Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 7204–7211
Marseille, 11–16 May 2020

MALT-IT2: A New Resource to Measure Text Difficulty in light of CEFR levels for Italian L2 learning

Luciana Forti, Giuliana Grego Bolli, Filippo Santarelli, Valentino Santucci, Stefania Spina
Department of Humanities and Social Sciences
University for Foreigners of Perugia
Piazza G. Spitella, 3 - 06123 Perugia (Italy)
{luciana.forti, giuliana.grego, filippo.santarelli, valentino.santucci, stefania.spina}@unistrapg.it

Abstract

This paper presents a new resource for automatically assessing text difficulty in the context of Italian as a second or foreign language learning and teaching. It is called MALT-IT2, and it automatically classifies inputted texts according to the CEFR level they are more likely to belong to. After an introduction to the field of automatic text difficulty assessment, and an overview of previous related work, we describe the rationale of the project, the corpus and computational system it is based on. Experiments were conducted in order to investigate the reliability of the system. The results show that the system is able to obtain a good prediction accuracy, while a further analysis was conducted in order to identify the categories of features which mostly influenced the predictions.

Keywords: text difficulty, Italian L2, assessment

1. Introduction

Determining the appropriateness of a text in relation to CEFR levels has a key role in the context of second or foreign language teaching and assessment. The suitability of a text for a certain learner group is generally established on the basis of its thematic content, as it needs to be adequate to the learners' personal motivation for learning, and to its linguistic content, as it needs to be in line with their level of proficiency. In regards to the latter criterion, evaluations of the difficulty of a text are generally conducted subjectively, both when a text needs to be chosen as a component of a language test, and when it needs to be chosen for classroom use. The formal and quantitative characteristics of a text, in particular, have a major role in determining the comprehensibility of that text, as they will impose specific cognitive demands upon the reader when approaching the text (Bachman and Palmer, 2010; Purpura, 2014). However, in the absence of an objective difficulty measure, not only do subjective judgments lead to potential discrepancies between different individual evaluations made by teachers and/or test developers (François et al., 2014), but it also makes it impossible to gain insight into a large number of quantitative features that have an impact on the difficulty of a text, and on the cognitive load involved.

A number of research projects have been conducted with the aim to automatically assess the difficulty of a text. Most involve English, both with the aim of simplifying administrative texts and of second/foreign language learning purposes, though recent years have seen a rise in projects involving languages other than English such French, Swedish, Dutch, German and Portuguese.

These projects are generally based on a collection of texts that is used as gold standard upon which a text classification system is trained. This gold standard collection of texts can be formed by second language coursebooks (François and Fairon, 2012; Pilán and Volodina, 2018), also with the addition of learner produced texts (Pilán and Volodina, 2018). Other approaches have considered exam texts (Branco et al., 2014), or exams texts and native texts combined (Xia et al., 2011). The texts composing the gold standard of the system have also been chosen in order to represent a range of reading skills (Velleman and Van der Geest, 2014), or to cater for different kinds of readers, both native and non-native, with both generalist and specialist reading needs (Vajjala and Meurers, 2016).

Flesch-Kincaid (Kincaid and Lieutenant Robert, 1975), Coh-metrix (Graesser et al., 2004) and CTAP (Xia et al., 2011) are possibly three of the most widely known automatic assessment systems for text difficulty, though all based on English. In the field of Italian, the three main approaches developed so far are: the Flesch-Vacca formula, an adaptation of the Flesch-Kincaid formula for English (Franchina and Vacca, 1986), the GulpEase index (Lucisano and Piemontese, 1988), and READ-IT (Dell'Orletta et al., 2011). In the first system, the complexity of a text is measured on the basis of average length of words, based on syllables, and average length of sentences, based on words. Moreover, the output provided by the formula indicates an approximate number of years that a reader needs to have spent in the education system in order to be able to comprehend a certain text. The information provided by the second system, the GulpEase index, has a number of distinctive characteristics compared to the Flesch-Kincaid formula. First of all, it is created directly on and for the Italian language. Secondly, though it includes the average length of words as well as the average length of sentences, similarly to what we find in the Flesch-Kincaid formula, the former is calculated on the basis of letters, not syllables, and this aids the automatic treatment of the text. Finally, READ-IT is based on a list of raw text, lexical, morpho-syntactic, and syntactic features, that are used with a training corpus of texts based on newspaper articles, in order to develop a statistical model able to provide an automatic assessment of an inputted text.

The three projects outlined above, related to the assessment of Italian texts in relation to their difficulty, are aimed to

create resources to simplify administrative texts and guarantee information accessibility to all, particularly to those with low literacy skill levels or with forms of mild cognitive impairment. To the best of our knowledge, the only other study in line with the present one, i.e. aiming to connect the automatic assessment of text difficulty with CEFR level categorisation, and specifically focused on Italian, is (Forti et al., 2019).
The main differences between the present work with the previous one are:

- a more extended set of linguistic features,
- a wider corpus,
- a more extended use of feature selection tools,
- a website where the system can be accessed and used by teachers, trainers and people involved in assessment activities.

To the best of our knowledge, this work includes the most comprehensive set of linguistic features used to develop an automatic text classification system for the purposes of Italian L2 learning and teaching. In addition to the more traditionally used features (Dell'Orletta et al., 2011), this work incorporates discursive features as well as the relatively newly developed morphological complexity index. As a result, it aims to push the boundaries of text difficulty classification systems in two respects: 1) by linking the automatic assessment of a text to Italian CEFR levels; 2) by incorporating the linguistic features previously found in separate studies into a single study.
The rest of the article is organised as follows. Section 2. describes the collected corpus of texts, while the design and the implementation of the computational system is illustrated in Section 3.. The experimental reliability of the proposed system is investigated in Section 4., while Section 5. analyses the impact of the different categories of linguistic features. Finally, Section 6. concludes the paper by also outlining possible future lines of research.

2. The Corpus

With the aim of creating a system able to assign Italian texts to a specific CEFR level, a corpus was collected in order to be used as a gold standard for training text classification systems. Most of the texts derive from language certification exams, extracted from the CELI (Certificati di conoscenza della lingua italiana) item bank maintained at the University for Foreigners of Perugia (Italy). The texts are taken from language certification materials at B1, B2, C1 and C2 levels. We consider text classification procedures performed for language certification purposes as considerably reliable because they respond to the specific needs of proficiency level assessment. In this respect, they exhibit a higher degree of external validity if compared to texts found in language coursebooks. For this reason, the 13 texts included in our corpus deriving from Italian language coursebooks underwent an additional phase of level assignment. To this end, we asked two professional language test developers to decide which level each text taken from the coursebook materials was most likely representative of.

Level	#Texts	#Types	#Tokens
B1	249	7 494	45 695
B2	185	12 743	90 133
C1	139	14 089	95 515
C2	119	15 709	104 679
Corpus	692	29 983	336 022

Table 1: Characteristics of the Corpus

When comparing the two sets of ratings, we observed that in two cases they were distanced by more than one level (e.g. B1 vs. C1). The two texts exhibiting this situation were removed from the text pool, thus leaving a total of eleven texts. This additional rating phase confirmed the potential vagueness that may be involved in text grading performed for publishing purposes, and this is why we chose not to use entire language coursebooks, without some sort of cross-verification of the level assigned to the texts included. Level A texts were excluded because of their very short length, which would have hindered the reliability of the text classification system.
With the entire set of 692 selected texts, including 336 022 tokens (see Table 1), the corpus was xml-annotated and post-tagged. The tagset was the same as the one used for the tagging of the Perugia corpus, a reference Italian corpus (see (Spina, 2014)). The main characteristics of the corpus are shown in Table 1.

3. The Computational System

In this section we describe the main architecture of the computational system and its main components: the classification model and the computation of the linguistic features.

3.1. System architecture

The problem of automatically measuring text complexity through CEFR proficiency levels was cast to a supervised classification problem. The labeled corpus discussed in Section 2. was used in order to train a classification model for automatically predicting the proficiency level of any previously unseen text in input.
Figure 1 depicts the high level architecture of the system. The classification model does not directly work with the texts in their pure form: any text is converted into a vector of numeric features – computed as described in Section 3.3. – and then passed on to the classification model.
First, the inner parameters of the classification model are trained using the labeled vectors corresponding to the texts in the considered corpus. Then, any unlabeled text is *vectorised* and fed to the trained model which predicts its proficiency level.
This architecture allows, on the one hand, to use the most common classification models available in the machine learning literature (Shalev-Shwartz and Ben-David, 2014) and, on the other hand, to build a classification model based only on the linguistic features of the texts that, we think, are what discriminating texts from the point-of-view of the CEFR levels.

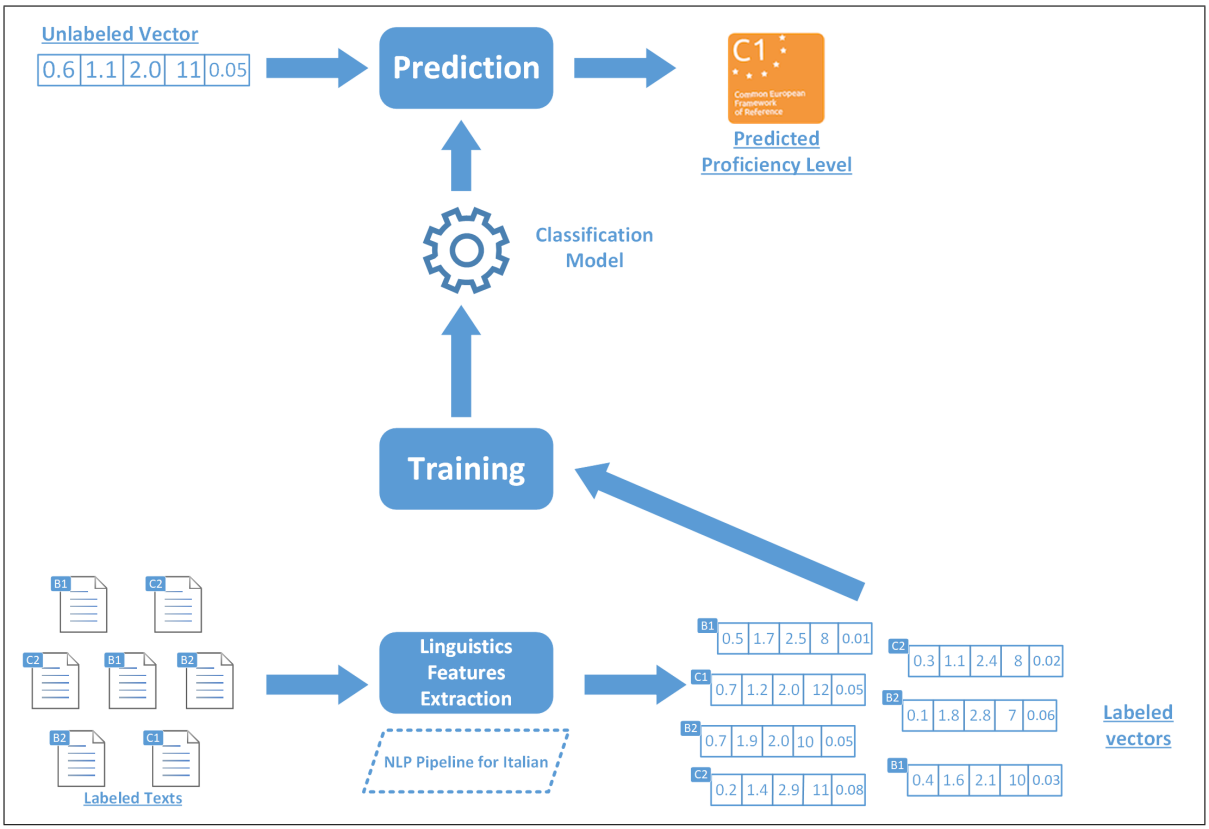

Figure 1: Architecture of the System

Finally, a user friendly web interface was developed as depicted in Figure 2: the user types or pastes a text of his/her choice in the provided text area, press the "Analyse" button, then the system transparently executes the prediction procedure of the trained model and shows the predicted CEFR level for the given text, together with a bar chart showing how the four different levels are represented within the text in terms of percentages. The developed resource is freely available on the internet at the following address: `https://lol.unistrapg.it/malt`. It may be accessed and explored by both fellow researchers and second/foreign language teachers.

3.2. The classification model

Regarding the classification model, we made some preliminary experiments using decision trees, random forests and support vector machines. A report of such experiments is provided in (Forti et al., 2019; Milani et al., 2019). According to the preliminary results, this work focuses on the Support Vector Machines (SVM) model.

An SVM (Shalev-Shwartz and Ben-David, 2014) is a supervised classification model which, given a (training) set of labeled numeric vectors, constructs a set of hyperplanes in a high-dimensional space, which identify the regions of the space corresponding to the different labels, i.e., the CEFR levels in our case.

The SVM implementation of the popular *Sci-Kit Learn* library (Pedregosa et al., 2011) has been used, while the Gaussian radial basin functions have been considered as kernel functions of the SVM.

3.3. Linguistic features

On the basis of a number of previous works (Dell'Orletta et al., 2011; Xiaobin and Meurers, 2016; Grego Bolli et al., 2017; Brezina and Pallotti, 2016; Gyllstad et al., 2014; Norris and Ortega, 2009), we defined a set of 139 linguistic features and implemented them by relying on well known NLP tools for the Italian language. In particular, UDPipe (Straka and Straková, 2017) was used for tokenization, lemmatisation, POS tagging and to build dependency trees, while OpenNER (García-Pablos et al., 2013) was adopted in order to compute the constituent trees.

For the sake of presentation, we divide the 139 features into six categories which are described in the following sections.

3.3.1. Raw text features

Raw text features are the most elementary type of features considered here and they were computed through the tokenisation of the inputted text. They include:

- *Sentence Length in Tokens* – mean and standard deviation, across all sentences, of the number of words in a sentence.
- *Token Length* – mean and standard deviation, across all tokens, of the number of characters in a token.
- *Text Length in Sentences* – number of sentences in the text.
- *Text Length in Lemmas* – number of lemmas in the text.

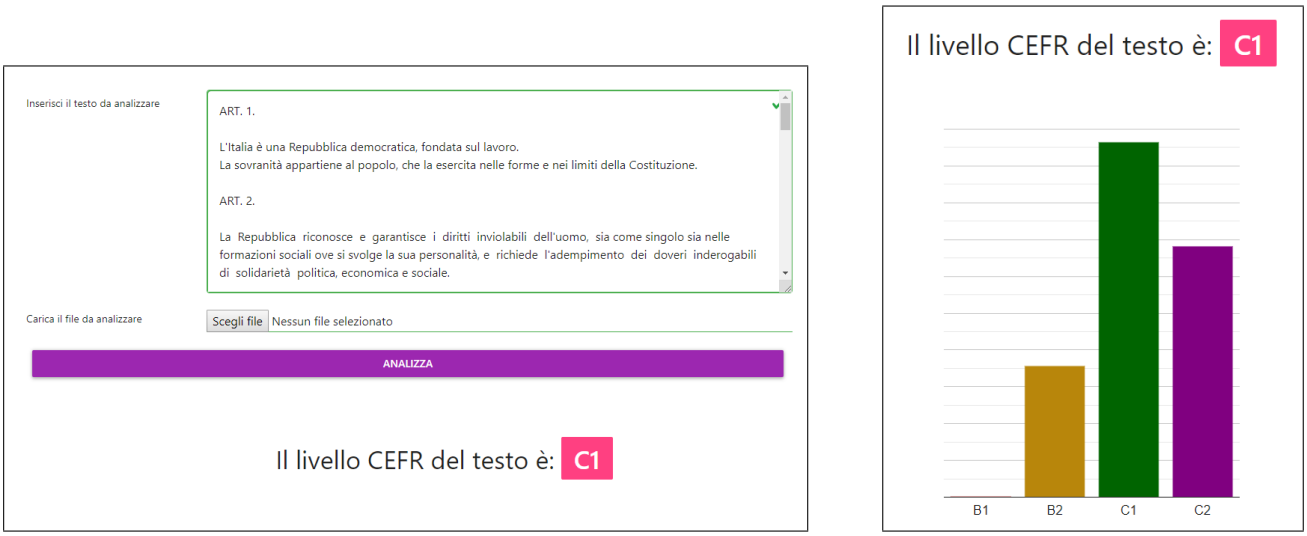

Figure 2: User interface: the input form and the results of the elaboration

3.3.2. Lexical features

Lexical features are mainly based on the lemmatization of the texts. They include:

- *Basic Vocabulary Rate* – with reference to the Nuovo Vocabolario di Base (New Basic Italian Vocabulary) (NVdB) (De Mauro and Chiari, forthcoming), the number of lemmas belonging to Fundamentals (the first 2000 most frequent words), High Usage (frequency ranks between 2000 and 4300) and High Availability (identified in (De Mauro and Chiari, forthcoming)) through a native speaker judgment questionnaire) wordlists.
- *Nouns Abstractness Distribution* – the number of nouns considered as Abstract, Semiabstract and Concrete (as used in (Grego Bolli et al., 2017)).
- *Lexical Diversity* – Ratio between the total number of words and the total number of unique words, within 100 random words. This unit of measure was chosen in order to have an homogeneous basis upon which to compute this feature.
- *Lexical Variation* – Type-Token Ratio (TTR), within 100 random words, where Type and Token may refer to different lexical categories. We considered TTR with types of a lexical category (adjective, verb, adverb, noun) and tokens either of the same category or of the four categories altogether.
- *Lexical Sophistication* – mean and standard deviation, across the sentences, of the occurring frequency of function tokens, function lemmas, lexical tokens and lexical lemmas. Information about frequency is taken from COLFIS dictionary (Bertinetto et al., 2005).

3.3.3. Morphological features

Morphological features are reflected by the Morphological Complexity Index (MCI) computed for two word classes: verbs and nouns. The MCI is operationalised by randomly drawing sub-samples of 10 forms of a word class (e.g. verbs) from a text and computing the average within-sample and across-samples of inflectional exponents. Further details can be found in (Brezina and Pallotti, 2016).

3.3.4. Morpho-syntactic features

Morpho-syntactic features are computed on the basis of part-of-speech (POS) tagging and the morphological analysis was performed on the inputted text (as largely used in (Dell'Orletta et al., 2011)). They include:

- *Subordinate Ratio* – Mean and standard deviation of the percentage of subordinate clauses over the total number of clauses.
- *POS Tags Distribution* – Normalized entropy and distribution of tokens over all the POS tags.
- *Verbal Moods Distribution* – Normalized entropy and distribution of the seven verbal moods (indicative, subjunctive, conditional, imperative, gerund, participle, infinite) among the verbal tokens.
- *Dependency Tags Distribution* – Normalized entropy and distribution of the tokens over all the dependency tags.

3.3.5. Discursive features

Discursive features comprise the main elements that are used to organise the text in terms of its cohesive structure (as most typically used in (Graesser et al., 2004)). They include:

- *Referential Cohesion* – This trait counts the nominal types that groups of three adjacent sentences have in common. We use its mean and standard deviation across all groups to condense the information.
- *Deep Causal Cohesion* – Normalized entropy and distribution of the eight classes of connectives (causal, temporal, additive, adversative, marking results, transitions, alternative or reformulation/specification), which play an important role in the creation of logical relations within text meanings, and provide clues about text organization.

3.3.6. Syntactic features

Our list of syntactic features aims to reflect the main characteristic of syntactic constituents,also in terms of dependency relations (as largely seen in (Xia et al., 2011)). It includes:

- *Depth of the Parse Trees* – Maximum depth among all the dependency trees.
- *Non-Verbal Chains Length* – Mean and standard deviation of the length of the paths without verbal nodes in the dependency trees.
- *Maximal Non-Verbal Phrase* – Mean and standard deviation of the dimension of the maximal nominal phrases in the constituent trees of the inputted text, where the dimension of a node is the number of terminal nodes beneath that node.
- *Verbal Roots* – Percentage of dependency trees with a verbal root.
- *Arity of Verbal Predicates* – Distribution of the arity of verbal nodes, where the arity is the number of dependency links with that node as head.
- *Relative Subordinate Order* – Distribution of the distance between the main clause and each subordinate clause.
- *Subordination Chains Length* – Distribution of the depth of chains of embedded subordinate clauses.
- *Dependency Links Length* – Mean, standard deviation and maximum of the number of words occurring between a syntactic head and a dependent.
- *Mean Length of Clauses (Tokens)* – Mean length of clauses expressed in tokens.
- *Number of Syntactic Constituents* – Counts the occurrences of a specific syntactic constituent in the text. The following elements are considered: clauses, nominal phrases per sentence (mean across sentences), coordinate phrases, and subordinate clauses.
- *Syntactic Complexity Feature* – Calculates the syntactic complexity of the text in terms of average coordinate phrases per clause, sentence complexity ratio (#clauses / #sentences), and sentence coordination ratio (#coordinating clauses / #sentences).

4. Experiments

In order to assess the accuracy of the prediction system, computational experiments were held using the corpus of texts described in Section 2. as dataset.

Every experiment – tuning the SVM hyper-parameters, selecting the features, and assessing the final accuracy of the system – was performed using 5 repetitions of a stratified 10-folds cross-validation executed on the whole dataset of 692 texts.

First, the hyper-parameters C and γ of the SVM model have been tuned by means of a grid search process aimed at optimising the F1 score measure. The whole set of 139 features was considered and the calibrated setting is $C = 2.24$ and $\gamma = 0.02$.

Actual \ Predicted	B1	B2	C1	C2
B1	**214.6**	30.2	4.2	0.0
B2	28.4	**129.8**	23.8	3.0
C1	9.6	28.8	**74.2**	26.4
C2	1.2	9.0	30.0	**78.8**

Table 2: Confusion Matrix

Then, a *features selection* additional phase was performed by means of the well known Recursive Features Elimination (RFE) algorithm (Guyon et al., 2002). RFE recursively fits the model and removes the weakest feature until a specified number of features is reached. In our work, the well known *permutation feature importance* technique (Fisher et al., 2018) was considered to measure the importance of every feature during the last model fitting. Moreover, to find the optimal number of features, cross-validation was used with RFE to score different feature subsets and select the best scoring collection of features. As depicted in Figure 3, a subset formed by 54 features – around the 39% of the whole set of features – has obtained the best F1 score in our experiments.

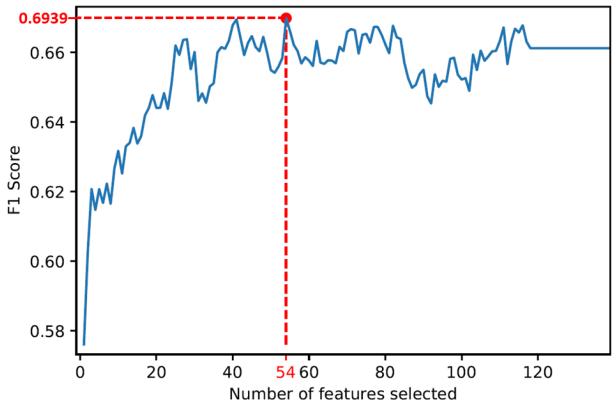

Figure 3: Features selection graph

This set of features is further analysed in Section 5., while the performances of the tuned SVM model trained using the selected features are shown in the confusion matrix provided in Table 2. In this table, each entry X, Y provides the average number – over the 5 repetitions of the 10-folds cross-validation process – of texts which are known to belong to the CEFR level X, but have been classified by our system to the CEFR level Y.

The correctly classified texts are those accounted in the diagonal of the confusion matrix. They are (in average) 497.4 out of 692, thus the accuracy of our system is about 71.88%.

The confusion matrix also allows to derive the precision and recall measures (Shalev-Shwartz and Ben-David, 2014) for all the considered CEFR levels. Our experiments reveal that the B1 level exhibits the highest precision and recall (respectively, 84.55% and 86.18%), while the weakest predictions are those regarding the C1 level (which has 56.13% and 53.38% as, respectively, precision and recall).

Furthermore, it is interesting to observe that most of the incorrectly classified texts are only one level away from their actual CEFR levels. In fact, by aggregating the pairs of levels B1,B2 and C1,C2 into the macro-levels B and C, respectively, we obtain that the average accuracy of the system increases up to 88.50%.

Finally, note that the results discussed in this section are also in line with the 2D visualisation of the dataset provided in Figure 4, where each point is the two-dimensional representation of a text in the dataset obtained by means of the well known dimensionality reduction technique t-SNE (Van der Maaten and Weinberger, 2012) executed on the 139-dimensional representation of the dataset.

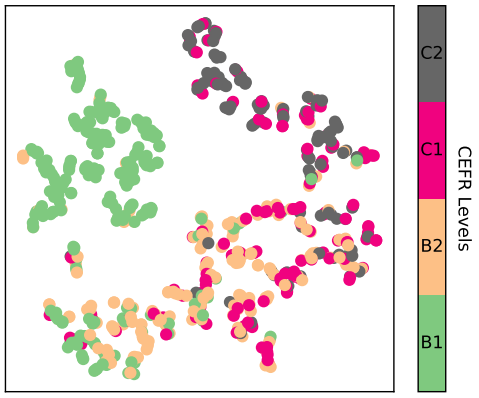

Figure 4: 2D visualisation of the dataset

5. Analysis of the Features

The features selection algorithm used identified a subset of 54 features within the total of 139 linguistic features as those with overall discriminatory power.

The graph in Figure 5 shows the distribution of selected features within each feature category (see Section 3.3.).

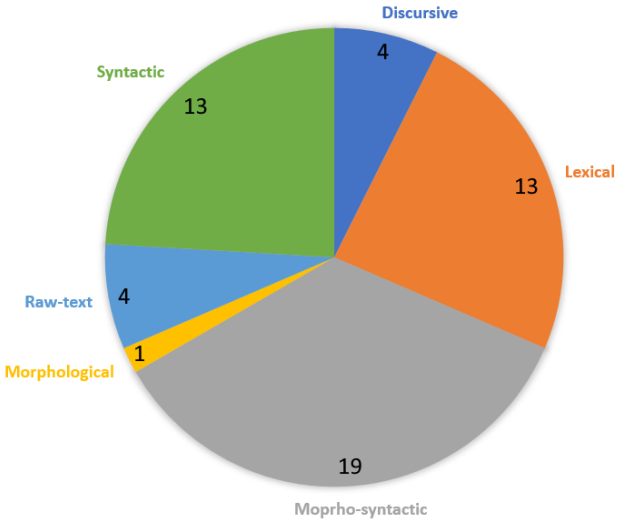

Figure 5: Category distribution in selected features

As we can see, morpho-syntactic features constitute the category with most features represented in the final set (n = 19), followed by lexical and syntactic features (n = 13 each), raw-text and discursive (n = 4 each) and finally morphological (n = 1). It should be said that the morpho-syntactic category of features was also initially the broadest one (n = 68), followed by the syntactic (n = 33), lexical (n = 21), discursive (n = 9), raw-text (n = 6) and morphological (n = 2) categories.

Furthermore, for each selected feature, we computed its *permutation feature importance* (PFI) score using the technique described in (Fisher et al., 2018), i.e. the PFI of a feature is the mean percentage decrease of F1 score obtained by replacing the considered feature with random noise and cross validating again the model. The graph in Figure 6 shows the distribution of the PFI scores averaged among the different categories of linguistic features in order to see which of them had a higher impact in discriminating the classification of the texts. As we can see, the category of features with the highest discriminating power is the category of syntactic features, very closely followed by raw-text and lexical features. Morpho-syntactic, discursive and morphological features represent the three categories with the lowest discriminatory power. As a result, even though the morpho-syntactic features are those most prominently selected by the algorithm, possibly also due to their numerosity to begin with, these were ultimately not the ones with most discriminatory power. Moreover, we note the closeness between the lexical and syntactic categories of features, indicating they both exhibit a considerable role is the automatic assessment of text difficulty. This can be particularly interesting in light of the studies regarding the lexis-syntax interface in language. Finally, these results appear to be in line with an earlier preliminary study conducted in 2017: morpho-syntactic features were, even then, the set of features which were least discriminatory in comparison to lexical and raw-text features; syntactic features were not included at that stage (Grego Bolli et al., 2017).

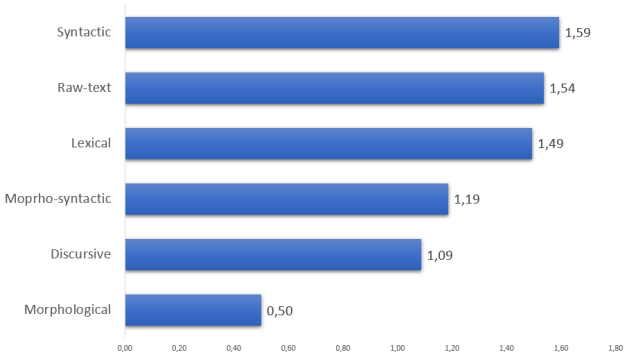

Figure 6: Average PFI scores across categories

6. Conclusion and Future Work

This study described the development and assessment of a new system trained to automatically classify written texts according to their difficulty, and assign them to a CEFR level. We showed how the developed system exhibits considerably good accuracy levels (71.88%), and we also ob-

served how by aggregating the level pairs into single B and C levels, the accuracy increased up to 88.50%. We also noted that the highest precision and recall values are obtained in relation to the B1 level, and that the texts that were incorrectly classified were only one level away from their actual CEFR level.

This study certainly lends itself to further investigation in relation to the single linguistic features composing each feature category.

7. Acknowledgements

This research was supported by the grant 2018.0424.021 *MALT-IT2. Una risorsa computazionale per Misurare Automaticamente la Leggibilità dei Testi per studenti di Italiano L2*, co-funded by the University for Foreigners of Perugia and by the Fondazione Cassa di Risparmio di Perugia.

8. Bibliographical References

Bachman, L. and Palmer, A. (2010). *Language Assessment in Practice*. Oxford University Press.

Branco, A., Rodrigues, J., Costa, F., Silva, J., and Vaz, R. (2014). Rolling out text categorization for language learning assessment supported by language technology. *Computational Processing of the Portuguese Language*, pages 256–261.

Brezina, V. and Pallotti, G. (2016). Morphological complexity in written L2 texts. *Second Language Research*, 35(1):99–119, July.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (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, Edinburgh, Scotland, UK, July. Association for Computational Linguistics.

Fisher, A., Rudin, C., and Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the "rashomon" perspective. *arXiv preprint arXiv:1801.01489*.

Forti, L., Milani, A., Piersanti, L., Santarelli, F., Santucci, V., and Spina, S. (2019). Measuring text complexity for Italian as a second language learning purposes. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 360–368, Florence, Italy, August. Association for Computational Linguistics.

Franchina, V. and Vacca, R. (1986). Adaptation of flesh readability index on a bilingual text written by the same author both in Italian and English languages. *Linguaggi*, 3:47–49.

François, T., Brouwers, L., Naets, H., and Fairon, C. (2014). AMesure: a readability formula for administrative texts (AMESURE: une plateforme de lisibilité pour les textes administratifs) [in French]. In *Proceedings of TALN 2014 (Volume 2: Short Papers)*, pages 467–472, Marseille, France, July. Association pour le Traitement Automatique des Langues.

François, T. and Fairon, C. (2012). An "AI readability" formula for french as a foreign language. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477.

García-Pablos, A., Cuadros, M., Gaines, S., and Rigau, G. (2013). Opener demo: Open polarity enhanced named entity recognition. *Come Hack with OpeNER*, page 12.

Graesser, A., McNamara, D., Louwerse, M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, And Computers*, 36:193–202.

Grego Bolli, G., Rini, D., and Spina, S. (2017). Predicting readability of texts for Italian L2 students: A preliminary study. *Learning and Assessment: Making the Connections*, page 272.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

Gyllstad, H., Granfeldt, J., Bernardini, P., and Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. 14:1–30.

Kincaid, P. and Lieutenant Robert, P., F. R. (1975). Derivation of new readability formulas for navy enlisted personnel. *Research Branch Report, Millington, TN: Chiefof Naval Training*, pages 8–75.

Lucisano, P. and Piemontese, M. (1988). Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città*, 31(3):110–124.

Milani, A., Spina, S., Santucci, V., Piersanti, L., Simonetti, M., and Biondi, G. (2019). Text classification for italian proficiency evaluation. In *Computational Science and Its Applications – ICCSA 2019*, pages 830–841, Cham. Springer International Publishing.

Norris, J. M. and Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4):555–578, November.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pilán, I. and Volodina, E. (2018). Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58.

Purpura, J. (2014). Cognition and language assessment. In *The Companion to Language Assessment volume III*, pages 1,453–1,476.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Spina, S. (2014). Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth*

International Workshop EVALITA 2014, pages 354–359, Pisa, Italy. Pisa University Press.

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, August. Association for Computational Linguistics.

Vajjala, S. and Meurers, D. (2016). Readability-based sentence ranking for evaluating text simplification. *Int. Jour. of Applied Linguistics*, pages 194–222.

Van der Maaten, L. and Weinberger, K. (2012). Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, Sep.

Velleman, E. and Van der Geest, T. (2014). Online test tool to determine the CEFR reading comprehension level of text. *Procedia Computer Science*, pages 350–358.

Xia, M., Kochmar, E., and Briscoe, T. (2011). Text readability assessment for second language learners. *Proc. of the 11th Works. on Innov. Use of NLP for Building Educ. Appl.*, pages 12–22.

Xiaobin, C. and Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. *Proc. of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 113–119.

9. Language Resource References

Bertinetto, P. M., Burani, C., Laudanna, A., Marconi, L., Ratti, D., Rolando, C., and Thornton, A. M. (2005). Corpus e lessico di frequenza dell'italiano scritto (CoLFIS). *Scuola Normale Superiore di Pisa*.

De Mauro, T. and Chiari, I. (forthcoming). *Il Nuovo Vocabolario di Base della Lingua Italiana*.