human annotations offers a more comprehensive understanding of its effectiveness. The choice of agreement measure depends on the data nature and and the objectives of the study. Since our data are ordinal, with complexity values ranging from 1 to 5, we chose to use Spearman correlation to calculate agreement and not kappa (Rau and Shih, 2021), which is more suited for nominal or categorical data. The Spearman correlation is suitable for ordinal variables as it accounts for the rank order of values without assuming a linear relationship, offering greater flexibility in measuring agreement. Furthermore, it is particularly adequate in cases where the order of the values is significant, but no specific assumptions can be made about the distribution of the data or the uniform intervals between the categories. Our annotation task involved ordinal ratings, where the magnitude of difference between ratings carries significance, thus making this measure a more appropriate choice for assessing agreement. The Spearman correlation coefficients were calculated using the `spearmanr` function from the `scipy.stats` module by iterating through combinations of annotator pairs. After calculat-

[9] https://github.com/MLSP2024/MLSP_Data/.

Model	Accuracy	Precision	Recall	F1
Frequency_based	0.8826	0.8995	0.9214	0.9103
Feature_based	0.9006	0.9137	0.9346	0.9243
Embedding_model	0.8943	0.9055	0.9289	0.9170
Total_model	**0.9149**	**0.9237**	**0.9466**	**0.9350**

Table 1: Classifier results

ing the correlation coefficients, we computed the overall average correlation coefficient across the entire dataset. We operated in the same way with the results of our best model. We also treated our model's predictions as an additional annotator to calculate its agreement with all human judgments. The final value is the result of the average of correlation values of our model with all the single value of complexity defined by annotators. The comparison between the two values is reported in Section (4.3).

4 Results and discussion

4.1 Model results on classification

The calculated performances of the four models on the test set are reported in Table 1. The Total_Model showed the best performances across all validation metrics, outperforming not only the simplest model based on frequency but also the more complex Feature-based and Embedding-based models. The Frequency-based model, even if inferior to the others, still demonstrated acceptable performance, thus highlighting the key significance of frequency. However, a model exclusively based on frequency has a key limitation: words that are not represented in the corpus considered will be labelled as complex with a probability of around 0.99. Thus it is essential to include more words classification features in the training of the model. The Feature-based Model and the Embedding Model exhibit comparable performance across various evaluation metrics. However, it is crucial to recognize the underlying differences in their methodologies and interpretability. The Feature-Based Model provides a transparent framework enabling granular analysis of the impact of individual features on prediction. This transparency facilitates the identification of specific features that contribute significantly to the model's predictive performance. In contrast, the Embedding model operates on distributed representations of words in a highly dimensional semantic space, making it inherently more opaque and difficult to interpret. The Total_Model, thanks to its adeptness in harnessing the respective strengths of each method, shows superior performance due to its capacity to leverage not only the linguistic features we selected but also the connected semantic representations in word embeddings.

4.2 Model results on complexity prediction

In Table 2, we reported the results of our models in comparison with the gold standard dataset, containing the human annotations. These results provide insights into the effectiveness of our models in predicting complexity, with the Total_Model demonstrating again an overall superior performance compared to the others. Pearson's Correlation evaluates the linear relationship between predicted and actual complexity values, indicating the strength of this relationship. For the Total_model, the Pearson Correlation is 0.5503, demonstrating a relatively strong linear relationship between predicted and actual values. On the other hand, Spearman's Rank assesses the monotonic relationship between predicted and actual complexity values, regardless of linearity. The Total_model achieved a Spearman's Rank of 0.5528, indicating a strong monotonic relationship between predicted and actual values. The two correlation coefficients are quite similar, suggesting that the model performs well in capturing both linear and non-linear trends in complexity prediction. Mean Squared Error (MSE) measures the average squared difference between predicted and actual complexity values, with lower values indicating better performance. For this parameter the Embedding_Model shows a slightly lower value than the Total_Model. Similarly, Mean Absolute Error (MAE) calculates the average absolute difference between predicted and actual complexity values. The Total_model achieved a MAE of 0.2393, suggesting that its predictions are closer to the true complexity values. Despite the slightly lower MSE for the Embedding_model, the Total_model still demonstrates superior performance overall, as evidenced by its higher correlation coefficients and

Model	Pearson Correlation	Spearman's	Mean squared error	Mean absolute error
Feature_based	0.5331	0.5403	0.1231	0.2718
Embedding_model	0.4762	0.4752	**0.0927**	0.2482
Total_model	**0.5503**	**0.5528**	0.0965	**0.2393**

Table 2: Results of models on complexity prediction

System agreement	Spearman's Rank
Inter-annotation agreement	**0.4196**
Total_Model agreement	0.4145

Table 3: Comparison between inter-annotator agreement and model predictions

lower error metrics compared to the other models.

4.3 Comparison with inter-annotator agreement

The comparison between the Spearman correlation coefficients obtained from the assessments of human annotators and those derived from the predictions of our best model reveals a notable similarity. The results are reported in Table 3. Both values, falling within the same range, demonstrate a significant degree of agreement between the model's predictions and human evaluations. The close proximity of these figures underscores the model's proficiency in capturing the complexity assessed by humans. These findings imply that there are opportunities for improvement both within our system and in fostering increased inter-agreement among human annotators, thereby potentially refining the model's ability to accurately capture the complexities inherent in the task.

5 Conclusions

In this study, we introduced the first system aimed at identifying complex words within the Italian language, marking the initial exploration of this task for this linguistic domain.

The absence of specific datasets prompted us to build a dictionary comprising approximately 13,000 words annotated for simplicity and complexity. An appropriate selection of descriptive features of word complexity made it possible to train a classification model in different configurations. We tested our models on a test set and on an external dataset, containing human judgements on word complexity, our gold standard. From the different validation analyses we saw that the best model is Total_Model that integrates the linguistic features with the word embedding.

We conducted a further analysis by comparing the Total_Model to our gold standard. This dataset was annotated by multiple native Italian speakers and for this reason we decided to calculate the inter-annotation agreement and compared it with model-human correlation. In this way we not only validated the reliability of our dataset and the fidelity of our predictive model but also established the basis for a meaningful comparison between human and machine assessments.

Our analysis revealed that the average correlation of each predicted value from our model with the inter-annotator agreement falls within the same range, suggesting that our system is as effective as human judgment in subjective tasks such as this. To enhance inter-annotator agreement and the robustness of our findings, future efforts will focus on increasing the sample size and the number of annotators. Expanding the sample size will cover a broader lexical domain and provide a diverse set of words and contexts, thereby improving the model's generalizability to unseen data. This broader coverage supports robust statistical testing and validation, minimizing the influence of outliers. Incorporating more annotators is crucial for enriching the diversity of perspectives in the evaluation process, which is particularly important in subjective assessments where personal experiences, linguistic backgrounds, and individual biases might skew judgments. A larger pool of annotators diminishes these biases, fostering a balanced and representative consensus on lexical complexity. Furthermore, this approach allows for more detailed inter-annotator agreement analyses, clearly highlighting areas of consensus and disagreement. Together, these strategies not only enhance the reliability of our annotations but also improve the overall accuracy and applicability of our model.

The main limitation of our approach resides in the characteristics of the dataset we used to train our models. Our dataset is composed by words presented in isolation, thus disregarding crucial

contextual cues essential for understanding word meanings and disambiguation. We acknowledge the critical role of context in complexity analysis and recognize the necessity of incorporating specific contextual information where the target word appears. Moving forward, our aim is to advance in this direction by expanding effective datasets that integrate contextual frameworks for training our word CWI systems. Although challenges persist in the field of CWI, our study lays some groundwork for exploring this task for Italian and underscores the potential of automated systems in this domain.

In the future, collaborative efforts and advancements in building datasets and refining models will be crucial for advancing the field and uncovering new insights into language complexity. This methodology could enhance the precision of readability measures (Dell'Orletta et al., 2011), particularly in terms of lexical range and lexical sophistication. Moreover, such a system can be an essential component in a text simplification pipeline. By identifying words that may pose comprehension challenges, the system not only flags these words for potential replacement but also assists in suggesting simpler alternatives. This functionality ensures that the replacements not only match the original words' meanings as closely as possible but also contribute to a text that is overall easier to understand.

While this study focuses on the Italian language, the methodologies and models we have developed have the potential to be adapted for other languages, especially those considered low-resource in the context of computational linguistic tools. By leveraging similar linguistic resources and adjusting the feature sets to accommodate language-specific characteristics, researchers can extend this approach to support complex word identification across diverse linguistic domains.

References

Herve Abdi. 2007. Multiple correlation coefficient. *Encyclopedia of measurement and statistics*, 648(651):19.

Simona Amenta and Davide Crepaldi. 2012. Morphological processing as we know it: An analytical review of morphological effects in visual word identification. *Frontiers in psychology*, 3:232.

Segun Taofeek Aroyehun, Jason Angel, Daniel Alejandro Pérez Alvarez, and Alexander Gelbukh. 2018. Complex word identification: Convolutional neural network vs. feature engineering. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 322–327.

Ron Artstein. 2017. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.

Matthew Baerman, Dunstan Brown, and Greville Corbett. 2015. *Understanding and measuring morphological complexity*. Oxford University Press, USA.

Ricardo Baeza-Yates, Luz Rello, and Julia Dembowski. 2015. CASSA: A context-aware synonym simplification algorithm. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1385, Denver, Colorado. Association for Computational Linguistics.

Laura Barca, Cristina Burani, and Lisa S Arduino. 2002. Word naming times and psycholinguistic norms for italian nouns. *Behavior research methods, instruments, & computers*, 34:424–434.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43:209–226.

Mokhtar Billami, Thomas François, and Núria Gala. 2018. ReSyf: a French lexicon with ranked synonyms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2570–2581, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Stefan Bott, Luz Rello, Biljana Drndarević, and Horacio Saggion. 2012. Can spanish be simpler? lexsis: Lexical simplification for spanish. In *Proceedings of COLING 2012*, pages 357–374.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. 2014. The merlin corpus: Learner language and the cefr. In *LREC*, pages 1281–1288. Reykjavik, Iceland.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Vaclav Brezina and Gabriele Pallotti. 2019. Morphological complexity in written l2 texts. *Second language research*, 35(1):99–119.

Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2022. Linguistically-based comparison of different approaches to building corpora for text simplification: A case study on italian. *Frontiers in Psychology*, 13:707630.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41.

Cristina Burani, Laura Barca, and Lisa Saskia Arduino. 2001. Una base di dati sui valori di età di acquisizione, frequenza, familiarità, immaginabilità, concretezza, e altre variabili lessicali e sublessicali per 626 nomi dell'italiano. *Giornale Italiano di Psicologia*, 28(4):839–856.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Association for the Advancement of Artificial Intelligence.

Xiaobin Chen and Detmar Meurers. 2019. Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, 32(4):418–447.

Pascale Colé, Juan Segui, and Marcus Taft. 1997. Words and morphemes as units for lexical access. *Journal of Memory and Language*, 37(3):312–330.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Council of Europe. 2001. *Common of European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Davide Crepaldi, Simona Amenta, Mandera Pawel, Emmanuel Keuleers, and Marc Brysbaert. 2015. Subtlex-it. subtitle-based word frequency estimates for italian. In *Proceedings of the Annual Meeting of the Italian Association For Experimental Psychology*, pages 10–12.

Dirk De Hertog and Anaïs Tack. 2018. Deep learning architecture for complex word identification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 328–334.

Tullio De Mauro and I Chiari. 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*.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read–it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.

ELI Publishing Group. 2020. *ELI Picture Dictionary Junior: Picture Dictionary Junior - Italian*. ELI Publishing.

Sian Gooding and Ekaterina Kochmar. 2018. Camb at cwi shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194.

Sian Gooding and Ekaterina Kochmar. 2019. Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153.

Nicola Grandi. 2011. Evaluative affixes between inflection and derivation: a typological survey. In *Societas Linguistica Europaea–44th Annual Meeting*.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

ISTAT. 2021. Livelli di istruzione e ritorni occupazionali. Technical report, Istituto nazionale di Statistica.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48:121–163.

Sofie Johansson Kokkinakis and Elena Volodina. 2011. Corpus-based approaches for the creation of a frequency based vocabulary list in the eu project kelly–issues on reliability, validity and coverage. *Proceedings of eLex*, 2011:129–139.

Stela Manova, Harald Hammarström, Itamar Kastner, and Yining Nie. 2020. What is in a morpheme? theoretical, experimental and computational approaches to the relation of meaning and form in morphology. *Word Structure*, 13(1):1–21.

Matti Miestamo, Fred Karlsson, and Kaius Sinnemäki. 2008. Language complexity. *Language Complexity*, pages 1–374.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Gabriele Pallotti. 2015. A simple view of linguistic complexity. *Second Language Research*, 31(1):117–134.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Charles Perfetti, Julie Van Dyke, and Lesley Hart. 2001. The psycholinguistics of basic literacy. *Annual review of applied linguistics*, 21:127–149.

Kathleen Rastle and Matthew Davis. 2003. Reading morphologically complex words. *Masked priming: The state of the art*, pages 279–305.

Gerald Rau and Yu-Shan Shih. 2021. Evaluation of cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of english for academic purposes*, 53:101026.

Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna. 2000. Italwordnet: a large semantic database for italian. In *LREC*.

Horacio Saggion and Graeme Hirst. 2017. *Automatic text simplification*, volume 32. Springer.

Dominiek Sandra. 1994. The morphology of the mental lexicon: Internal word structure viewed from a psycholinguistic perspective. *Language and cognitive processes*, 9(3):227–269.

Juan Segui, Jacques Mehler, Uli Frauenfelder, and John Morton. 1982. The word frequency effect and lexical access. *Neuropsychologia*, 20(6):615–627.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st annual meeting of the association for computational linguistics proceedings of the student research workshop*, pages 103–109.

Matthew Shardlow. 2014. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1583–1590, Reykjavik, Iceland. European Language Resources Association (ELRA).

Matthew Shardlow, Fernando Alva-Manchego, Riza Theresa Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Laura Occhipinti, et al. 2024. An extensible massively multilingual lexical simplification pipeline dataset using the multils framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)@ LREC-COLING 2024*, pages 38–46.

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

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical simplification benchmarks for english, portuguese, and spanish. *Frontiers in Artificial Intelligence*, 5:991242.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016*, pages 291–296.

Vallardi. 2016. *Dizionario delle parole difficili e difficilissime*. Vallardi Editore.

Katrin Wisniewski, Karin Schöne, Lionel Nicolas, Chiara Vettori, Adriane Boyd, Detmar Meurers, Andrea Abel, and Jirka Hana. 2013. Merlin: An online trilingual learner corpus empirically grounding the european reference levels in authentic learner data. In *ICT for Language Learning 2013, Conference Proceedings, Florence, Italy. Libreriauniversitaria. it Edizioni*.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 66–78.